Medical Artificial Intelligence Laboratory
At Yonsei University

MR Image Synthesis

ResViT

TMI

2022

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

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur*, *Senior Member*

***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

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

MR Image Synthesis

ResViT

Background & Goal

Background

- CNN based models: Only learns local information and fails to utilize long-range contextual information
- Transformer based models: Leverages long-range contextual information, but has high computational cost

Goal

- ResViT that translates between multi-modal imaging data combines the sensitivity of Vision Transformers to gl obal context, the localization power of CNNs, and the realism of Adversarial learning

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

MR Image Synthesis

ResViT

Network Architecture

Generator subnetworks

- Encoder: To capture a hierarchy of localized features of source images using Convolutional layers
- Information Bottleneck: To distill task-relevant information in the encoded features using CNN + Transformer
- Decoder: To distill multi-modal images in separate channels using Transposed convolutional layers
- Parameter Sharing Transformers: To avoid inevitably elevate memory demand and risk of overfitting, weight sha ring strategy is adopted where the model weights for the transformer encoder are tied across separate ART blo cks.

Discriminator subnetworks

- Discriminator is based on a conditional PatchGAN architecture

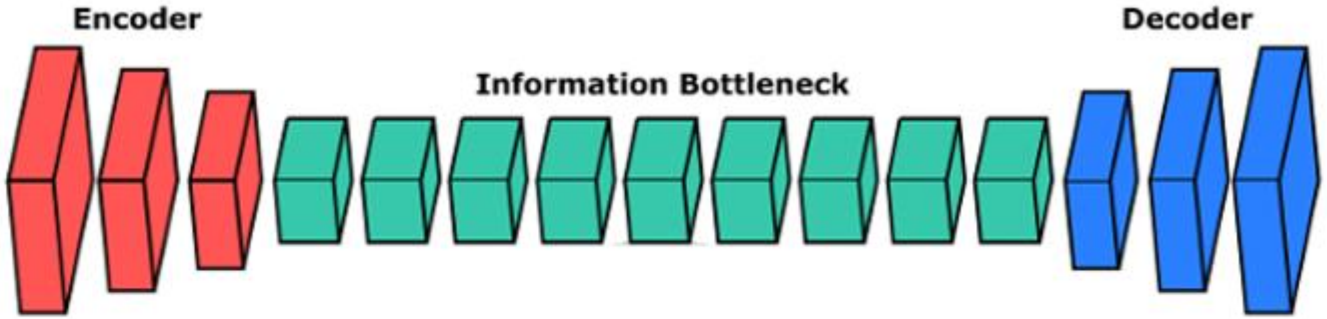

Fig 1. Overview of ResViT Generator

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

ResViT

Network Architecture

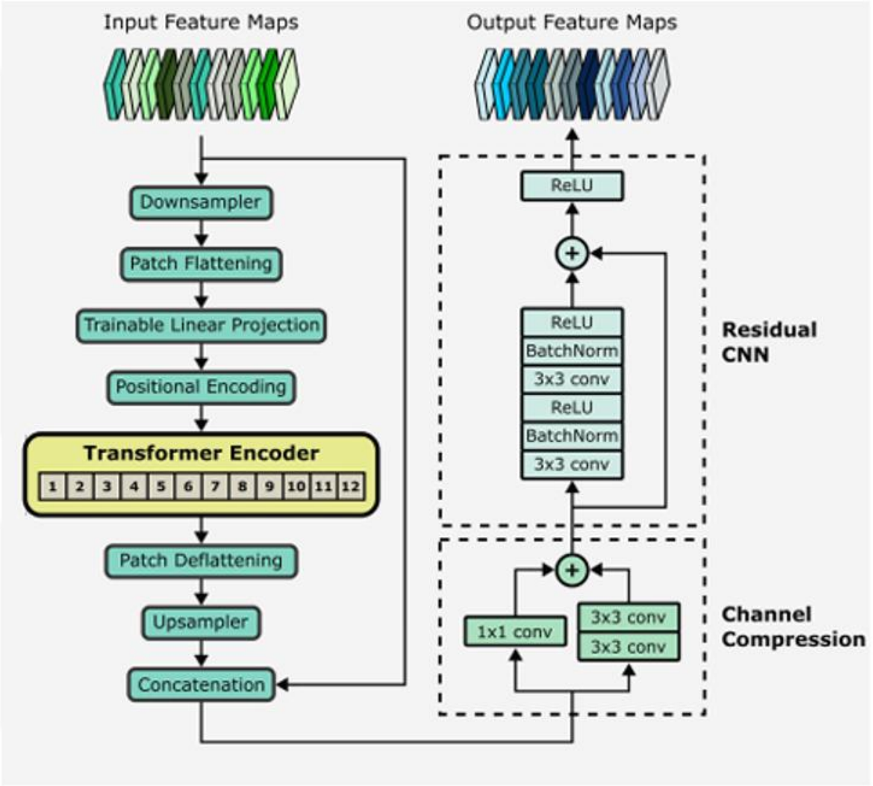

Fig 2. Overview of ART Block

Information Bottleneck

- To maintain both localization power and contextual sensitivity, ART blocks that aggregate information from residual convolutional and transformer branches

Residual Blocks

- The reason for using Residual Transformer and Residual CNN module is to learn a unified representation that considers local + contextual information together

Channel Compression

- The process of optimally combining the outputs of CNN and Transformer

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging, 41*(10), 2598-2614.

MR Image Synthesis

ResViT

Network Architecture

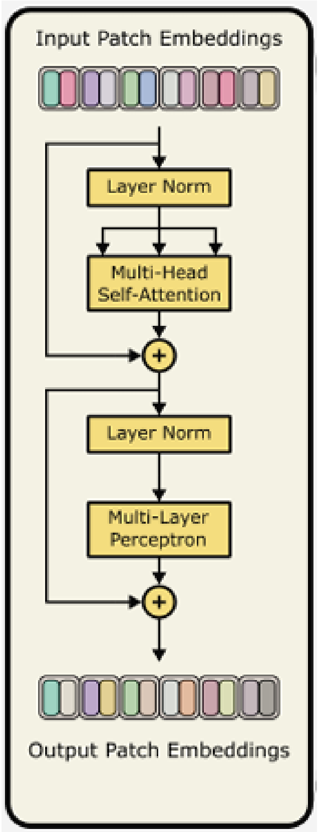

Fig 3. Overview of Transformer

Transformer

- The role of learning long range context information

Multi-Head Self Attention

- Learning the relationship between input features with multip le attention heads

Multi-layer Perceptron

- Learn more useful representations by nonlinearly transformi ng the output of the transformer

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

MR Image Synthesis

ResViT

Network Architecture

Decoder

- Synthesize all contrasts within the multi-modal protocol regardless of the specific source-target configuration

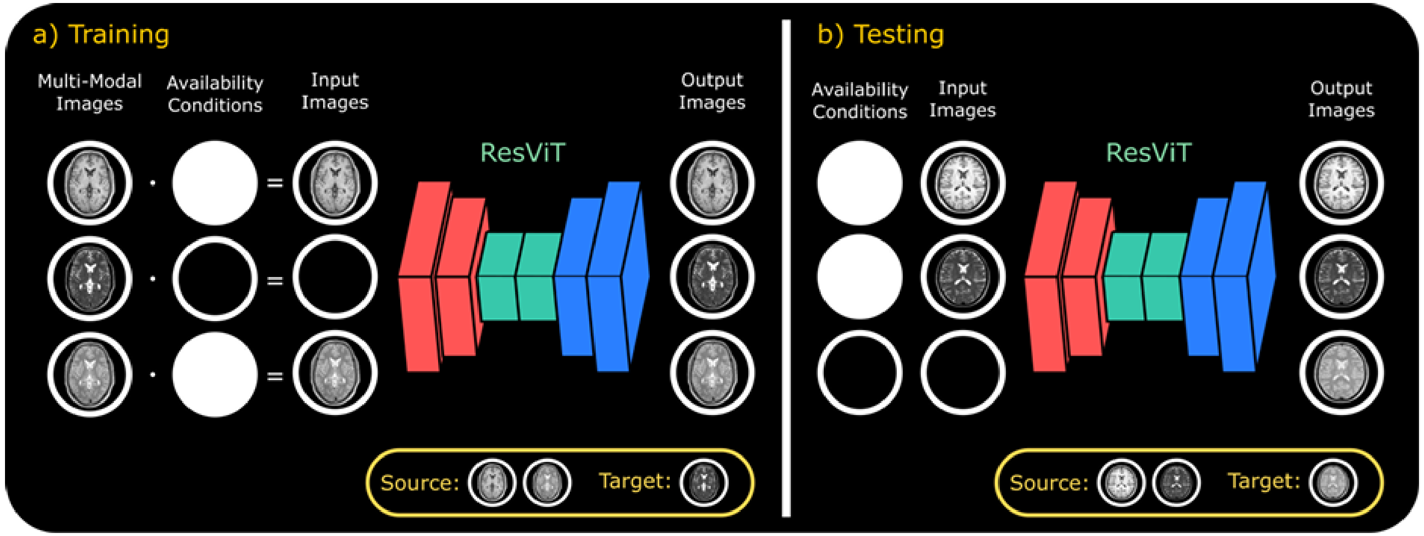

Fig 2. a) Training Process: ResViT takes as input the entire set of images within the multi-modal protocol, Multiple configurations of source-target modalities are expressed as availability conditions in ResViT.
b) Inference Process: For each test subject, a specific source-target configuration is determined through availability conditions. This allows the model to perform appropriate modality transformation.

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

MR Image Synthesis

ResViT

Loss Functions

Pixel-wise L1 loss

- Minimize the difference between the real target image and the synthetic image

Pixel-wise consistency loss

- Minimize the difference between the real source image and the reconstructed source image

Adversarial loss

- Adversarial loss is based on Least-Squares GAN loss

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

Eq 1. pixel-wise L1 loss

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

Eq 2. pixel-wise consistency loss

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

Eq 3. adversarial loss

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

MR Image Synthesis

ResViT

Results

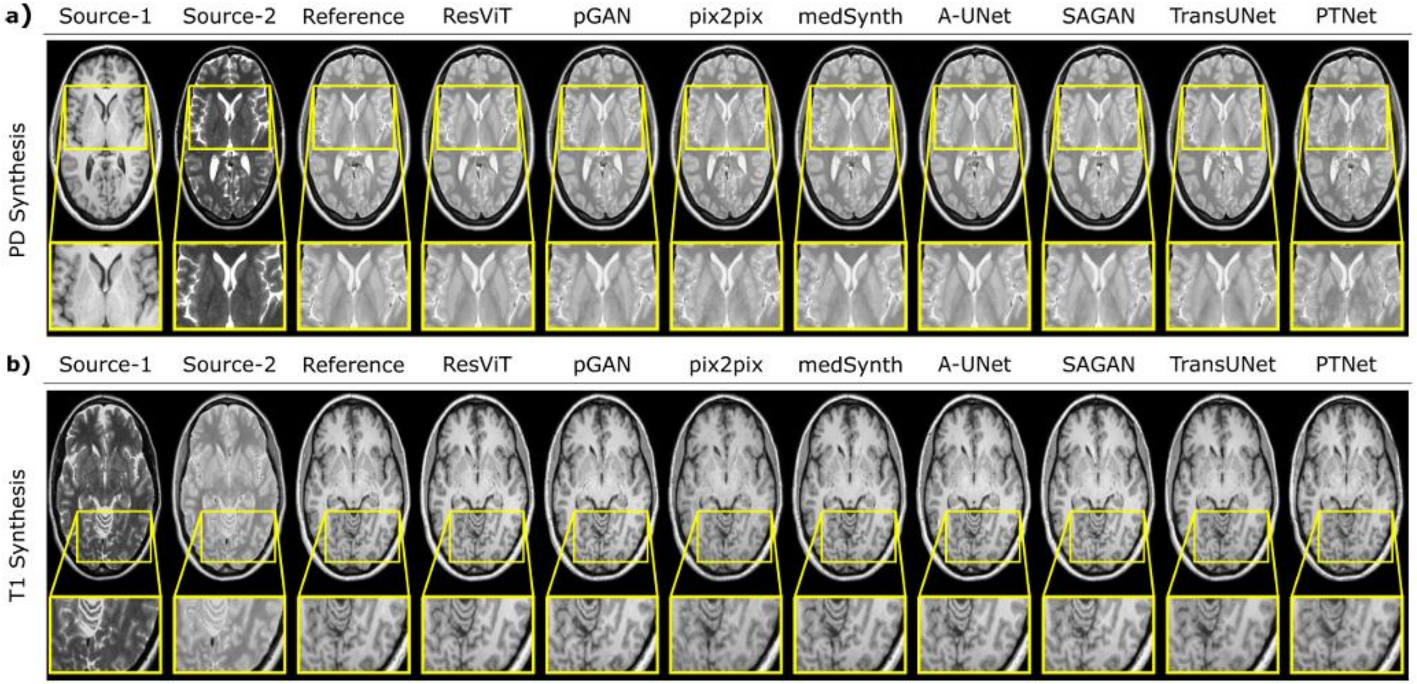

Fig 3. Demonstrated on the IXI dataset for two representative many-to-one synthesis tasks

	$T_1, T_2 \rightarrow PD$		$T_1, PD \rightarrow T_2$		$T_2, PD \rightarrow T_1$		$T_2 \rightarrow PD$		$PD \rightarrow T_2$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ResViT	**33.92 ±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 1. Performance of task-specific synthesis models in the IXI dataset

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

MR Image Synthesis

ResViT

Results

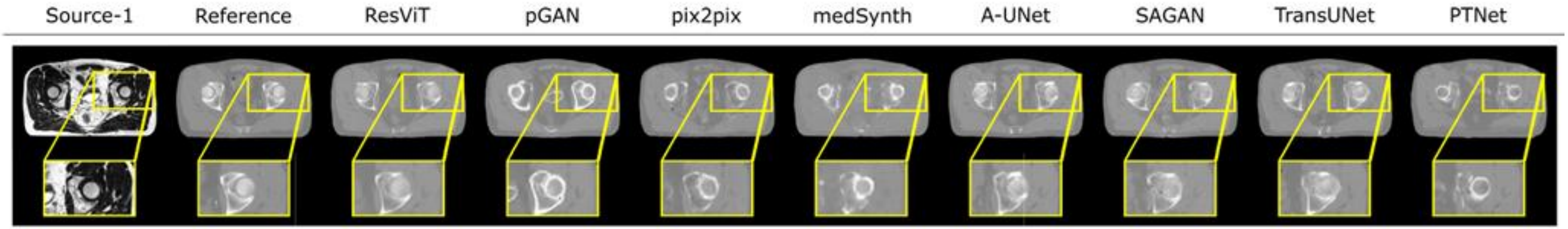

Fig 4. Demonstrated on the pelvic MRI-CT dataset for the T2-weighted MRI→CT task

		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

Table 2. Performance for the across-modality synthesis task(T2-w MRI → CT)

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

Medical Artificial Intelligence Laboratory
At Yonsei University

MR Image Synthesis
ResViT

Results

	$T_1, T_2 \rightarrow$ PD			$T_1, T_2 \rightarrow$ FLAIR			MRI $\rightarrow$ 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 3.Performance of ResViT and Ablated of Transformer Modules, Convolutional Modules, or Adversarial loss

	$T_1, T_2 \rightarrow$ PD		$T_1, T_2 \rightarrow$ FLAIR		MRI $\rightarrow$ 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 4. Performance of ResViT and Ablated of weight tying and individual transformer modules

	$T_1, T_2 \rightarrow$ 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 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

Table 5. Performance of ResViT and Ablated of pre-training and delayed insertion procedures for Transformers

	$T_1, T_2 \rightarrow$ 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

Table 6. Performance of ResViT and Variants ablated

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.

MR Image Synthesis

ResViT

Implications & Limitations

Implications

- Uniquely introduce many-to-one synthesis models and a unified model that generalizes across multiple source -target configurations
- Propose a hybrid architecture that combines localization capabilities of CNNs with contextual sensitivity of tran sformers

Limitations

- Learn the distribution of the target modality implicitly, without explicitly evaluating the likelihood → Mode coll apse
- Lack of reliability in network mapping because images are generated using the "One-Shot Sampling" method

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598-2614.