edaGAN: Encoder-Decoder Attention Generative Adversarial Networks for Multi-contrast MR Image Synthesis

Onat Dalmaz [1,2§], Baturay Saglam[1§], Kaan Gönç[3] and Tolga Çukur[1,2,4]
[1] Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
[2] National Magnetic Resonance Research Center, Bilkent University, Ankara,Turkey
[3] Department of Computer Science, Bilkent University, Ankara, Turkey
[4] Neuroscience Program, Sabuncu Brain Research Center, Bilkent University, Ankara, Turkey
onat@ee.bilkent.edu.tr, baturay@ee.bilkent.edu.tr, kaan.gonc@bilkent.edu.tr, cukur@ee.bilkent.edu.tr

2022 9th International Conference on Electrical and Electronics Engineering (ICEEE) | 978-1-6654-6754-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICEEE55327.2022.9772555

***Abstract*—Magnetic resonance imaging (MRI) is the preferred modality among radiologists in the clinic due to its superior depiction of tissue contrast. Its ability to capture different contrasts within an exam session allows it to collect additional diagnostic information. However, such multi-contrast MRI exams take a long time to scan, resulting in acquiring just a portion of the required contrasts. Consequently, synthetic multi-contrast MRI can improve subsequent radiological observations and image analysis tasks like segmentation and detection. Because of this significant potential, multi-contrast MRI synthesis approaches are gaining popularity. Recently, generative adversarial networks (GAN) have become the de facto choice for synthesis tasks in medical imaging due to their sensitivity to realism and high-frequency structures. In this study, we present a novel generative adversarial approach for multi-contrast MRI synthesis that combines the learning of deep residual convolutional networks and spatial modulation introduced by an attention gating mechanism to synthesize high-quality MR images. We show the superiority of the proposed approach against various synthesis models on multi-contrast MRI datasets.**

Keywords*—*MRI, synthesis, attention, generative, adversarial

I. Introduction

Multi-modal medical imaging uses a variety of scans on the body to get complementing tissue information, boosting diagnosis accuracy and confidence. Because of its superior soft-tissue contrast, magnetic resonance imaging (MRI) is the preferred modality in clinical neuroimaging. Its capability to capture the anatomy under a variety of different contrasts allows it to gather additional diagnostic information over the course of an exam [1]. However, such multi-contrast MRI exams take a long time to scan, resulting in the collection of an insufficient number of contrasts [2]. Hence, subsequent scans' time and financial expenses significantly restrict its use. This significant constraint has generated interest in synthesis methods that can restore missing scans from a subset of available scans in multi-contrast MRI protocols. Synthetic multi-contrast MR images, in turn, can improve subsequent radiological observations, as well as image analysis tasks like segmentation and detection [3].

The goal of MRI synthesis is to predict target-contrast images for a subject based on available source-contrast images [4]. Since MR images are high dimensional, target-modality data is missing during inference, and there are nonlinear changes in tissue contrast across different contrasts, this is an ill-posed inverse problem [5]–[7]. However, the use of deep models to solve this complex problem has resulted in significant performance gains [8]–[14]. Generative Adversarial Networks (GAN) that leverage an adversarial loss have been a de-facto choice for MRI synthesis, with their increased capture of detailed tissue structure [8], [15]–[17]. The pioneering works in the area [8], [9], employed Convolutional Neural Network (CNN) based generator and discriminator architectures to leverage adversarial learning and emphasize the capture of details in the synthesized images. They also included perceptual [18] and gradient-difference losses [19], along with the pixel-wise and adversarial losses. In [10], the multi-modal nature of multi-contrast MRI was explored through employing sophisticated modality-fusion techniques to increase the performance in many-to-one synthesis tasks. A novel weakly-supervised synthesis technique is proposed in [11], to learn synthesis tasks from source-target pairs which were undersampled in k-space. In [12], the authors introduce a 3D volumetric approach to mitigate the data and computation limitations. In [13], the authors propose a hybrid CNN-transformer [20] architecture to increase the capture of long-range context in medical images. In [14], the authors leverage cross-attention transformers [21] for unconditional modeling of generative MR image priors.

Inspired by the powerful pGAN [8] and Attention U-net [22] models, we introduce **E**ncoder-**D**ecoder **A**ttention **G**enerative **A**dversarial **N**etworks (edaGAN), a novel deep generative model for high quality multi-contrast MRI synthesis. Similar to pGAN [8], edaGAN comprises a convolutional encoder and residual blocks to capture task-critical features and leverages Residual Encoder-Decoder attention (RED) blocks to fuse low-level features and modulate decoder feature maps effectively. RED blocks leverage gated attention modules along with concatenation and channel compression subblocks.

[§] Equal contribution

978-1-6654-6754-4/22/$31.00 ©2022 IEEE

Convolutional layers within the existent approaches [8], [16], [23] are limited in learning the global context within the images. We employ gated attention modules within RED blocks to exploit long-range spatial interactions within feature maps to overcome this. Attention U-net [22] model offers such an attention mechanism within the network for improving medical image segmentation tasks. Different from Attention U-net [22] proposed for CT segmentation, we propose the edaGAN model, which is specialized for multi-contrast MRI synthesis. edaGAN differs from Attention U-net [22] in two critical perspectives:

- Originally proposed for CT segmentation, Attention U-net [22] employs bilinear upsampling modules in the decoder for increasing the spatial resolution of the latent feature maps. This upsampling, in turn, hinders the frequency distribution of the generated images in image generation tasks [24]. edaGAN instead employs transposed-convolutional layers for learnable upsampling of the feature maps.
- Instead of an encoder-decoder pipeline, edaGAN leverages residual blocks between the two modules to deepen the network further. This way, it learns to extract more critical and abstract features for image synthesis.

We performed extensive experiments for synthesizing missing sequences in multi-contrast MRI. We use IXI (https://braindevelopment.org/ixi-dataset) and BRATS [25]–[27] datasets for benchmarking. MRI datasets from healthy and diseased subjects demonstrate the proposed method's advantage over competing methods.

II. Theory and Methods

A. edaGAN

edaGAN is a conditional and adversarial image synthesis model that consists of generator and discriminator subnetworks. The generator follows an encoder-residual blocks-decoder pipeline, depicted by Fig. 1.

- **Encoder:** Given the input contrast, the encoder utilizes strided convolutional layers to extract hidden representations.
- **Residual Blocks:** Following the encoder, residual blocks [8], [28] leverage convolutional operations with additive residual skip connections to further distill structural representations extracted by the encoder.
- **Decoder:** The decoder uses feature maps extracted by the residual blocks, along with the feature maps from the encoder. It leverages transposed convolutional layers in conjunction with RED blocks (see Fig. 2) to effectively fuse high-level feature maps with low-level information from the encoder feature maps.

B. RED Blocks

RED blocks comprises gated attention units [22], concatenation, and channel compression blocks. RED blocks gets inputs x_i^l and g_i^l from the decoder and encoder, respectively.

*1) **Gated attention**:* Attention gate (AG) assists the network in learning contextual relationships. Attention operation exploits the long-range spatial interactions within the image. Although convolution operators are better at learning localized features, attention operators stand out at learning contextual features [13], [20], [22], [29]. In turn, learning global context in medical images improves synthesis performance [13]. Given x_i^l and g_i^l, gated attention [22] performs the following operations:

$$q_{attn}^l = \psi^T(\sigma(W_{x,l}^T x^l + W_{g,l}^T g^l + b_{g,l})) + b^\psi, \tag{1}$$

$$\alpha^l = \sigma(q_{attn}^l(x^l, g^l; \Theta_{att})), \tag{2}$$

where $\sigma(x') = \frac{1}{1+e^{-x'}}$ denotes the sigmoid function, x^l denotes the output of the l-th encoder layer, and y^l denotes the output of the residual blocks when $l = 1$ and denotes the $(l-1)$-th decoder layer when $l > 1$. The attention gate consists of parameters, Θ_{att}, including linear transformations $W_x \in \mathbb{R}^{F_l \times F_{int}}$, $W_g \in \mathbb{R}^{F_g \times F_{int}}$, $\psi \in \mathbb{R}^{F_{int} \times 1}$ and bias terms $b_\psi \in \mathbb{R}$, $b_g \in \mathbb{R}^{F_{int}}$. The linear transformations are produced through channel-wise $1 \times 1 \times 1$ convolutions for input tensors. The output of AG $\hat{x_i}^l$ is the element-wise multiplication of the input x_i^l and attention coefficients α_i^l.

2) Concatenation and Channel Compression: After AG, the output map $\hat{x_i}^l$ is channel-wise concatenated with the input map x_i^l. The number of channels is halved through two convolutional layers with kernel size 3, parallel with a convolutional layer with kernel size 1. The output is then fed to the next decoder layer.

C. Loss Function

We incorporated two loss functions to optimize the edaGAN model, namely, pixel-wise and adversarial losses. Enforcing L_2 norm-based losses for generator networks in image-to-image translation tasks usually results in blurrier output images [16]. Hence, pixel-wise loss implemented via the L_1 norm of the difference between the synthesized and target images:

$$\mathcal{L}_1(G) = \mathbb{E}_{x,y}[||y - G(x)||_1] \tag{3}$$

L_1 norm-based loss enforces G to generate images consistent with the target contrast. However, networks that rely solely on pixel-wise loss terms tend to fail to capture high-frequency details in the synthesized images [16]. Therefore, an adversarial loss is employed to enforce realism on synthesized images:

$$\begin{aligned}\mathcal{L}_{condGAN}(G,D) = &- \mathbb{E}_{x,y}[(D(x,y)-1)^2] \\ &- \mathbb{E}_x[(D(x,G_x))^2],\end{aligned} \tag{4}$$

where x is the source contrast, y is the target contrast, and G and D are the generator and discriminator networks, respectively. Combining these two loss functions, the overall objective can be expressed as:

$$\mathcal{L}_{edaGAN} = \mathcal{L}_{GAN}(G,D) + \lambda_{L_1}\mathcal{L}_1(G) \tag{5}$$

where λ_{L_1} is a hyperparameter to determine the prioritization of the pixel-wise loss. Discriminator subnetwork is designated to be patchGAN [16].

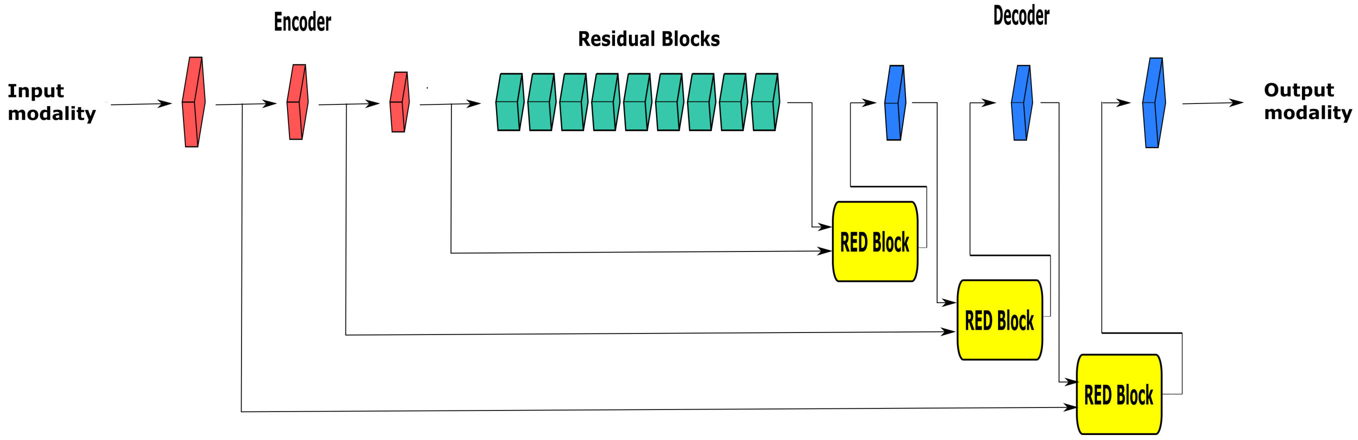

Fig. 1. The generator subnetwork of edaGAN consists of the encoder, residual blocks, RED blocks, and decoder. RED Blocks synergistically fuse low-level information from the encoder with the high-level structural representations learned via the residual blocks.

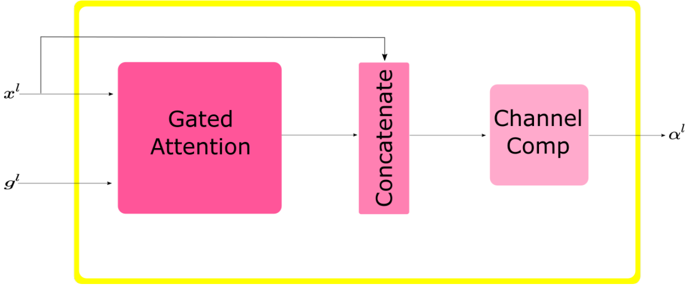

Fig. 2. Structure of the l-th RED block in which the input x^l is modulated by g^l via the gated attention module. The output of the gated attention is channelwise concatenated with input x^l. To keep the number of channels fixed throughout the network, channel compression block is employed to halve the number of channels.

D. Datasets

We demonstrate the effectiveness of the edaGAN model on two multi-contrast brain MRI datasets: IXI and BRATS [25]–[27].

1) IXI: We consider T$_1$-weighted, T$_2$-weighted brain MR images from 53 healthy subjects. 25, 10 and 18 subjects were set aside for training, validation, and testing, respectively. 100 axial cross-sections containing brain tissues were chosen from each subject. The following were the acquisition parameters: $TE = 4.603ms$, $TR = 9.813ms$, $spatial\ resolution = 0.94 \times 0.94 \times 1.2mm^3$ in T$_1$-weighted images. $TE = 100ms$, $TR = 8178.34ms$, $spatial\ resolution = 0.94 \times 0.94 \times 1.2mm^3$ in T$_2$-weighted images. Prior to analysis, T$_2$ -weighted images were spatially registered onto T$_1$-weighted images. The registration was done using mutual information and via affine transformation in FSL [30].

2) BRATS: We consider T$_1$-weighted, T$_2$-weighted brain MR images from 55 patients. 25, 10, and 20 subjects were set aside for training, validation, and testing, respectively. 100 axial cross-sections containing brain tissues were chosen from each subject. Note that the BRATS collection [25]–[27] comprises scans obtained at many institutions using varied clinical methods and scanners [25]–[27]. Multi-contrast images are co-registered to the same anatomical template, interpolated to $1 \times 1 \times 1mm^3$ resolution, and skull-stripped.

E. Competing Methods

We test the proposed edaGAN model against various state-of-the-art image synthesis methods. Implementation details are as follows:

- **pGAN:** We consider pGAN with ResNet [8], [28] generator which is a convolutional GAN model [8]. pGAN [8] consists of generator and discriminator submodules with CNN backbones [2]. Its generator comprises an encoder, 9 residual blocks [28] and decoder. The encoder comprises a cascade of 3 strided convolutional layers, where the decoder consists of 3 transposed-convolutional layers.
- **pix2pix:** We consider pix2pix [16] model with U-Net [23] generator which is an another convolutional GAN model [8]. U-net generator [23] is an encoder-decoder architecture with skip connections between correspondent layer of encoder and decoder.
- **Attention U-Net:** A CNN-based U-Net architecture with additive attention gates is considered [22]. Here we adopt the original Attention U-Net [22] model as the generator of a conditional GAN model [15].

All models were trained in an adversarial setup. Each competing method's hyperparameters were tuned using the same cross-validation procedures. For a fair evaluation procedure, each method employed the same loss function, and patchGAN [16] discriminator.

F. Implementation Details

The encoder of edaGAN consists of three convolutional layers with kernel size $7, 3, 3$ in cascade. We used a total of nine residual blocks which comprise two convolutional with

TABLE I: Test results of synthesis models for $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_1$ tasks on IXI and BRATS datasets [25]–[27]. PSNR (dB) and SSIM (%) measurements are listed as mean ± std across test subjects. For a given task, boldface indicates the model with the best performance.

Task	**pGAN**	**pix2pix**	**Attention U-Net**	**edaGAN (ours)**
$T_1 \rightarrow T_2$	**PSNR**: 28.54 ± 1.94 **SSIM:** 0.925 ± 0.0305	**PSNR**: 26.96 ± 1.87 **SSIM:** 0.908 ± 0.037	**PSNR**: 27.01 ± 1.59 **SSIM:** 0.920 ± 0.032	**PSNR**: **29.17 ± 2.11** **SSIM: 0.933 ± 0.028**
$T_2 \rightarrow T_1$	**PSNR**: 28.28 ± 2.32 **SSIM:** 0.936 ± 0.030	**PSNR**: 27.35 ± 2.46 **SSIM:** 0.926 ± 0.035	**PSNR**: 24.73 ± 2.31 **SSIM:** 0.924 ± 0.037	**PSNR**: **28.72 ± 2.39** **SSIM: 0.940 ± 0.030**

(a) IXI dataset

Task	**pGAN**	**pix2pix**	**Attention U-Net**	**edaGAN (ours)**
$T_1 \rightarrow T_2$	**PSNR**: 25.95 ± 1.40 **SSIM:** 0.911 ± 0.026	**PSNR**: 25.83 ± 1.38 **SSIM:** 0.913 ± 0.025	**PSNR**: 23.48 ± 3.18 **SSIM:** 0.892 ± 0.042	**PSNR**: **26.38 ± 1.42** **SSIM: 0.915 ± 0.025**
$T_2 \rightarrow T_1$	**PSNR**: 25.36 ± 3.0 **SSIM:** 0.915 ± 0.026	**PSNR**: 25.60 ± 2.81 **SSIM:** 0.917 ± 0.024	**PSNR**: 24.78 ± 2.47 **SSIM:** 0.907 ± 0.028	**PSNR**: **25.77 ± 2.84** **SSIM: 0.918 ± 0.026**

(b) BRATS datasets [25]–[27]

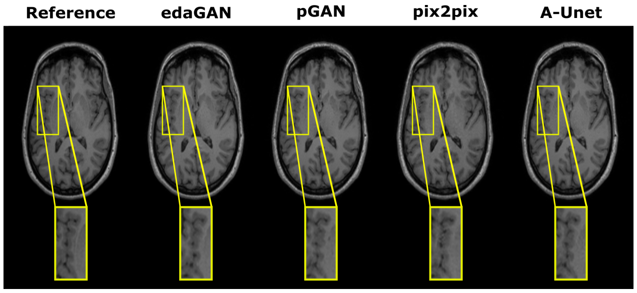

Fig. 3. T_1 synthesis.

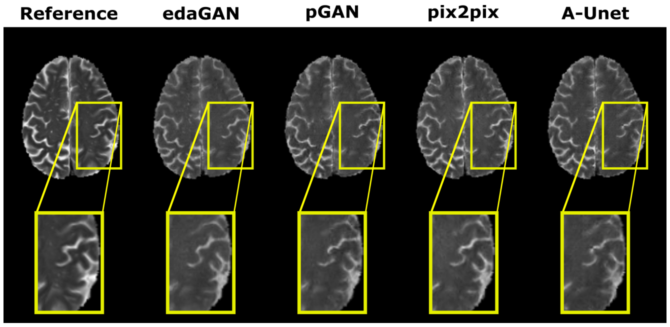

Fig. 4. T_2 synthesis.

kernel size 3 [28]. The decoder comprises three RED blocks and three transposed-convolutional layers with kernel size $3, 7, 7$ in cascade. All models were adversarially trained with the same PatchGAN discriminator [16]. Adversarial and pixel-wise losses were utilized in learning the synthesis models. All competing methods use the same optimization process and loss-term weightings for a fair comparison. Cross-validation is used to choose the learning rate schedules, the number of epochs, and loss-term weightings. In all approaches, selected parameters consistently deliver near-optimal results. For optimization, we used Adam optimizer [31] with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The models were trained for a total of 100 epochs. In the first 50 epochs, the learning rate was set to 0.0002, where in the last 50 epochs, it is linearly decayed to 0. We tuned λ_{L_1} to 100 via cross-validation experiments.

III. Results

We considered the brain images of healthy subjects in the IXI and patients in BRATS datasets [25]–[27]. We perform various experiments to show the performance of edaGAN in learning multi-contrast MR image synthesis. We compared edaGAN against the state-of-the-art synthesis models, pGAN [8], pix2pix [16], and [22]. We assessed the performance of the models in terms of Peak Signal-to-Noise Ratio(PSNR) and Structural Similarity Index Measure (SSIM) [32].

We first considered $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_1$ $(source \rightarrow target)$ tasks on IXI dataset. Test results in terms of metrics are reported in Table I a. In all tasks, edaGAN outperforms other competing methods in terms of PSNR and SSIM measurements ($p < 0.05$). On average across two tasks, we observe that edaGAN outperforms pGAN [8] by 0.56 dB PSNR and 0.010 % SSIM, pix2pix [16] by 1.80 dB PSNR and 0.020 % SSIM, and Attention U-net [22] by 3.08 dB PSNR and 0.015 % SSIM ($p < 0.05$). Representative images for all models for T_1 synthesis is given in Fig 3. Due to synergistic fusion of information captured by encoder and decoder feature maps, edaGAN yields superior depiction in regions that are depicted sub-optimally by baseline methods.

We then considered $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_1$ $(source \rightarrow target)$ tasks on BRATS datasets [25]–[27]. Test results in terms of metrics are reported in Table I b. Similarly, edaGAN outperforms other competing methods in all tasks in terms of PSNR and SSIM measurements ($p < 0.05$). On average across two tasks, we observe that edaGAN outperforms the pGAN [8] by 0.42 dB PSNR and 0.4 % SSIM, pix2pix [16] by 0.37 dB PSNR and 0.2 % SSIM, Attention U-net [22] by 1.95 dB PSNR and 1.8 % SSIM ($p < 0.05$). Representative images for all models for T_2 synthesis is given in Fig. 4. edaGAN offers preferable depiction of tissues in regions that are depicted poorly by baseline methods.

IV. Discussion and Conclusion

In this paper, we proposed a novel deep learning method for multi-contrast MRI synthesis. The proposed model employs RED blocks, which fuses the extracted latent maps with the low-level feature maps from the encoder via an additive attention gating mechanism. edaGAN synthesizes high-quality MRI images via learning global context in the input source contrast. We demonstrated the performance leap that edaGAN offers over the baseline synthesis models. We hypothesize that the introduced method has much potential for improving multi-contrast MRI synthesis both in practice and in clinical settings.

Acknowledgment

This study was supported in part by a TUBA GEBIP 2015 fellowship, a BAGEP 2017 fellowship, TUBITAK BIDEB 2020 grant, and by NVIDIA under a GPU grant.

References

[1] 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 *MICCAI*, pp. 631–638, Springer, 2013.

[2] B. Thukral, "Problems and preferences in pediatric imaging," *Indian Journal of Radiology and Imaging*, vol. 25, p. 359, 11 2015.

[3] B. E. Dewey, C. Zhao, J. C. Reinhold, A. Carass, K. C. Fitzgerald, E. S. Sotirchos, S. Saidha, J. Oh, D. L. Pham, P. A. Calabresi, P. C. van Zijl, and J. L. Prince, "Deepharmony: A deep learning approach to contrast harmonization across scanner changes," *Magnetic Resonance Imaging*, vol. 64, pp. 160–170, 2019. Artificial Intelligence in MRI.

[4] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and challenges in super-resolution," *International Journal of Imaging Systems and Technology*, vol. 14, pp. 47–57, 2004.

[5] C. Catana, A. van der Kouwe, T. Benner, C. J. Michel, M. Hamm, M. Fenchel, B. Fischl, B. Rosen, M. Schmand, and A. G. Sorensen, "Toward implementing an MRI-based PET attenuation-correction method for neurologic studies on the MR-PET brain prototype," *Journal of Nuclear Medicine*, vol. 51, no. 9, pp. 1431–1438, 2010.

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

[7] Y. Huang, L. Shao, and A. F. Frangi, "Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 815–827, 2018.

[8] S. U. 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. on Medical Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[9] 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 Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1750–1762, 2019.

[10] 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," *Medical Image Analysis*, vol. 70, p. 101944, 2021.

[11] 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," *arXiv:2011.14347*, 2021.

[12] M. Yurt, M. Özbey, S. U. H. Dar, B. Tınaz, K. K. Oğuz, and T. Çukur, "Progressively volumetrized deep generative models for data-efficient contextual learning of MR image recovery," *arXiv:2011.13913*, 2020.

[13] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *arXiv:2106.16031*, 2021.

[14] Y. Korkmaz, S. U. Dar, M. Yurt, M. Özbey, and T. Çukur, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *arXiv:2105.08059*, 2021.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Proceedings of NIPS*, vol. 24, 2014.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of CVPR*, pp. 1125–1134, 2017.

[17] A. Beers, J. Brown, K. Chang, J. Campbell, S. Ostmo, M. Chiang, and J. Kalpathy-Cramer, "High-resolution medical image synthesis using progressively grown generative adversarial networks," *arXiv:1805.03144*, 2018.

[18] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[19] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, and Q. Wang, "Medical image synthesis with deep convolutional adversarial networks," vol. 65, no. 12, pp. 2720–2730, 2018.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2021.

[21] D. A. Hudson and C. L. Zitnick, "Generative adversarial transformers," *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 2021.

[22] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," *arXiv:1804.03999*, 2018.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[24] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions," 2020.

[25] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, and et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[26] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Nature Scientific Data*, vol. 4, p. 170117, 2017.

[27] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, and et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv:1811.02629*, 2019.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of ICML* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 7354–7363, 2019.

[30] M. Jenkinson and S. Smith, "A global optimisation methof for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, pp. 143–156, 2001.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[32] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.