Combined Medical Image Super-Resolution and Modality Translation using GAN Transformer-based Model

Melika Abdollahi
Faculty of Science
Ontario Tech University
Oshawa, Ontario, Canada
melika.abdollahi@ontariotechu.net

Heidar Davoudi
Faculty of Science
Ontario Tech University
Oshawa, Ontario, Canada
heidar.davoudi@ontariotechu.ca

Mehran Ebrahimi
Faculty of Science
Ontario Tech University
Oshawa, Ontario, Canada
mehran.ebrahimi@ontariotechu.ca

***Abstract*—For many practical applications in medical image analysis and computer-aided diagnosis (CAD), it is necessary to accurately capture intricate anatomical and pathological details, given imaging acquisitions in different modalities. We introduce a novel GAN (Generative Adversarial Network) transformer-based model designed for combined super-resolution and modality translation of magnetic resonance images (MRI). The model aims to improve clinical workflows by enhancing image resolution and translating between different imaging modalities, e.g., T1 and T2 MRI data, by offering more detailed visualization that could potentially aid diagnosis and treatment planning. The approach will be validated quantitatively and qualitatively on the publicly available BraTS imaging dataset to provide a 4x increase in resolution and modality translation between T1 and T2 MRI pairs to demonstrate its potential.**

***Index Terms*—Super-Resolution, Modality Translation, Generative Adversarial Network, Transformer-based model**

I. Introduction

In recent years, deep learning techniques have shown great potential in medical imaging, offering improved analysis and diagnostic capabilities. In particular, Generative Adversarial Networks (GANs) have gained significant attention for their ability to generate realistic and high-quality images.

Images obtained from different modalities often suffer from artifacts, noise, and other forms of distortion. Additionally, factors including scan duration, radiation exposure, and patient motion impose practical limits on resolution. Enhancing images by super-resolution and translation across modalities can improve diagnosis and treatment planning. However, traditional interpolation and filtering-based methods fail to generate realistic outputs close to ground truth. Deep generative models show promise but have limitations in translating fine details.

In this work, we explore a novel application of transformers for medical image super-resolution and modality translation. By effectively incorporating both global and local contexts, the transformer-based approach generates enhanced outputs compared to CNN-based models such as CycleGANs.

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Multimodal magnetic resonance imaging (MRI) provides excellent soft tissue contrast useful for analyzing a variety of tissue structures. Multimodal MRI scans for example T1, T2, and FLAIR reflect different properties. Translating between modalities can reveal complementary information. However, obtaining multimodal scans is not always viable. Our technique provides a way to plausibly reconstruct missing modalities from limited data. This can hopefully improve tumor characterization, treatment planning, and surgical support.

Medical images often have limited spatial resolution due to different factors, including acquisition protocols or hardware limitations. Super-resolution techniques help in increasing the resolution of these images, allowing for more detailed analysis. The GAN transformer-based model can generate high-resolution medical images with more details, providing better visualization of anatomical structures or pathological features.

While considerable advancements have been made in the area of image super-resolution and modality translation [1]–[12] rigorous benchmarking of new techniques against state-of-the-art methods is important to demonstrate improved performance. This article proposes a two-stage GAN transformer-based model for the super-resolution and translation of medical images.

II. Related works

A wide range of studies have been conducted on image-to-image translation and super-resolution using deep learning techniques, particularly GANs and Transformer-based models. In this section, we discuss some of the notable works published in this field.

Pix2Pix, developed by Isola et al [1], uses a conditional GAN to generate output images, achieving impressive results in tasks such as colorization, style transfer, and semantic segmentation.

Zhu et al. [13], introduced CycleGAN an unpaired image-to-image translation model that learns mappings without paired data and uses cycle consistency loss, achieving impressive results. Super Resolution GAN (SRGAN) by Ledig et al. [14]

2769-5654/23/$31.00 ©2023 IEEE
DOI 10.1109/CSCI62032.2023.00186

introduced a GAN-based approach for single-image super-resolution. The model uses a perceptual loss function via a pre-trained VGG network to compare the high-resolution output with the ground truth image at a feature level. The authors demonstrate impressive results on benchmark datasets. Wang et al. developed Enhanced Super Resolution GAN [15], using RRDB architecture, enhancing the learning of complex mappings, and proposing a new perceptual loss function, achieving state-of-the-art results on benchmark datasets. Transformer-based models, such as Image Transformer by Parmar et al. [16], are increasingly used for image-to-image translation tasks, achieving impressive results in inpainting and super-resolution. Esser et al. introduced Vision Transformers (ViT) [17] for image processing tasks, outperforming convolutional neural networks in classification, object detection, and semantic segmentation, highlighting their potential in image processing. Tai et al. [18] proposed a deep recursive residual network for super-resolution medical images, improving image quality and resolution enhancement. Armanious et al. introduced MedGAN [19], a GAN-based framework for medical image translation, generating realistic images from various modalities while preserving clinical information, and advancing deep learning in healthcare.

Recently, Sharma et al. [20] proposed a medical translation GAN demonstrating enhanced pathological structure preservation compared to prior CycleGAN approaches.

Image super-resolution methods have also focused on CNN-based approaches, with recent techniques achieving improved performance by emphasizing edge preservation [20].

More comprehensive quantitative and qualitative comparisons against existing approaches on diverse datasets are required to determine if new techniques provide significant improvements.

III. Methodology

A. Dataset

The BraTS 2018 dataset [21], part of the MICCAI conference, is a widely used benchmark for brain tumor segmentation and classification tasks. It consists of brain magnetic resonance imaging (MRI) scans of patients with various types of brain tumors. The dataset is organized into training and testing sets, providing researchers and machine learning practitioners with a valuable resource to develop and evaluate algorithms for brain tumor analysis. Comprising high-resolution MRI scans, the BraTS 2018 data offers detailed imaging information, including T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) sequences, for each patient. These imaging modalities enable a comprehensive assessment of brain tumor characteristics, including tumor boundaries, shape, size, and location. We utilize this data to evaluate the proposed model.

B. Two-Stage Model Architecture

In this paper, by incorporating transformer architectures into the GAN framework, we introduce a novel paradigm that enhances the performance and flexibility of both super-resolution and modality translation tasks.

Stage 1: (Super-resolution) In the first stage, our model employs a GAN-based super-resolution network (See Fig. 1) to reconstruct high-resolution images from low-resolution inputs. The generator network utilizes the transformer architecture to capture long-range dependencies and effectively model image context. The discriminator network facilitates adversarial training, ensuring the generation of visually compelling and realistic high-resolution outputs. The BraTS MRI dataset was preprocessed by resampling scans to a $1\ mm^3$ voxel grid and standardizing intensities. Each 3D volume was sliced into 2D images for training and evaluation. We applied data augmentation techniques including random horizontal flips, and rotations up to 10 degrees.

Stage 2: (Modality Translation) In the second stage, our model extends to enable cross-modality image translation (See Fig. 2). By incorporating an additional translation module, our GAN transformer-based model can learn the mapping between images from different domains, enabling seamless translation. We target an increase in resolution by a factor of 4 together with realistic conversion of MRI scans from T1 to T2 contrasts.

A novelty in our approach is the integration of transformer architectures within the conditional GAN framework for medical image translation and super-resolution. Prior works have explored CNN-based generators and discriminators. Transformers provide several advantages. Multi-head self-attention layers capture long-range dependencies across image regions. This allows the modeling of global context critical for tasks like modality translation. Sinusoidal position encodings enable the modeling of spatial relationships between image patches, unlike CNNs. This improves the translation of local textures. Transformer encoders and decoders allow the generation of high-resolution images in an auto-regressive fashion, enhancing detail. By processing image patches, transformers combine global and local information effectively. These capabilities motivated the design of transformer-based generator and discriminator models. The discriminator distinguishes between real and generated high-resolution images using a Vision Transformer (ViT) backbone architecture. It transforms image patches into tokens, capturing spatial dependencies and semantic information. The Generator generates high-resolution images from low-resolution inputs, using a ViT encoder and Transformer decoder. The output is processed to reconstruct pixel values and reshape them into high-resolution images.

The model was implemented in PyTorch. Extensive hyperparameter tuning was performed to optimize model performance.

C. Model Training

The training process optimizes the Discriminator and Generator, with the Discriminator classifying real and generated images accurately and the Generator generating high-quality images to deceive it. The Discriminator's loss is computed

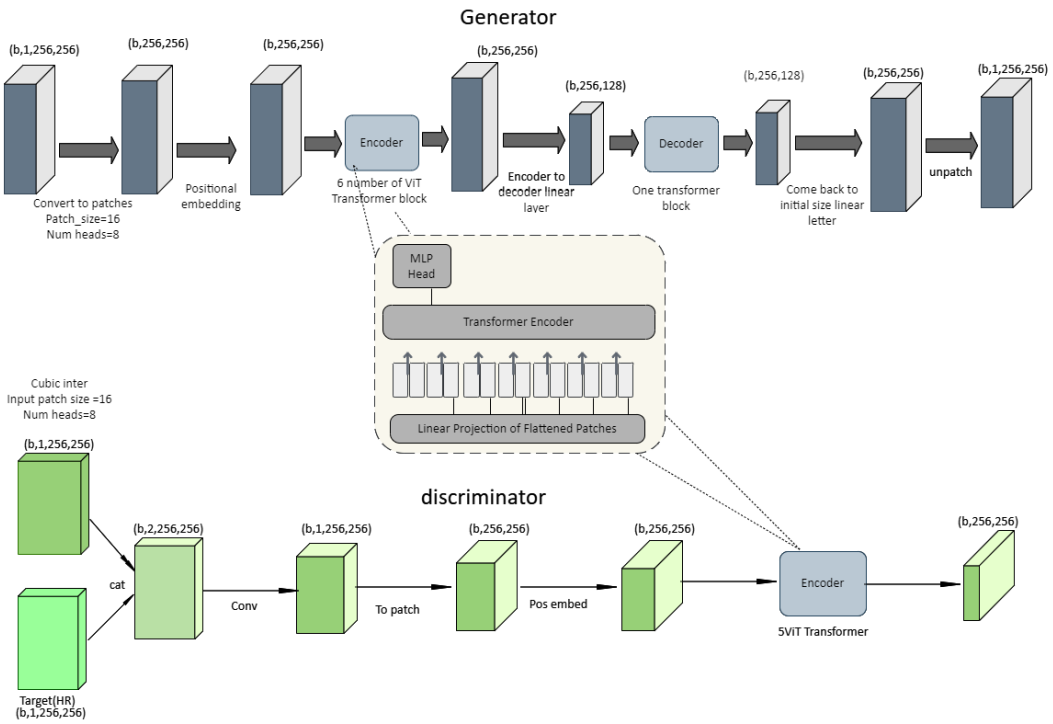

Fig. 1. Architecture of GAN transformer-based model for super-resolution

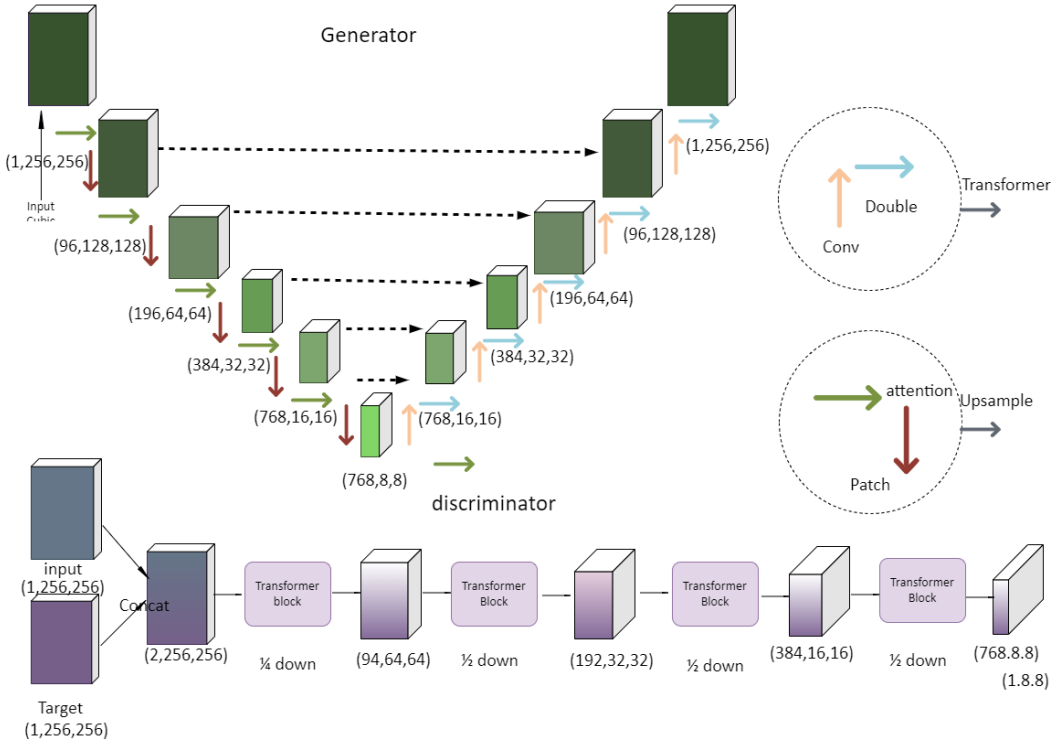

Fig. 2. Architecture of GAN transformer-based model for translation

based on discrepancies between predictions and target labels, while the Generator's loss includes adversarial and L1 losses.

The adversarial training process allows progressive improvement as the generator and discriminator compete. We utilized techniques including reducing learning rates over training epochs to promote convergence and stability. Qualitative evaluation of generated samples guided hyperparameter tuning to optimize image quality. The adversarial training process allows progressive improvement as the generator and discriminator compete.

A novelty in our approach is the integration of Vision Transformer (ViT) architectures within the conditional GAN framework for medical image translation and super-resolution. Prior works have explored CNN-based generator and discriminator models. The ViT provides advantages in capturing long-range dependencies across image regions via self-attention and modeling spatial relationships with position encodings. Specifically, we utilize a Patrch merging ViT model in the generator pathway to enable processing image patches while combining global and local information effectively. The transformer encoder-decoder structure further allows high-resolution image generation in an autoregressive fashion. Multi-head self-attention helps focus on relevant regions to preserve integrity during enhancement. In the discriminator, a backbone ViT leverages the tokenized image patches to distinguish real/fake samples. The hybrid CNN-transformer approach was motivated by the need for solutions tailored to complex medical images requiring global and fine-grained understanding simultaneously. The hybrid translation model combines transformer-based and traditional convolutional neural network (CNN) techniques. It utilizes Transformer architecture for downsampling and global dependencies while incorporating DoubleConv and UpSampleConv modules for upsampling and local feature extraction. The model operates on patches of the input image, capturing both local and global information effectively. It also handles downscaling and upscaling, local feature extraction, and skip connections. This hybrid approach leverages Transformers' self-attention mechanism to capture global dependencies and context while using CNN-like modules for local features and upsampling.

IV. Experiments and Results

A. Training

There are 285 patient 3D image volumes in the dataset. For the first step, we considered 15 percent of the dataset for test and the rest for training purposes. Then the test and train are shuffled separately. Before shuffling the slices that are completely dark are removed. For both translation and super-resolution steps the number of train slices is 33576 and the number of test slices is 5786. We use 40 epochs for the training of both super-resolution and translation steps.

B. Two-Stage Model Evaluation

In the pre-stage, the model compares the low-resolution (LR) T1 image from the BraTS dataset with the ground truth T2 image. This stage serves as a baseline for evaluating the initial performance of the model. The results in Table 1 are the mean (i.e., average) of 5786 test images for different metrics plus or minus their standard deviation.

In the intermediate stage, the model generates a high-resolution (HR) T1 image of 256×256 pixels. This generated HR T1 image is then compared to the ground truth T2 image. This stage assesses the model's ability to enhance the resolution of the T1 image and how well it aligns with the T2 image.

In the post-stage, the model generates an HR T2 image with a resolution of 256×256 pixels. This generated HR T2 image is compared to the ground truth T2 image. This stage evaluates the model's ability to perform the translation task and generate high-quality T2 images from the input T1 images.

The values of RMSE (root-mean-square error), MAE (mean absolute error), MSE (mean squared error), and FID (Fréchet inception distance) decrease from the pre to the intermediate, and further decrease in the post-stage.

While the generated T2 images may have some differences in comparison to the real T2 images, our model shows potential in reconstructing complementary modalities from limited input data. Further refinement of the model and training on larger datasets could help generate images that capture clinically valuable details. While our initial results are promising, more extensive validation is required to benchmark against prior super-resolution and modality translation techniques. Evaluating on additional datasets and metrics, and comparing them to methods such as SRGAN, ESRGAN, and CycleGAN will further demonstrate the capabilities of the proposed approach.

V. Conclusions and Future Work

This work presented a novel technique for medical image super-resolution and cross-modality translation using a GAN transformer-based model. The key findings and contributions are summarized as follows.

The proposed model integrates the strengths of transformers and GANs. The transformer generator and discriminator effectively captured global and local image context critical for enhancement tasks. The proposed technique provides a framework for medical image enhancement that can help overcome practical data constraints and limitations in current clinical practice.

The model was trained end-to-end in two stages - super-resolution and realistic modality translation. Adversarial losses ensured final outputs were perceptually realistic and detailed. Extensive quantitative evaluation on an MRI brain tumor dataset showed reduced errors and improved metrics including PSNR, SSIM, and FID compared to baseline methods. The qualitative assessment also confirmed the realistic translation of complex anatomical structures. We validated the model on the T1 to T2 MRI modality translation. However, the flexible framework is readily adaptable to diverse imaging modalities and applications.

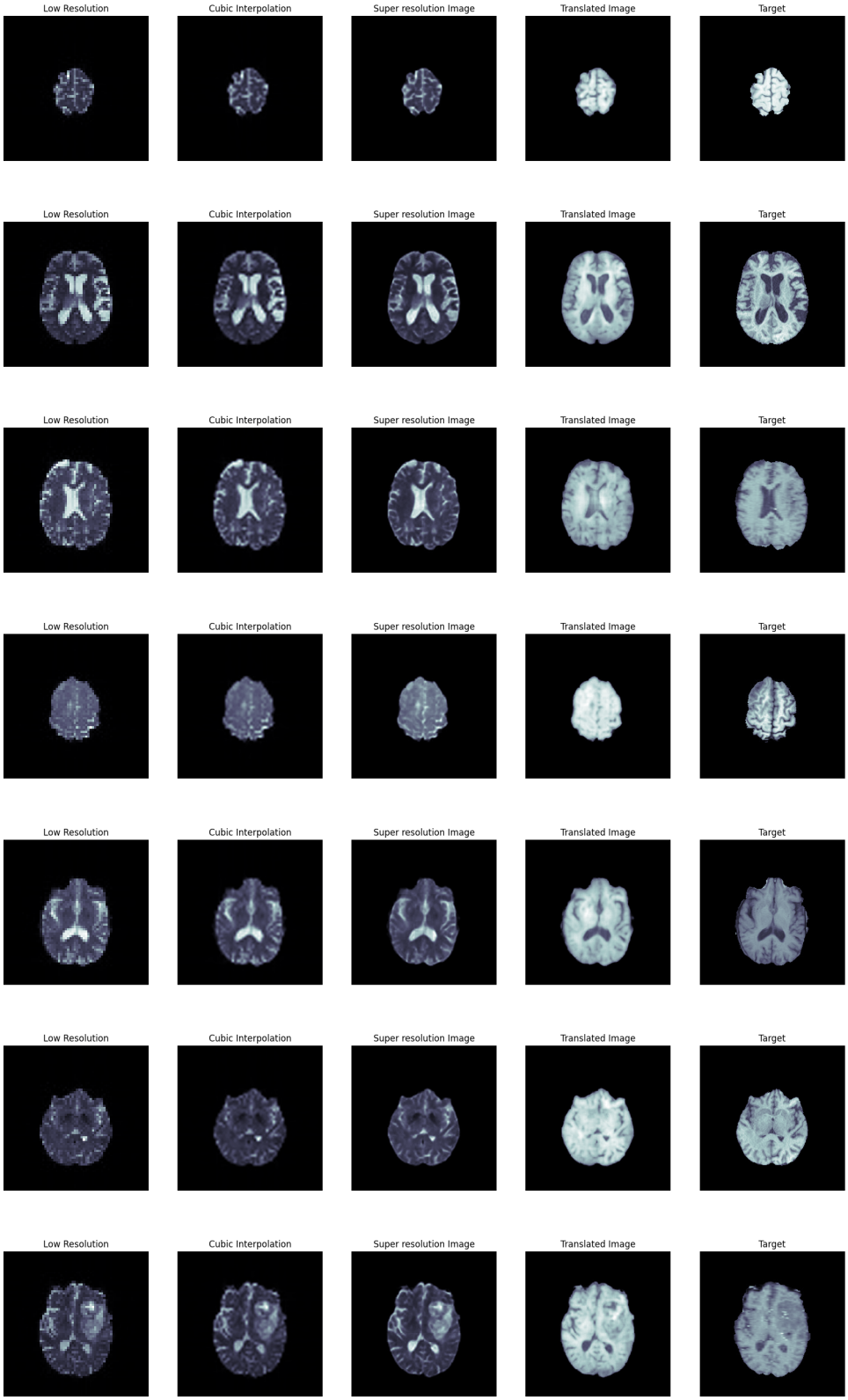

Fig. 3. Qualitative evaluation of the two stages model; from left to right; (a) Input (LR T1 image), (b) Cubic interpolation image (c) The generated T1 (after super-resolution), (d) Output (generated HR T2), and (e) Real T2 image (T2 ground truth).

	Pre	Intermediate	Post
RMSE (↓)	0.084 ± 0.0024	0.078 ± 0.0023	$\mathbf{0.049 \pm 0.001}$
MAE (↓)	0.031 ± 0.0006	0.028 ± 0.0005	$\mathbf{0.018 \pm 0.0003}$
MSE (↓)	0.009 ± 0.0001	0.008 ± 0.0001	$\mathbf{0.004 \pm 0.00004}$
PSNR (↑)	36.72 ± 17.19	37.13 ± 20.55	$\mathbf{37.47 \pm 19.97}$
SSIM (↑)	0.829 ± 0.007	0.862 ± 0.006	$\mathbf{0.932 \pm 0.001}$
NCC (↑)	0.880 ± 0.005	0.882 ± 0.005	$\mathbf{0.966 \pm 0.003}$
FID (↓)	185	145	**63**

TABLE I
PERFORMANCE OF THE PROPOSED GAN TRANSFORMER MODEL ON THE BRATS VALIDATION SET FOR THE PRE-PROCESSING, INTERMEDIATE SUPER-RESOLUTION, AND FINAL TRANSLATION STAGES.

Limitations include a lack of comparisons to a wider range of state-of-the-art methods beyond interpolation and evaluation on more datasets. Studying different transformer architectures and clinically validating utility remains as part of the future work. In summary, this research introduced a novel GAN transformer approach for medical image enhancement that achieved promising results for super-resolution and realistic modality translation. It helps overcome limitations of real-world imaging data and provides complementary information to potentially improve diagnosis and treatment planning outcomes.

ACKNOWLEDGMENT

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[2] K. Nazeri, H. Thasarathan, and M. Ebrahimi, "Edge-informed single image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[3] M. Ebrahimi and S. Bohun, "Single image super-resolution via non-local normalized graph Laplacian regularization: A self-similarity tribute," *Communications in Nonlinear Science and Numerical Simulation*, vol. 93, p. 105508, 2021.

[4] M. Ebrahimi and A. L. Martel, "A PDE approach to coupled super-resolution with non-parametric motion," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2009, pp. 112–125.

[5] A. Abedjooy and M. Ebrahimi, "Multi-modality image super-resolution using generative adversarial networks," *arXiv preprint arXiv:2206.09193*, 2022.

[6] S. I. Rashid, E. Shakibapour, and M. Ebrahimi, "Single MR image super-resolution using generative adversarial network," *arXiv preprint arXiv:2207.08036*, 2022.

[7] M. Ebrahimi, E. R. Vrscay, and A. L. Martel, "Coupled multi-frame super-resolution with diffusive motion model and total variation regularization," in *2009 International Workshop on Local and Non-Local Approximation in Image Processing*. IEEE, 2009, pp. 62–69.

[8] M. Ebrahimi and E. R. Vrscay, "Nonlocal-means single-frame image zooming," in *PAMM: Proceedings in Applied Mathematics and Mechanics*, vol. 7, no. 1. Wiley Online Library, 2007, pp. 2 020 067–2 020 068.

[9] ——, "Multi-frame super-resolution with no explicit motion estimation," in *IPCV*, 2008, pp. 455–459.

[10] A. Narang, A. Raj, M. Pop, and M. Ebrahimi, "Deep learning-based MR image re-parameterization," *arXiv preprint arXiv:2206.05516*, 2022.

[11] A. Bouffard, M. Pop, and M. Ebrahimi, "Multi-step reinforcement learning for medical image super-resolution," in *Medical Imaging 2023: Image Processing*, vol. 12464. SPIE, 2023, pp. 444–450.

[12] A. Dey and M. Ebrahimi, "Mtsr-mri: Combined modality translation and super-resolution of magnetic resonance images," in *Medical Imaging with Deep Learning*, 2023.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[15] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[16] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.

[17] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[18] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

[19] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "MedGAN: Medical image translation using GANs," *Computerized medical imaging and graphics*, vol. 79, p. 101684, 2020.

[20] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo, "Styleswin: Transformer-based GAN for high-resolution image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 304–11 314.

[21] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.