

# VIGU: VISION GNN U-NET FOR FAST MRI

Jiahao Huang<sup>1,2,\*</sup>, Angelica I. Aviles-Rivero<sup>3</sup>, Carola-Bibiane Schönlieb<sup>3</sup>, Guang Yang<sup>1,2,\*</sup>

<sup>1</sup> National Heart and Lung Institute, Imperial College London, United Kingdom

<sup>2</sup> Cardiovascular Research Centre, Royal Brompton Hospital, United Kingdom

<sup>3</sup> Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom

\* Send correspondence to {j.huang21,g.yang}@imperial.ac.uk

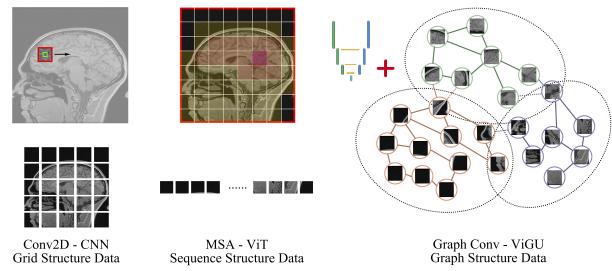
## ABSTRACT

Deep learning models have been widely applied for fast MRI. The majority of existing deep learning models, e.g., convolutional neural networks, work on data with Euclidean or regular grids structures. However, high-dimensional features extracted from MR data could be encapsulated in non-Euclidean manifolds. This disparity between the go-to assumption of existing models and data requirements limits the flexibility to capture irregular anatomical features in MR data. In this work, we introduce a novel Vision GNN type network for fast MRI called Vision GNN U-Net (ViGU). More precisely, the pixel array is first embedded into patches and then converted into a graph. Secondly, a U-shape network is developed using several graph blocks in symmetrical encoder and decoder paths. Moreover, we show that the proposed ViGU can also benefit from Generative Adversarial Networks yielding to its variant ViGU-GAN. We demonstrate, through numerical and visual experiments, that the proposed ViGU and GAN variant outperform existing CNN and GAN-based methods. Moreover, we show that the proposed network readily competes with approaches based on Transformers while requiring a fraction of the computational cost. More importantly, the graph structure of the network reveals how the network extracts features from MR images, providing intuitive explainability.

**Index Terms**— Fast MRI, Graph Neural Network (GNN)

## 1. INTRODUCTION

Magnetic Resonance Imaging (MRI) is one of the most important clinical tools. It provides high-resolution and non-invasive imaging for diagnosis and prognosis in a harmless manner. However, MRI has an inherently slow scanning time, since the raw data is acquired in  $k$ -space, and the minimum scanning time is decided by the selection of temporal and spatial resolution as well as the field of view, constraining by the Nyquist theorem. The prolonged scanning time leads to artefacts from the voluntary and involuntary physiological



**Fig. 1.** Data Structures of Convolutional Neural Networks, Vision Transformers and Vision GNN U-Net.

movements of the patients [1].

With the thriving development of artificial intelligence technologies, deep learning-based models have been promptly developed for fast MRI [2, 3]. Convolutional neural networks (CNNs) have dominated research studies in computer vision (CV) and medical image analysis, including MRI reconstruction [3, 4, 5], taking advantage of the inductive biases of locality and weight sharing, and their hierarchical structures. Recently, Transformers [6] have shown superiority for CV tasks bolstered by their global sensitivity and long-range dependency. Transformer-based MRI reconstruction methods [7, 8, 9, 10] have been proposed and achieved promising results, even though their increased computational cost is still a challenge for a wider application.

General CNNs and Transformers backbones treat image data differently (Figure 1). The 2D convolution (Conv2D) in CNNs applies sliding operation kernel on pixels in a regular grid, exploiting the shift-invariance and local prior. The multi-head self-attention (MSA) in Transformers (specifically ViT [6]), embeds different ranges of pixels into patches, then converts them into sequences, introducing global sensitivity and long-range dependency. However, both Conv2D and MSA operations are usually based on the regular pixel grid in the Euclidean space [11].

Recently, Han et al. in [11] proposed the Vision GNN (ViG) backbone. ViG combines, combining the patch em-

bedding from ViT [6] and the idea of Graph Convolutional Networks (GCNs) [12], treating images with more flexibility from the graph perspective. GCNs are originally designed for tackling specific tasks for non-Euclidean data, e.g., point cloud, social network, and biochemical graphs. Vision GNN fills the technological gap between GNNs and image data for computer vision tasks, and achieves state-of-the-art results in high-level tasks like classification and detection tasks.

For MR images, the shape of anatomical structures are irregular, leading to redundancy and inflexibility when using the conventional grid or sequence data structure. We hypothesise that treat MR images as graphs (Figure 1) can provide a comprehensive understanding of the anatomical structures in MR images. Specifically, the image is first converted into patches by a shallow CNN and then regarded as nodes in a graph. Nodes with similar features can be gathered and connected using the K-nearest neighbours (KNN) algorithm, where information exchange can be conducted. Different anatomical structures can be recognised as sub-graphs of the whole graph (for an image). The edge connections within and between sub-graphs can be learnt to reflect the intra- and inter-relationship of anatomical structures.

In this paper, we exploit how ViG works for a specific low-level image restoration task, i.e., MR reconstruction, by introducing a ViG-based U-Net, namely ViGU, and its variants based on Generative Adversarial Network (GAN), namely ViGU-GAN. Experiments have shown that our proposed ViGU and ViGU-GAN can outperform CNN-based and GAN-based MRI reconstruction methods and can achieve comparable results with Transformer-based methods with much a lower computational cost. The edge connection of ViGU shows that the proposed ViGU can learn the intra- and inter-relationship of different anatomical structures, providing model explainability.

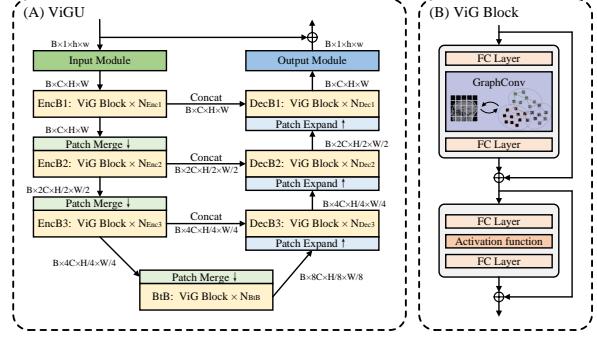
## 2. METHODS

This section describes in detail the key parts of the proposed ViGU network and variant.

### 2.1. U-Net Based Architecture

The architecture of the proposed ViGU is displayed in Fig. 2 (A). CNN-based input and output modules are applied at the beginning and end of our ViGU converting between images  $\mathbb{R}^{h \times w \times 1}$  and patch vectors  $\mathbb{R}^{N \times C}$ . We denote  $r$  and  $C$  as the patch size and embedding channel number respectively. We then define the number of patches as  $N = H \times W = h/r \times w/r$ . Relative position embedding is applied for each patch, which is omitted for brevity.

Three encoder blocks (EncB) and three decoder blocks (DecB) are symmetrically arranged in the encoder and decoder path correspondingly, between which a bottleneck block (BnB) is placed. The EncB, DecB and BnB are com-



**Fig. 2.** (A) The network architecture of our ViGU; (B) The structure of the ViG Blocks.

posed of one or multiple ViG Blocks, which are the basic computation blocks for ViGU. The resolution of feature maps is gradually decreased and increased along the encoder and decoder paths. Information is passed, via the skip connection and concatenation operation, from the encoder to the decoder paths between feature maps with the same resolution. Residual connection is applied to convert the ViGU into a refinement function:  $\hat{x}_u = H_{ViGU}(x_u) + x_u$ .

### 2.2. Graph-level Operation

A key step is how to transform an image as a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  composed of a set of nodes  $\mathcal{V}$  connected by a set of edges  $\mathcal{E}$ . For each feature map, we have a group of patches  $X = \{x_1, x_2, \dots, x_N\}$ , which are viewed as a set of unordered nodes  $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ . For a single node  $v_i$ ,  $K$  edges  $\mathcal{E}_i = \{e_{1i}, e_{2i}, \dots, e_{Ki}\}$  are acquired from its  $K$  nearest neighbours  $\mathcal{N}(v_i)$ , where  $e_{ji}$  indicates the edge from node  $v_j$  to node  $v_i$ .

The graph representation of feature map  $X$  can be expressed as  $\mathcal{G}(X)$ . A graph convolution operation  $H_{GConv}$  is expressed as:

$$\begin{aligned} \mathcal{G}' &= H_{GConv}(\mathcal{G}(X), W) \\ &= H_{Update}(H_{Aggregate}(\mathcal{G}(X), W_{Aggregate}), W_{Update}), \end{aligned} \quad (1)$$

in which  $H_{Aggregate}$  and  $H_{Update}$  refer to the Aggregate and Update operations in graph convolution with learnable parameters  $W_{Aggregate}$  and  $W_{Update}$  [11].

### 2.3. ViG Block

As Figure 2 (B) shows, ViG Block adopted the structure from ViT Block [6], which can be expressed as:

$$X' = FC(GraphConv(FC(X))) + X \quad (2)$$

$$X'' = MLP(X) + X', \quad (3)$$

where  $X$  and  $X''$  are the input and output of ViG Block.  $GraphConv(\cdot)$  and  $MLP(\cdot)$  denote the graph convolution and

the multi-layer perceptron.  $\text{FC}(\cdot)$  denotes the full connected layer, which is applied before and after the graph convolution; with the purpose to keep the domain consistency between node and image features and increase the feature diversity. All the normalisation and activation functions are omitted for brevity.

#### 2.4. Optimisation Scheme

To train our proposed ViGU, Charbonnier loss is applied to the image and frequency domains, which are denoted as  $\mathcal{L}_{\text{img}}(\theta)$  and  $\mathcal{L}_{\text{freq}}(\theta)$  respectively. They allow for constraining the ground truth MR images  $x$  and reconstructed MR images  $\hat{x}_u$ . Moreover, a  $l1$  loss is applied for perceptual-based,  $\mathcal{L}_{\text{perc}}(\theta)$ , constraints using a pre-trained VGG  $f_{\text{VGG}}(\cdot)$ . Formally, they read:

$$\min_{\theta} \mathcal{L}_{\text{img}}(\theta) = \sqrt{\|x - \hat{x}_u\|_2^2 + \epsilon^2}, \quad (4)$$

$$\min_{\theta} \mathcal{L}_{\text{freq}}(\theta) = \sqrt{\|\mathcal{F}x - \mathcal{F}\hat{x}_u\|_2^2 + \epsilon^2}, \quad (5)$$

$$\min_{\theta} \mathcal{L}_{\text{perc}}(\theta) = \|f_{\text{VGG}}(x) - f_{\text{VGG}}(\hat{x}_u)\|_1, \quad (6)$$

where  $\epsilon$  is empirically set to  $10^{-9}$ . We denote  $\theta$  as the network parameter of ViGU, and  $\mathcal{F}$  refers to the Fourier transform. The total loss of ViGU,  $\mathcal{L}_{\text{ViGU}}(\theta)$ , using is computed as:

$$\mathcal{L}_{\text{ViGU}}(\theta) = \alpha \mathcal{L}_{\text{img}}(\theta) + \beta \mathcal{L}_{\text{freq}}(\theta) + \gamma \mathcal{L}_{\text{perc}}(\theta), \quad (7)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are weighting parameters balancing the importance of each term.

Our ViGU can also benefit from GAN principles yielding to a new variant called ViGU-GAN. For the GAN-based variant, the proposed ViGU is the generator  $G_{\theta_G}$  parameterised by  $\theta_G$  (same with the  $\theta$  in ViGU), and a U-Net based discriminator [13],  $D_{\theta_D}$ , is applied for adversarial training. The adversarial loss  $\mathcal{L}_{\text{adv}}(\theta_G, \theta_D)$  is then given by:

$$\begin{aligned} & \min_{\theta_G} \max_{\theta_D} \mathcal{L}(\theta_G, \theta_D) \\ &= \mathbb{E}_{x \sim p_t(x)} [\log D_{\theta_D}(x)] - \mathbb{E}_{x_u \sim p_u(x_u)} [\log D_{\theta_D}(\hat{x}_u)]. \end{aligned} \quad (8)$$

The total loss of ViGU-GAN,  $\mathcal{L}_{\text{ViGU-GAN}}(\theta)$ , reads:

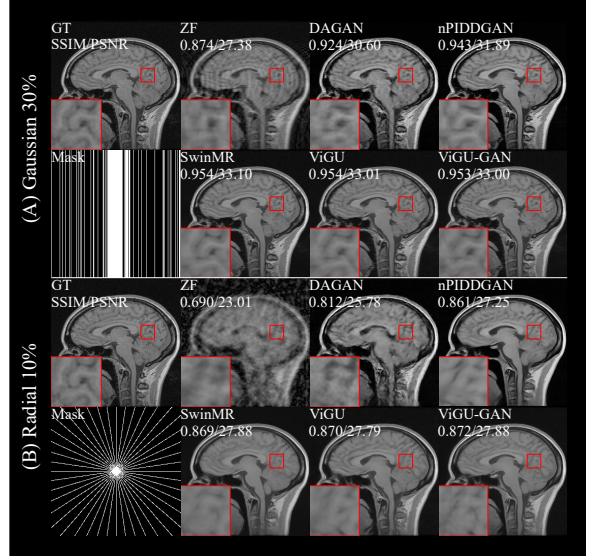
$$\mathcal{L}_{\text{ViGU-GAN}}(\theta_G, \theta_D) = \mathcal{L}_{\text{ViGU}}(\theta_G) + \mathcal{L}(\theta_G, \theta_D). \quad (9)$$

### 3. EXPERIMENTAL SETTINGS AND RESULTS

This section describes in detail the set of experiments conducted to validate the proposed ViGU and variant.

#### 3.1. Implementation Details

We evaluate our approach using the Calgary-Campinas Public Dataset [14]. It is composed of 67 cases of T1-weight 3D brains, and randomly divided into training, validation and



**Fig. 3.** Visual comparison of our ViGU/ViGU-GAN vs existing techniques. Results display SSIM and PSNR results.

testing datasets following a ratio of 6:1:3. The multi-channel data was converted into single-channel MR images using the root sum square method. The top and bottom slices in each case were discarded, and the rest of the 100 slices were chosen for experiments.

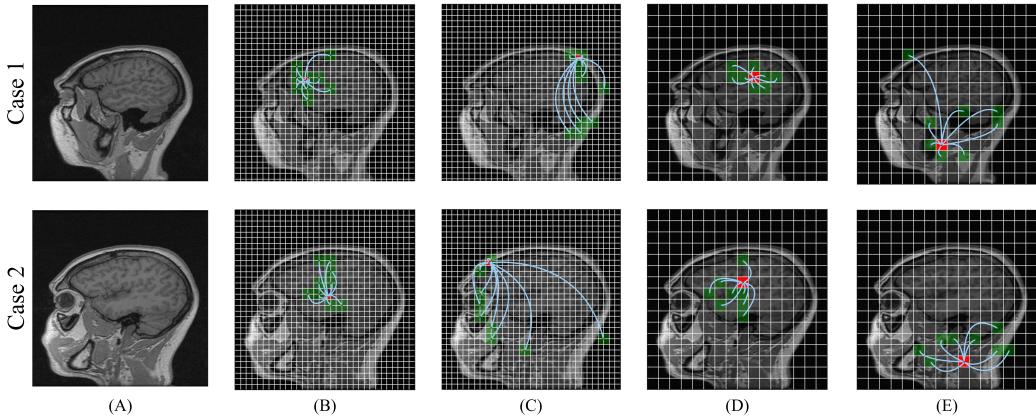
The number of ViG Blocks and embedding channels was set to [3, 3, 3, 1, 3, 3, 3] and [96, 192, 384, 768, 384, 192, 96] respectively.  $\text{ViGU}_x$  indicated the proposed ViGU with a patch size of  $x$ . The initial learning rate was set to  $6 \times 10^{-4}$  and decays every 10,000 steps by 0.5 from the 50,000<sup>th</sup> step. The weighting parameters in the loss function  $\alpha$ ,  $\beta$  and  $\gamma$  were set to 15, 0.1 and 0.0025. For training the ViGU-GAN, the parameter of the discriminator is updated every 5 steps, to prevent training an “overly strong” discriminator and compromising the training of the generator.

We compared the proposed ViGU and ViGU-GAN against MRI reconstruction methods of DAGAN [4], nPIDD-GAN [5] and SwinMR [8] with Gaussian 1D 30% (G1D30%) and radial 10% (R10%) masks.

For quantitative results, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Fréchet Inception Distance (FID) [15]. Multiply Accumulate Operations (MACs) were utilised to estimate the computational complexity with an input size of  $1 \times 256 \times 256$ .

#### 3.2. Comparison Experiments

Table 1 and Figure 3 show the quantitative results and visualised samples of the comparison experiments, respectively. The proposed ViGU and ViGU-GAN outperformed other CNN and GAN-based methods, and achieved comparable results compared to the Transformer-based method SwinMR,



**Fig. 4.** Visualised graphs of the proposed ViGU. (A): the original MR images; (B-C): Graph connection from EnvB2; (D-E): Graph connection from EnvB3. A chosen node (red) and its first-order neighbours (green) are connected by edges (green line). In (B-C),  $2 \times 2$  maximum pooling operation was applied for the neighbour nodes to reduce the computational cost.

**Table 1.** Quantitative results of the comparison experiments.

Method	MACs		G1D30%			R10%		
	(G) ↓	SSIM ↑	PSNR ↑	FID ↓	SSIM ↑	PSNR ↑	FID ↓	
ZF	-	0.883 (0.012)	27.81 (0.82)	156.38	0.706 (0.022)	23.53 (0.82)	319.45	
DAGAN	33.97	0.924 (0.010)	30.41 (0.82)	56.04	0.822 (0.024)	25.95 (0.85)	132.58	
nPIDD-GAN	56.44	0.943 (0.009)	31.81 (0.92)	26.15	0.864 (0.023)	27.17 (0.97)	82.86	
SwinMR	800.73	<b>0.955 (0.009)</b>	<b>33.05 (1.09)</b>	<b>21.03</b>	<b>0.876 (0.022)</b>	<b>27.86 (1.02)</b>	59.0	
ViGU <sub>4</sub>	15.07	0.954 (0.009)	32.85 (1.05)	26.06	0.868 (0.025)	27.60 (1.03)	63.43	
ViGU <sub>4</sub> -GAN	15.07	0.949 (0.009)	32.41 (1.02)	22.44	0.841 (0.024)	26.86 (0.91)	87.03	
ViGU <sub>2</sub>	73.02	<b>0.955 (0.009)</b>	<b>32.95 (1.07)</b>	<b>22.73</b>	0.872 (0.023)	27.72 (1.02)	<b>58.61</b>	
ViGU <sub>2</sub> -GAN	73.02	0.954 (0.009)	32.88 (1.07)	<b>16.62</b>	0.873 (0.022)	27.75 (1.00)	<b>50.19</b>	

with only 1.9% and 9.1% MACs depending on the patch size.

For the patch size setting, ViGU and ViGU-GAN with small patch sizes (larger patch resolution) tend to have better reconstructed results, whereas at the cost of larger MACs.

For the GAN-based variant ViGU-GAN, the utilisation of adversarial training mainly improves the perceptual experiments and reflects a better FID score. However, the proposed ViGU-GAN leads to an unstable training process (abnormal pool performance of ViGU<sub>4</sub>-GAN using R10% mask in Table 1), prolonged convergence time and enlarged GPU memory requirements. Further research and optimisation of GAN-based ViGU should be conducted.

### 3.3. Visualised Graph & Explainability

Figure 4 shows the visualised graph connection of the proposed ViGU, including reference MR images (A), and graph connection from EnvB2 (B-C) and EnvB3 (D-E). For better visualisation, we only display a chosen node (red) and its first-order neighbours (green) connected by an edge (green line). In Figure 4 (B-C),  $2 \times 2$  maximum pooling operation was applied for the neighbour nodes to reduce the computational cost, which led that the neighbour node area being bigger than the chosen node area.

The graph connection of the proposed ViGU model can

provide an explainability of how the network recognises and extracts the feature of MR images. Figures 4 (B) and (D) show that a node of brain tissue tends to have more neighbour nodes containing brain tissue, which proves that the network can be trained to gather the node with similar features and create the connection between them. However, since there is no tag information added to the network, it is hard for the proposed ViGU to learn the accurate border of different without any supervision. Different anatomical structures with similar textures can also mislead the network. A node at the edge (border of the anatomical structures, not the edge in the graph) tends to have a neighbour node that is also at the edge, regardless of the anatomical structures (Figure 4 (C) and (E)).

## 4. DISCUSSION

This work has exploited how ViG works for MRI reconstruction, treating the MR images as graphs instead of conventional grid or sequence structure data. Using graph-based operation our proposed network can extract and process the feature more flexibly and efficiently since the irregular anatomical structures leads to redundancy and inflexibility using regular grid-based or sequence-based operations like CNN and transformers. In addition, the proposed ViGU can learn a comprehensive understanding of the feature of MR images in latent non-Euclidean space, gathering and linking different parts with similar features globally.

In conclusion, we can envisage that our proposed ViGU and ViGU-GAN to be served as a UNet-based backbone for the graph-based MRI reconstruction, super-resolution and segmentation. For future work, segmentation information would be incorporated into the ViGU, guiding the network to build clinically-meaningful graphs, and improving the reconstruction performance while providing better explainability.

## 5. ACKNOWLEDGEMENTS

This study was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC\NSFC\211235), the NVIDIA Academic Hardware Grant Program, and the UKRI Future Leaders Fellowship (MR/V023799/1). CBS acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute.

## 6. REFERENCES

- [1] Yutong Chen, Carola-Bibiane Schönlieb, Pietro Liò, Tim Leiner, Pier Luigi Dragotti, Ge Wang, Daniel Rueckert, David Firmin, and Guang Yang, “AI-based reconstruction for fast MRI-a systematic review and meta-analysis,” *Proceedings of the IEEE*, vol. 110, no. 2, pp. 224–245, 2022.
- [2] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu, “Deep ADMM-Net for compressive sensing MRI,” in *Advances in Neural Information Processing Systems*. 2016, vol. 29, Curran Associates, Inc.
- [3] Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony N. Price, and Daniel Rueckert, “A deep cascade of convolutional neural networks for dynamic MR image reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 491–503, 2 2018.
- [4] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, and David Firmin, “DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction,” *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1310–1321, 6 2018.
- [5] Jiahao Huang, Weiping Ding, Jun Lv, Jingwen Yang, Hao Dong, Javier Del Ser, Jun Xia, Tiaojuan Ren, Stephen Wong, and Guang Yang, “Edge-enhanced dual discriminator generative adversarial network for fast MRI with parallel imaging using multi-view information,” *Applied Intelligence*, 2021.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv e-prints*, p. arXiv:2010.11929, Oct. 2020.
- [7] Yilmaz Korkmaz, Mahmut Yurt, Salman Ul Hassan Dar, Muzaffer Özbeý, and Tolga Cukur, “Deep MRI reconstruction with generative vision transformers,” in *Machine Learning for Medical Image Reconstruction*, Cham, 2021, pp. 54–64, Springer International Publishing.
- [8] Jiahao Huang, Yingying Fang, Yinze Wu, Huanjun Wu, Zhifan Gao, Yang Li, Javier Del Ser, Jun Xia, and Guang Yang, “Swin transformer for fast MRI,” *Neurocomputing*, vol. 493, pp. 281–304, 2022.
- [9] Jiahao Huang, Yinze Wu, Huanjun Wu, and Guang Yang, “Fast MRI reconstruction: How powerful transformers are?,” *arXiv e-prints*, p. arXiv:2201.09400, Jan. 2022.
- [10] Jiahao Huang, Xiaodan Xing, Zhifan Gao, and Guang Yang, “Swin deformable attention U-Net transformer (SDAUT) for explainable fast MRI,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, Eds., Cham, 2022, pp. 538–548, Springer Nature Switzerland.
- [11] Kai Han, Yunhe Wang, Jianyu Guo, Yehui Tang, and Enhua Wu, “Vision GNN: An image is worth graph of nodes,” *arXiv e-prints*, p. arXiv:2206.00272, June 2022.
- [12] Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem, “DeepGCNs: Can GCNs go as deep as CNNs?,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [13] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva, “A U-Net based discriminator for generative adversarial networks,” in *CVPR*, June 2020.
- [14] Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo, “An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement,” *NeuroImage*, vol. 170, pp. 482–494, 2018.
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.