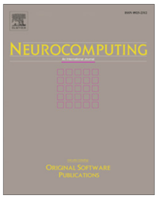
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Swin transformer for fast MRI

Jiahao Huang [a,b,*], Yingying Fang [a], Yinzhe Wu [a,c], Huanjun Wu [a,c], Zhifan Gao [d], Yang Li [e], Javier Del Ser [f,g], Jun Xia [h], Guang Yang [a,b,*]

[a] *National Heart and Lung Institute, Imperial College London, London SW7 2AZ, United Kingdom*
[b] *Cardiovascular Research Centre, Royal Brompton Hospital, London SW3 6NP, United Kingdom*
[c] *Department of Bioengineering, Imperial College London, London SW7 2AZ, United Kingdom*
[d] *School of Biomedical Engineering, Sun Yat-sen University, Guangzhou 510275, China*
[e] *School of Automation Sciences and Electrical Engineering, Beihang University, Beijing 100190, China*
[f] *Department of Communications Engineering, University of the Basque Country UPV/EHU, Bilbao 48013, Spain*
[g] *TECNALIA, Basque Research and Technology Alliance (BRTA), Derio 48160, Spain*
[h] *Department of Radiology, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University Health Science Center, Shenzhen 518037, China*

ARTICLE INFO

Article history:
Received 10 January 2022
Revised 15 March 2022
Accepted 3 April 2022
Available online 12 April 2022

Keywords:
MRI reconstruction
Transformer
Compressed sensing
Parallel imaging

ABSTRACT

Magnetic resonance imaging (MRI) is an important non-invasive clinical tool that can produce high-resolution and reproducible images. However, a long scanning time is required for high-quality MR images, which leads to exhaustion and discomfort of patients, inducing more artefacts due to voluntary movements of the patients and involuntary physiological movements. To accelerate the scanning process, methods by *k*-space undersampling and deep learning based reconstruction have been popularised. This work introduced SwinMR, a novel Swin transformer based method for fast MRI reconstruction. The whole network consisted of an input module (IM), a feature extraction module (FEM) and an output module (OM). The IM and OM were 2D convolutional layers and the FEM was composed of a cascaded of residual Swin transformer blocks (RSTBs) and 2D convolutional layers. The RSTB consisted of a series of Swin transformer layers (STLs). The shifted windows multi-head self-attention (W-MSA/SW-MSA) of STL was performed in shifted windows rather than the multi-head self-attention (MSA) of the original transformer in the whole image space. A novel multi-channel loss was proposed by using the sensitivity maps, which was proved to reserve more textures and details. We performed a series of comparative studies and ablation studies in the Calgary-Campinas public brain MR dataset and conducted a downstream segmentation experiment in the Multi-modal Brain Tumour Segmentation Challenge 2017 dataset. The results demonstrate our SwinMR achieved high-quality reconstruction compared with other benchmark methods, and it shows great robustness with different undersampling masks, under noise interruption and on different datasets. The code is publicly available at https://github.com/ayanglab/SwinMR.

1. Introduction

Magnetic resonance imaging (MRI) is an important non-invasive imaging technique, which enables excellent assessments of structural and functional conditions with no radiation in a reproducible manner. Basically, MRI is aimed to reconstruct the images from the observed signals whose degradation process can be formulated as follows:

$$y = \mathscr{F}x + n, \tag{1}$$

where $x, y \in \mathbb{C}^N$ are the vectors denoting the latent image to reconstruct in the image domain and the observed measurements in k-space, $\mathscr{F} \in \mathbb{C}^{N \times N}$ is the two-dimensional (2D) discrete Fourier transform (DFT) and n is the noise inevitably appearing in the signal acquisition process.

However, acquiring the full measurements of y to construct a high-quality MR image x is highly time-consuming. Moreover, the long scanning time will bring about the artefacts arising from the voluntary movements of the patients and involuntary physiological movements [1]. In order to mitigate the long acquisition time of MRI as well as alleviate the aliasing artefacts, a range of methods has been developed for accelerating MRI to obtain accurate reconstructions. Traditionally, gradient refocusing [2] and multiple-radio frequency mediated [3] approaches were proposed.

[*] Corresponding author at: National Heart and Lung Institute, Imperial College London, London SW7 2AZ, United Kingdom.

E-mail addresses: j.huang21@imperial.ac.uk (J. Huang), g.yang@imperial.ac.uk (G. Yang).

https://doi.org/10.1016/j.neucom.2022.04.051
0925-2312/© 2022 The Author(s). Published by Elsevier B.V.

Under constraints of the Nyquist-Shannon sampling theorem, they did reduce the scanning time although by only a limited factor. With the development of the parallel imaging (PI) and the compressed sensing (CS), the fast MRI based on these two theories attracted much research and advancements.

Parallel imaging was introduced to take advantage of spatial sensitivity distribution derived from an array of carefully distributed receiver surface coils, to reduce the measurement from each coil, alleviating the need of enhancing gradient performance and hence reducing the acquisition time [4]. The undersampled k-space signal using PI-MRI can be represented by a general model as:

$$y^q = \mathscr{F}_u(\mathscr{S}^q \otimes x) + n^q, \quad q = 1, \ldots S, \tag{2}$$

where $\mathscr{S}^q$ and n^q are the sensitivity map and inevitable noise of the q^{th} coil (S coils in total). $\otimes$ denotes the pixel-wise multiplication. $\mathscr{F}_u \in \mathbb{C}^{M\times N}$ is the undersampled 2D DFT matrix with $M \ll N$ to reduce the measurements of each y^q. With S coils applied parallelly, one can obtain $y^1, \ldots, y^S$ simultaneously to reconstruct the latent image x. To reconstruct these PI acquired images, great progress in developing PI reconstruction techniques has taken place, proposing popular methods such as the simultaneous acquisition of spatial harmonic (SMASH) [5], sensitivity encoding (SENSE) [6] and generalized auto-calibrating partially parallel acquisition (GRAPPA) [7].

The invention of CS theory [8] further advanced the sampling efficiency of MRI. The CS-MRI utilises the non-linear methodology and sparse transformation to reconstruct the latent image from only a small portion of k-space measurement under a much smaller downsampling rate than the Nyquist rate. The general problem of MRI using the CS-MRI is to find the minimiser image to the following problem:

$$\arg\min_x ||\Phi x||_1, \quad \text{s.t.} \ \ y = \mathscr{F}_u x, \tag{3}$$

where Φ is the sparsifying transformation, $\mathscr{F}_u \in \mathbb{C}^{M\times N}$ is undersampled 2D DFT with $M \ll N$, and $y \in \mathbb{C}^M$ is the observed undersampled measurements in k-space. A range of non-linear reconstruction methods has demonstrated success in resolving this, including some fixed sparsifying methods such as total variation [9], curvelets [10] and double-density complex wavelet [11], and a few adaptive sparsifying models taking the advantage of dictionary learning [12]. While both CS-MRI and PI-MRI can significantly reduce the required number of measurements in k-space, the iterative algorithms are required to derive the image however prolong the recon-

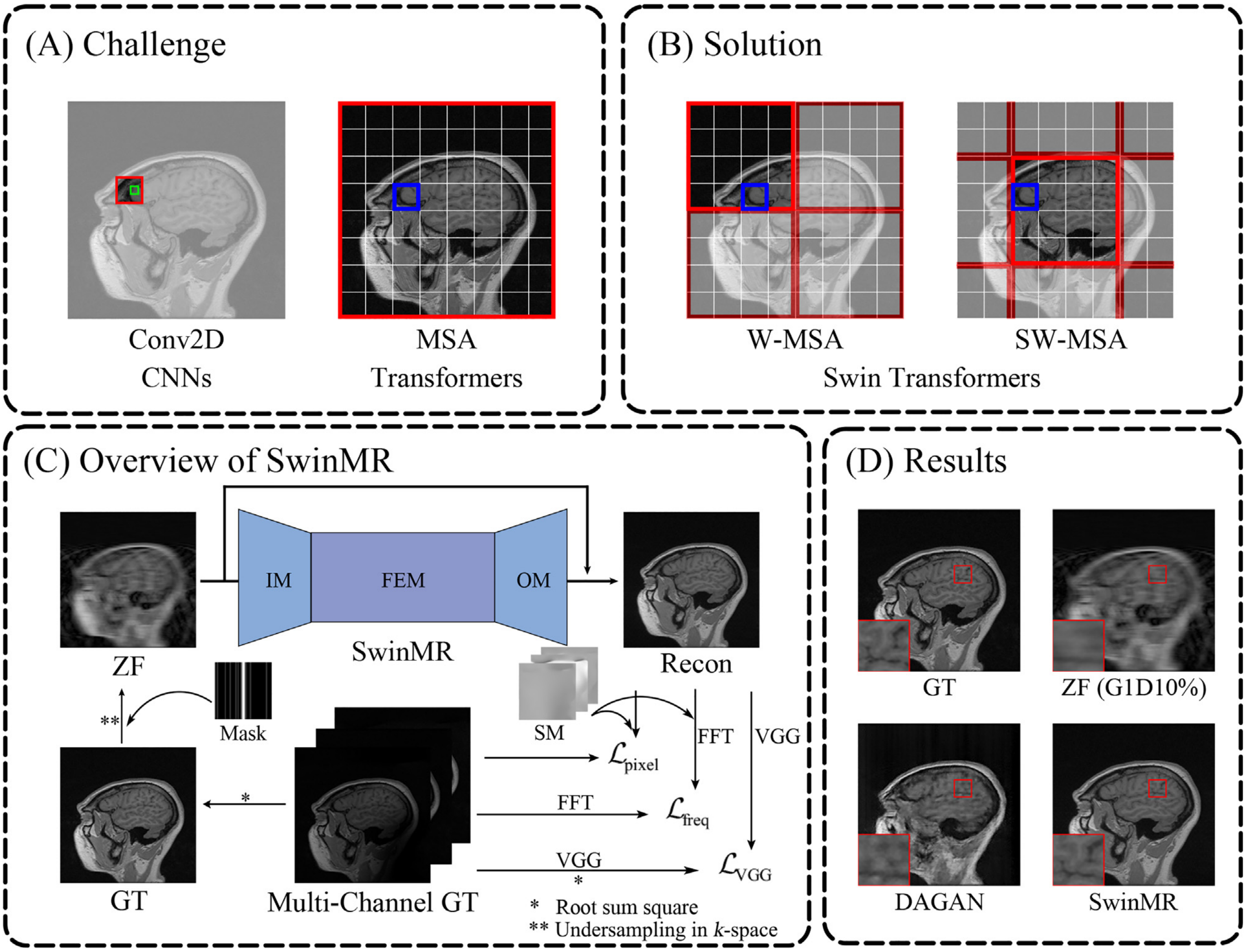


Fig. 1. Overview of the proposed SwinMR. (A) and (B) are the schematic diagrams of the receptive field for 2D convolution (Conv2D), multi-head self-attention (MSA) and shifted windows based multi-head self-attention (W-MSA/SW-MSA). Conv2D is locally sensitive and lacks long-range dependency. Compared with Conv2D, MSA and (S)W-MSA have larger receptive fields. MSA is performed in the whole image space, while W-MSA and SW-MSA are alternatingly used in Swin transformer [36], and performed in shifted windows. (Red box: the receptive field of the operation; green box: the pixel; blue box: the patch in self-attention.) (C) is the overview of SwinMR. (D) shows the results of the proposed SwinMR compared with GT, ZF and another method DAGAN [37]. (IM: the input module; FEM: the feature extraction module; OM: the output module. ZF: undersampled zero-filled MR images; Recon: reconstructed MR images; Multi-Channel GT: multi-channel ground truth MR images; GT: single-channel ground truth MR images; Mask: the undersampling mask; SM: sensitivity maps.).

struction time and hence cause concerns when transferred for actual clinical uses.

As a modern popular method for general image analysis, deep learning has been very successful by exploiting the non-linear and complex natures of the network with supervised or unsupervised learning, and widely applied in medical image research [13–17]. Convolutional neural networks (CNNs) as a special type of deep learning networks enable enhanced latent feature extraction by their very deep hierarchical structure. CNN has demonstrated its superiority in multiple tasks, including detection [18], classification [19], segmentation [20] and super-resolution [21]. Wang et al. [22] became the pioneer to take advantage of CNNs by extracting latent correlations between undersampled and fully sampled *k*-space data for MRI reconstruction. Yang et al. [23] further improved the network structures by re-applying the alternating direction method of multipliers (ADMM), which was originally used for CS-MRI reconstruction methods. A cascaded structure was developed by Schlemper et al. [24] for the more targeted reconstruction of dynamic sequences in cardiac MRI. To enable further latent mapping in the reconstruction model, Zhu et al. [25] developed a novel framework to provide more dense mapping through domains via its proposed automated transform by manifold approximation.

For a long time, CNNs have had a dominant position in the field of computer vision (CV) since convolutions are effective feature extractors. Most deep learning-based MRI reconstruction methods are based on CNNs, including the GAN-based model. As Fig. 1(A) shows, the feature extraction of CNNs is based on convolution, which is locally sensitive and lacks long-range dependency. The receptive field of CNNs is limited by the convolutional kernel and the network depth. Oversized convolutional kernel brings huge computational cost, and overly-deep network depth can cause gradient vanishing.

A novel structure, transformer, taking advantage of even deeper mapping, sequence-to-sequence model design [26] and adaptive self-attention setting [27–30] with expanding receptive fields (Fig. 1(A)) [31,32] has been proposed recently and been popularised in natural language processing (NLP) initially [33]. Then it has been applied to object detection [34], image recognition [35] and extended to super-resolution [31] for general image analysis.

With its superior ability in image reconstruction and synthesis as demonstrated in natural images, we could see transformers applied in MRI in many different ways. For synthesis, it has greatly enhanced cross-modality image synthesis (PET-to-MR by directional encoder [38], T1-to-T2 by a pyramid structure [39], and MR-to-CT and T1/T2/PD by a novel aggregated residual transformer block [40]). Variants of the transformer also enabled improved performance in reconstruction and super-resolution tasks. It was first applied on the reconstruction of brain MR imaging [41]. Korkmaz et al. [42] developed an unsupervised adversarial method to alleviate the scarce training sample populations. To further improve the quality of imaging, Feng et al. [43] enabled an end-to-end joint reconstruction and super-resolution. Feng et al. [44] further advanced the model for these dual tasks by incorporating the model with task-specific novel cross-attention modules.

However, the shift from NLP tasks to CV tasks leads to challenges: (1) Difference in scale: visual elements (e.g., pixels) in CV tasks tend to vary substantially in scale unlike language elements (e.g., word tokens) in NLP tasks. (2) Higher resolution: the resolution of pixels in images (or frames) tend to be much higher than words in sentences. [36] Therefore, it is a trade-off for less computational complexity to limit the scale of self-attention in a local window, as Fig. 1(A) and (B) shows. **S**hifted **win**dows (Swin) transformer [36] replaced the traditional multi-head self-attention (MSA) by the shifted windows based multi-head self-attention (W-MSA/SW-MSA). W-MSA and SW-MSA were alternatingly used in consecutive transformer layers, since if all attention operations are conducted in fixed windows, the cross-window relationship may be ignored. Based on the Swin transformer module, Liang et al. [45] proposed SwinIR for image restoration tasks.

In this work, we introduced the SwinMR, a novel parallel imaging coupled Swin transformer based model for fast CS-MRI reconstruction, as Fig. 1(C) shows. The main contributions can be summarised as follows:

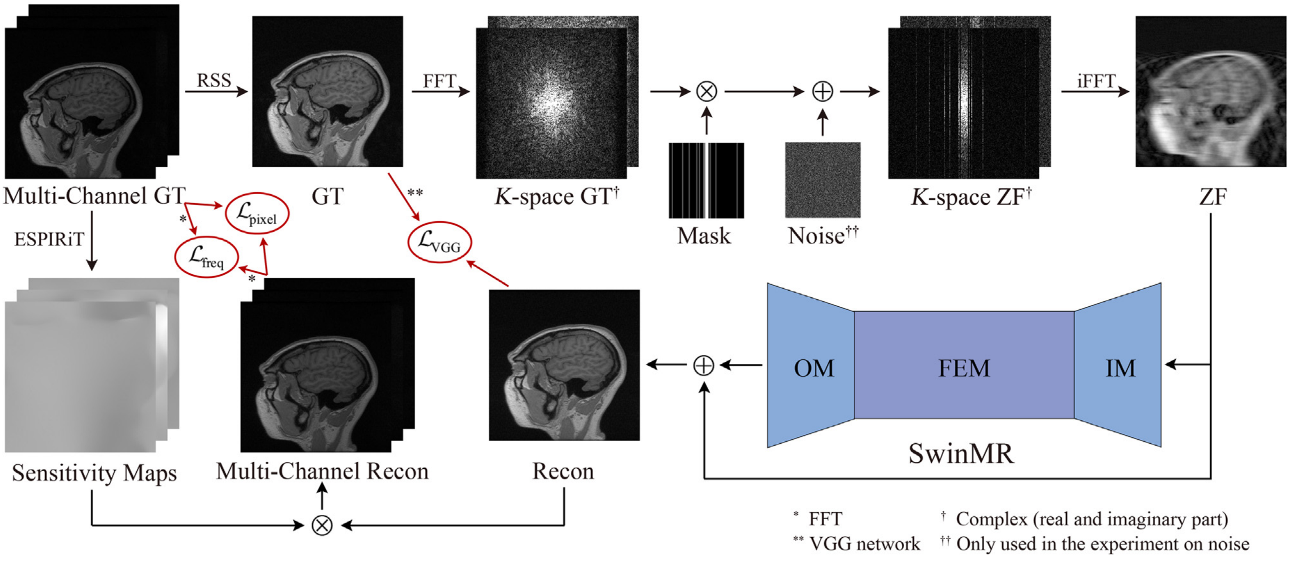


Fig. 2. The dataflow of proposed SwinMR. Root sum square (RSS) is applied to combine the multi-channel ground truth MR images (Muti-Channel GT) to single-channel ground truth MR images (GT). Undersampling and noise interruption are performed in *k*-space using fast Fourier transform (FFT) and inverse fast Fourier transform (iFFT) to convert the GT to undersampled zero-filled MR images (ZF) as the input of our proposed SwinMR. Multi-channel reconstructed MR images (Muti-Channel Recon) are calculated by the pixel-wise multiplication of single-channel reconstructed MR images (Recon), which are the output of the proposed SwinMR, and sensitivity maps, which are estimated by ESPIRiT from the Multi-Channel GT.

- A novel parallel imaging coupled Swin transformer-based model for fast MRI reconstruction was proposed, as Fig. 1(C) shows.
- A novel multi-channel loss was proposed by using the sensitivity maps, which was proved to preserve more textures and details in the reconstruction results.
- A series of ablation studies and comparison experiments were conducted. Experimental studies using different undersampling trajectories with various noises were performed to validate the robustness of our proposed SwinMR.
- A downstream task experiment using a segmentation network was conducted. A pre-trained segmentation network was applied to test the segmentation score for reconstructed images.

2. Method

2.1. Classic model-based CS-MRI reconstruction

To recover better spatial information with less artefacts from the undersampled k-space data, traditional CS-MRI methods usually consider solving the following optimisation problem:

$$\min_{x} \frac{1}{2}||\mathscr{F}_u x - y||_2^2 + \lambda R(\Phi x), \tag{4}$$

where Φ is the sparsifying transform, e.g., discrete wavelet transform [46], gradient operator [9,47] and dictionary-based transform [48]. $R(\cdot)$ is the regularisation function imposed on the sparsity, e.g, l_1-norm and l_0-norm, and λ is the weight parameter to balance the

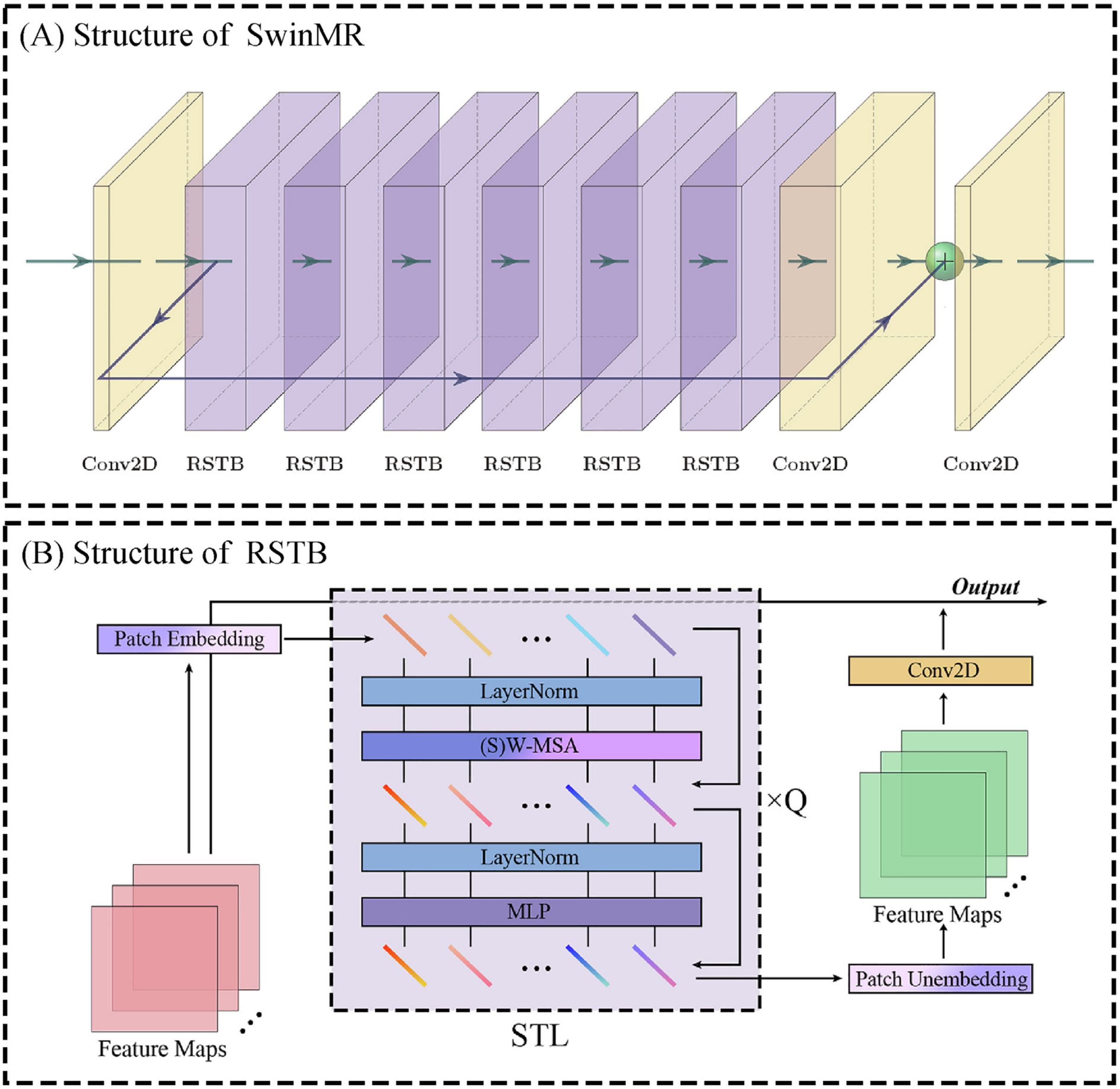


Fig. 3. The structure of proposed SwinMR. (A) shows the overall structure of SwinMR. In SwinMR architecture, two Conv2Ds are placed at the beginning and the ending. A cascade of RSTBs and a Conv2D with a residual connection are placed between the two Conv2Ds. (B) shows the structure of RSTB. The RSTB consists of a patch embedding operator, Q cascaded STLs, a patch unembedding operator, a Conv2D, and a residual connection between the input and output of RSTB. An STL consists of an LN, an (S)W-MSA, an LN and an MLP, with two residual connections. (RSTB: the residual Swin transformer block; STL: the Swin transformer layer; Conv2D: the 2D convolutional layer; LN: the layer normalisation layer; MLP: the multi-layer perceptron; (S)W-MSA: the (shifted) windows multi-head self-attention. W-MSA and SW-MSA are altinatively used in consecutive STLs.)

two terms. The solution of the above problem can be derived by the non-linear optimisation solvers such as gradient-based algorithms [49] and variable splitting methods [50,51]. Depending on the manually designed regularisation, some models may suffer from a long reconstruction time for better reconstruction quality. Additionally, the manually selected sparsifying transform Φ could also introduce undesirable artefacts, e.g., total variation based regularisation which is well-known for removing the noise and preserving the sharp edges can introduce staircase artefacts [10] and the tight wavelet frame transform increases the reconstruction efficiency but may lead to the blocky artefacts [52].

2.2. CNN-based fast MRI reconstruction

To relieve the artefacts brought by the hand-crafting regularisation and the long reconstruction time of classic models, the deep CNNs which are well-known as the powerful features extractors, were firstly applied in the CS-MRI in [22]. In this work, a deep CNN was applied to learn the mapping from down-sampled reconstruction images to fully sampled reconstruction images directly. Following that, several networks have been proposed to further improve the reconstruction quality.

Some works attempted to bridge the classic models with deep CNNs by mimicking the iterative algorithm in their network architectures. Deep ADMM Net [23] was firstly trained by unfolding the optimisation algorithm ADMM to derive the solution to the general model Eq. (4) by network blocks. In [24], the reconstruction of the deep CNN from lower-quality images was adopted as the prior information to approximate in a classic CS-model as follows:

$$\min_{x} \frac{1}{2}||y - \mathscr{F}_u x||_2^2 + \lambda||x - f_{\mathrm{CNN}}(x_u|\theta)||_2^2, \tag{5}$$

where the solution of the above function was further adopted into the network architecture iteratively to improve the reconstruction result of f_{CNN} which takes the zero-filled reconstruction x_u as the input.

On top of the CNNs, conditional generative adversarial networks (cGANs) exploited the advantages of deep learning further and proved to enhance the quality of the MR image reconstruction to a large extent [53,54]. Such a competitive network introduced a two-player generator-discriminator training mechanism to competitively improve reconstruction performance by alternatingly optimising θ_G and θ_D of the generator G and the discriminator D, in a general form as:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x\sim p_{\mathrm{gt}}}\left[\log D_{\theta_D}(x)\right] + \mathbb{E}_{x_u\sim p_u}\left[\log\left(1 - D_{\theta_D}\left(G_{\theta_G}(x_u)\right)\right)\right], \tag{6}$$

where G_{θ_G} and D_{θ_D} denote the generator and the discriminator with parameters θ_G and θ_D, respectively. x and x_u denote the ground truth MR images and undersampled zero-filled MR images with aliasing

Table 1
Quantitative results of the comparison experiment with other methods using Gaussian 1D 30% mask (mean (std)). †: $p < 0.05$; ††: $p < 0.01$ (compared with SwinMR (PI) by paired t-Test). ‡: $p < 0.05$; ‡‡: $p < 0.01$ (compared with SwinMR (nPI) by paired t-Test). *: #PARAMs for only the generator/for both the generator and discriminator. PSNR: Peak signal-to-noise ratio; SSIM: Structural similarity index; FID: Fréchet inception distance; Inference Time: The average time for one inference in an Intel Core i9-10980XE CPU or an NVIDIA RTX 3090 GPU; #PARAMs: The parameters number of models; MACs: Multiply-Accumulate Operations.

Methods	PSNR	SSIM	FID	Inference Time		#PARAMs	MACs
				CPU (s)	GPU (s)	(M)	(G)
ZF	27.81 (0.83)††‡‡	0.884 (0.012)††‡‡	156.39	–	–	–	–
Deep ADMM Net	29.24 (0.99)††‡‡	0.922 (0.012)††‡‡	54.56	0.459 (0.052)	–	–	–
U-Net	31.48 (0.86)††‡‡	0.939 (0.009)††‡‡	46.90	0.166 (0.007)	0.006 (0.000)	32.31	56.44
DAGAN	30.41 (0.83)††‡‡	0.924 (0.010)††‡‡	56.05	0.089 (0.003)	0.003 (0.000)	98.59/127.18*	33.97
PIDDGAN	31.23 (0.93)††‡‡	0.936 (0.010)††‡‡	17.55	0.166 (0.007)	0.006 (0.000)	32.31/89.50*	56.44
SwinMR (nPI)	**33.06 (1.09)**††	**0.956 (0.009)**††	21.03	19.310 (0.115)	0.041 (0.001)	11.40	800.73
SwinMR (PI)	32.07 (1.02)‡‡	0.945 (0.010)‡‡	**8.70**	19.310 (0.115)	0.041 (0.001)	11.40	800.73

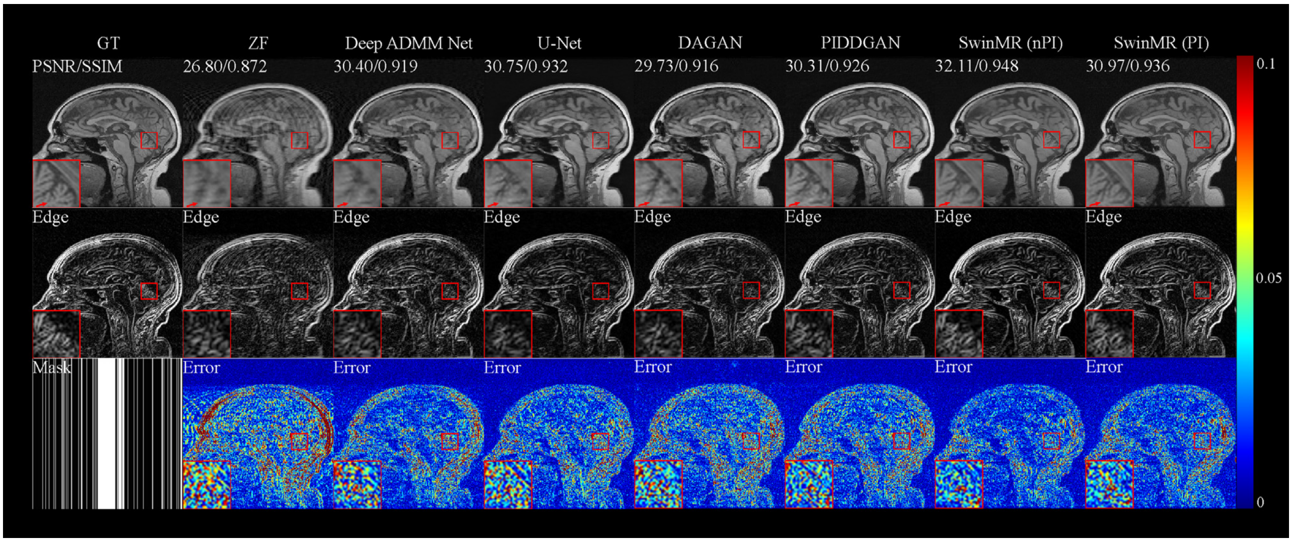


Fig. 4. Samples of the comparison experiment with ground truth images (GT), undersampled zero-filled images (ZF) and reconstructed images by other methods. Row 1: GT, ZF and reconstructed images by different methods; Row 2: Edge information extracted by Sobel operator; Row 3: Gaussian 1D 30% mask and the absolute differences between reconstructed (or ZF) images and GT images (10×).

artefacts. After the training, the generator can yield the corresponding reconstruction from x_u to reconstructed images $G_{\theta_G}(x_u)$.

Variants of generators and discriminators have been developed to cope with multiple flaws in the original architecture of GAN – for improved generator [55], improved discriminator [56], loss functions [57], regularisation [58], training stability by Wasserstein GAN [59,60] and attention mechanism [61]. DAGAN [37], by substituting the residual networks with a modified U-Net [62], combined the advantage of U-Net in latent information extraction with competitive training and pre-trained VGG based transfer learning. Furthermore, PIDDGAN [56] considered edge information into their model and further enhance the edge information in the reconstruction, which are clinically important when interpreting MR images. The utilisation of transfer learning improved the generalisability of a network trained with a small dataset [63].

Table 2
Fréchet inception distance (FID) of the experiment on different masks. Five undersampling masks including Gaussian 1D 10% (G1D10%), Gaussian 1D 30% (G1D30%), Gaussian 1D 50% (G1D50%), radial 10% (R10%) and spiral 10% (S10%) were applied in this experiment.

Mask	SwinMR (PI)	DAGAN	ZF
G1D10%	**28.27**	169.83	326.00
G1D30%	**8.70**	56.05	156.38
G1D50%	**5.11**	19.26	86.25
R10%	**34.19**	132.58	319.45
S10%	**28.97**	115.98	333.40

CNN-based MR reconstruction methods showed their superiority both on reconstruction quality and efficiency compared to classical MR reconstruction methods. However, the performance of

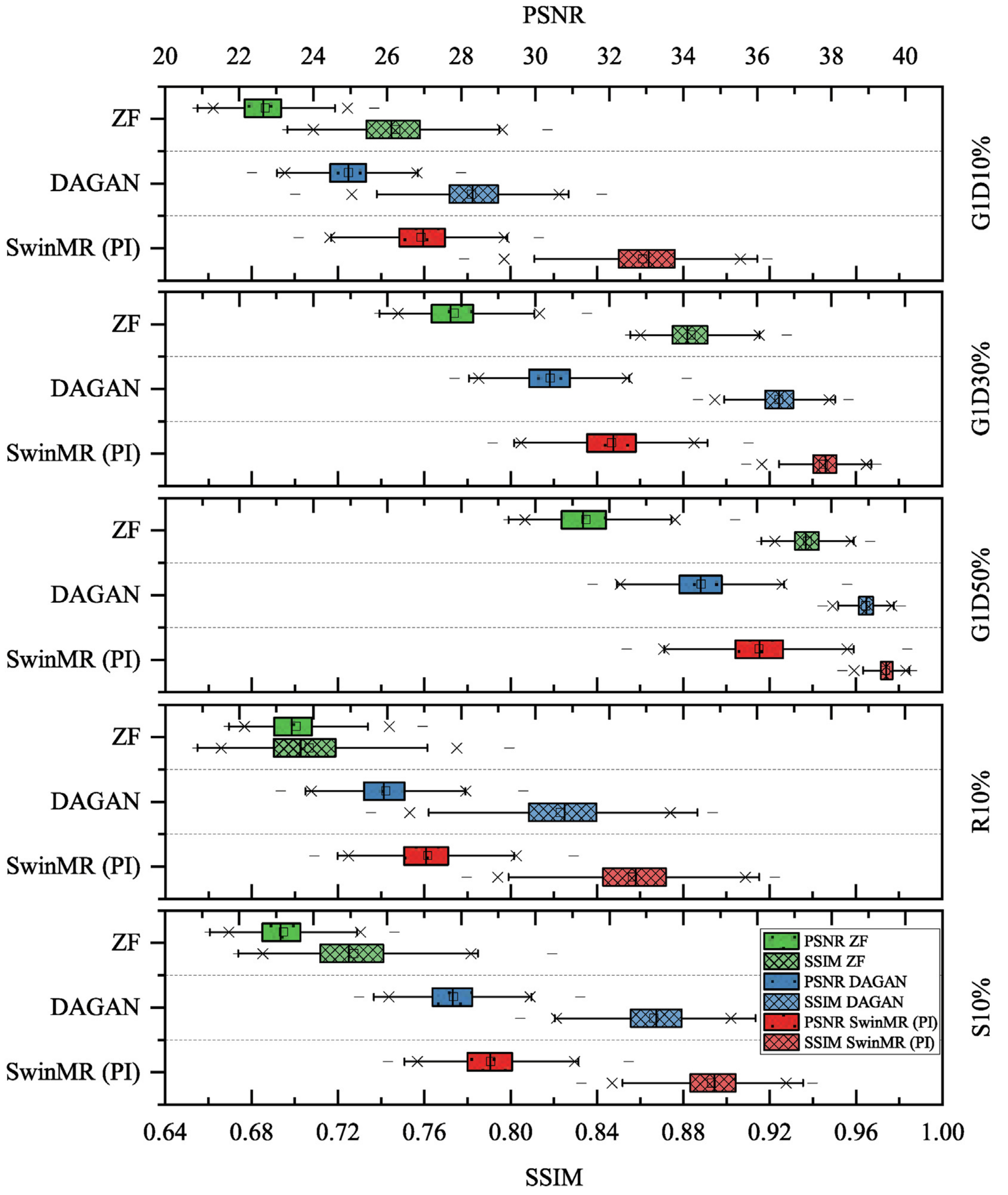


Fig. 5. Peak signal-to-noise ratio (PSNR) and Structural similarity index (SSIM) of the experiment on different masks. Five undersampling trajectories including Gaussian 1D 10% (G1D10%), Gaussian 1D 30% (G1D30%), Gaussian 1D 50% (G1D50%), radial 10% (R10%) and spiral 10% (S10%) were applied in this experiment. (Box range: interquartile range; ×:1% and 99% confidence interval; –: maximum and minimum; □: mean; |: median.) The SwinMR (PI) outperforms the DAGAN using different undersampling masks with significantly higher PSNR, SSIM ($p < 0.05$ by paired t-Test).

those CNN-based methods was limited by the local sensitivity of the convolutional operation. Motivated by this limitation, we proposed a Swin transformer based MR reconstruction method SwinMR.

2.3. SwinMR: Swin transformer for MRI reconstruction

2.3.1. Overall architecture

The overall architecture is shown in Fig. 1(C) and the data flow of SwinMR is shown in Fig. 2. Root sum square (RSS) is applied to combine the multi-channel ground truth MR images x^q to single-channel ground truth MR images x (q denotes the q^{th} coil). Sensitivity maps $\mathscr{S}^q$ are estimated by ESPIRiT [64] from multi-channel ground truth MR images x^q. Undersampling and noise interruption are performed in k-space using fast Fourier transform (FFT) and inverse fast Fourier transform (iFFT) (Gaussian noise is added in the noise experiments), which converts single-channel ground truth MR images x to undersampled zero-filled MR images x_u.

The proposed SwinMR model can produce reconstructed MR images $\hat{x}_u$ from undersampled zero-filled MR images x_u, where the residual connection is applied to accelerate the convergence and stable the training processing. It can be expressed by

$$\hat{x}_u = \text{SwinMR}(x_u|\theta) + x_u, \tag{7}$$

where the SwinMR network is parameterised by θ.

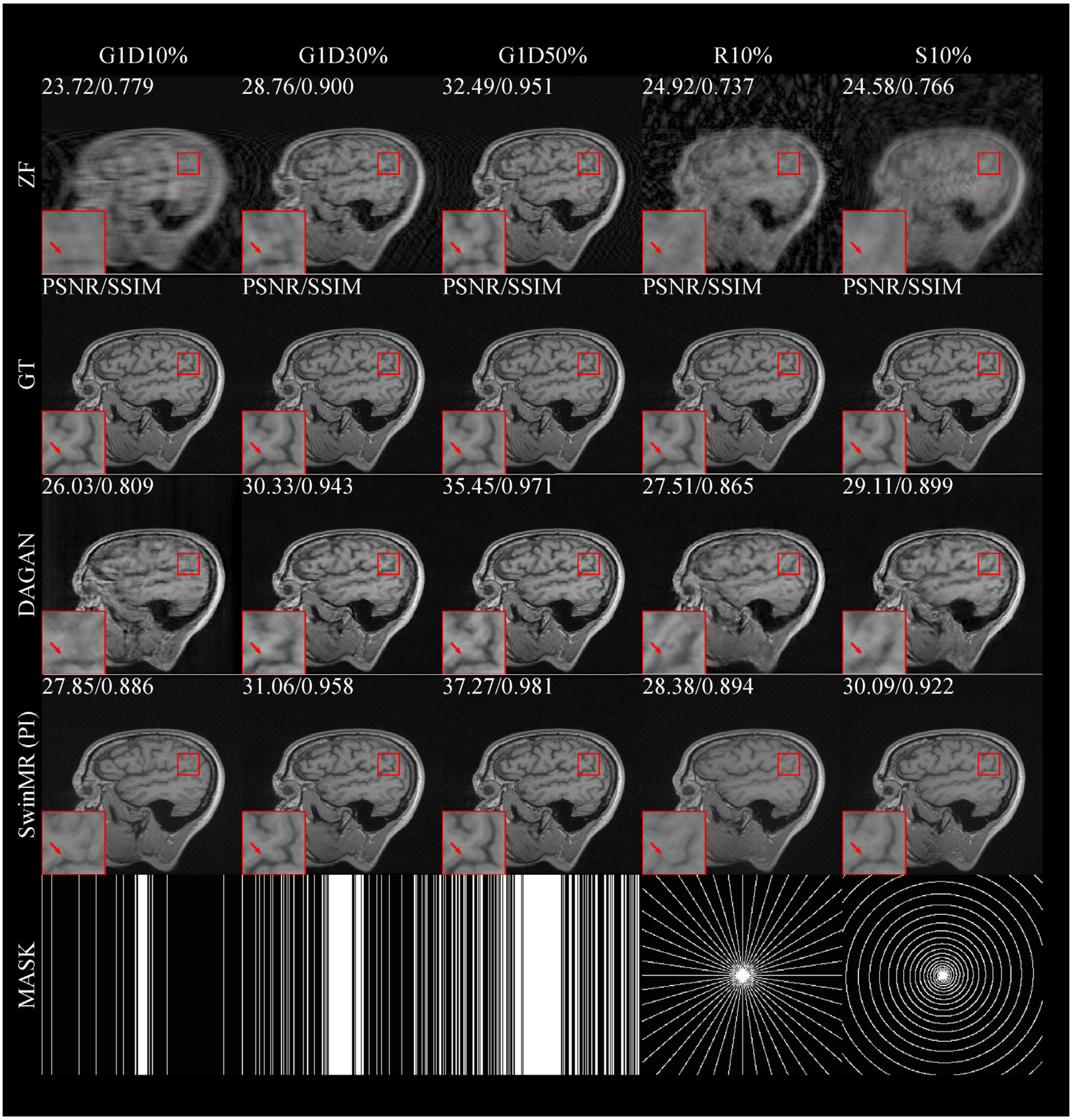


Fig. 6. Samples of the experiment on different masks. Five undersampling trajectories including Gaussian 1D 10% (G1D10%), Gaussian 1D 30% (G1D30%), Gaussian 1D 50% (G1D50%), radial 10% (R10%) and spiral 10% (S10%) were applied in this experiment. Row 1: Undersampled zero-filled MR images (ZF) using different masks; Row 2: Ground truth MR images (GT); Row 3: Reconstructed MR images by DAGAN; Row 4: Reconstructed MR images by SwinMR (PI); Row 5: Undersampling masks. The Peak signal-to-noise ratio (PSNR) and Structural similarity index (SSIM) of reconstructed and ZF images are shown in the top-left corner.

Fig. 3(A) shows the structure of SwinMR, which is composed of an input module (IM), a feature extraction module (FEM) and an output module (OM). The IM and OM are at the beginning and the end of the whole structure, and the FEM is placed between the IM and OM with a residual connection. The structure can be expressed by

$$F_{\mathrm{IM}} = \mathrm{H}_{\mathrm{IM}}(x_u), \tag{8}$$

$$F_{\mathrm{FEM}} = \mathrm{H}_{\mathrm{FEM}}(F_{\mathrm{IM}}), \tag{9}$$

$$F_{\mathrm{OM}} = \mathrm{H}_{\mathrm{OM}}(F_{\mathrm{FEM}} + F_{\mathrm{IM}}), \tag{10}$$

where the $\mathrm{H}_{\mathrm{IM}}(\cdot)$, $\mathrm{H}_{\mathrm{FEM}}(\cdot)$ and $\mathrm{H}_{\mathrm{OM}}(\cdot)$ denote the IM, FEM and OM respectively. $F_{\mathrm{IM}}, F_{\mathrm{FEM}}$ and F_{OM} denote the output of the IM, FEM and OM respectively.

2.3.2. Input module and output module

The IM is used for early visual processing and mapping from the input image space to higher dimensional feature space for the following FEM. The IM applies a 2D convolutional layer (Conv2D) mapping $x_u \in \mathbb{R}^{H\times W\times 1}$ to $F_{\mathrm{IM}} \in \mathbb{R}^{H\times W\times C}$. In contrast, the OM is used to map the higher dimensional feature space to the output image space by a Conv2D mapping $F_{\mathrm{FEM}} \in \mathbb{R}^{H\times W\times C}$ to $F_{\mathrm{OM}} \in \mathbb{R}^{H\times W\times 1}$.

In the training stage, the input image is randomly cropped to a fixed size $H \times W$ $(H = W)$. In the inference stage, H, W denote the height and weight of the input image. Here we define H (or W) as the patch number and C as the channel number for the self-attention processing.

2.3.3. Feature extraction module

The FEM is composed of a cascade of residual Swin transformer blocks (RSTBs) and a Conv2D at the end. It can be expressed as

$$F_0 = F_{\mathrm{IM}}, \tag{11}$$

$$F_i = \mathrm{H}_{\mathrm{RSTB}_i}(F_{i-1}), \quad i = 1, 2, \ldots, P, \tag{12}$$

$$F_{\mathrm{FEM}} = \mathrm{H}_{\mathrm{CONV}}(F_P), \tag{13}$$

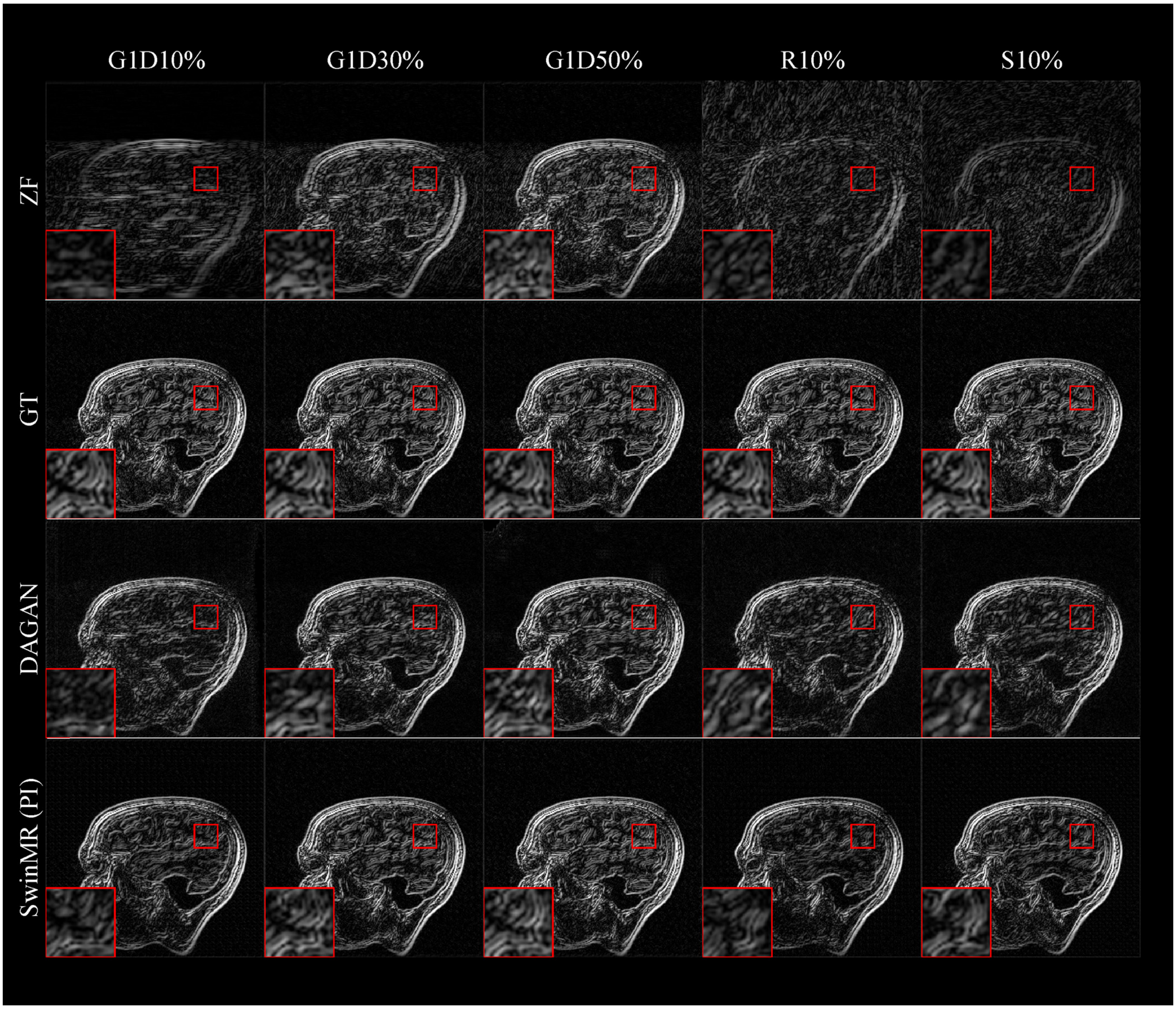


Fig. 7. Edge information of the experiment on different masks. Five undersampling trajectories including Gaussian 1D 10% (G1D10%), Gaussian 1D 30% (G1D30%), Gaussian 1D 50% (G1D50%), radial 10% (R10%) and spiral 10% (S10%) were applied in this experiment. Row 1: Edge information of undersampled zero-filled MR images (ZF) using different masks; Row 2: Edge information of ground truth MR images (GT); Row 3: Edge information of reconstructed MR images by DAGAN; Row 4: Edge information of reconstructed MR images by SwinMR (PI). The edge information was extracted by the Sobel operator.

where F_{IM} and F_{FEM} are the input and output of the FEM. $\mathrm{H}_{\mathrm{RSTB}_i}(\cdot)$ denotes the i^{th} RSTB (P RSTBs in total) in the FEM. $\mathrm{H}_{\mathrm{CONV}}(\cdot)$ denotes the Conv2D after a series of RSTBs.

Fig. 3(B) shows the structure of the RSTB. An RSTB consists of Q Swin transformer layers (STLs) and a Conv2D, and a residual connection is linked between the input and output of the RSTB. It can be expressed as

$$F_{i,0} = \mathrm{H}_{\mathrm{Emb}_i}(F_{i-1}), \tag{14}$$

$$F_{i,j} = \mathrm{H}_{\mathrm{STL}_{i,j}}(F_{i,j-1}), \quad j = 1,2,\ldots,Q, \tag{15}$$

$$F_i = \mathrm{H}_{\mathrm{CONV}_i}(\mathrm{H}_{\mathrm{Unemb}_i}(F_{i,Q}) + F_{i-1}), \tag{16}$$

where $\mathrm{H}_{\mathrm{Emb}_i}(\cdot)$ is the patch embedding from $F_{i-1} \in \mathbb{R}^{H\times W\times C}$ to $F_{i,0} \in \mathbb{R}^{HW\times C}$, and $\mathrm{H}_{\mathrm{Unemb}_i}(\cdot)$ is the patch unembedding from $F_{i,Q} \in \mathbb{R}^{HW\times C}$ to $\mathbb{R}^{H\times W\times C}$.

$\mathrm{H}_{\mathrm{STL}_{i,j}}(\cdot)$ and $\mathrm{H}_{\mathrm{CONV}_i}(\cdot)$ denote the j^{th} STL and the Conv2D in the i^{th} RSTB, respectively.

2.3.4. Swin transformer layer

The whole process of the STL can be expressed as

$$X' = \mathrm{H}_{\mathrm{(S)W-MSA}}(\mathrm{H}_{\mathrm{LN}}(X)) + X, \tag{17}$$

$$X'' = \mathrm{H}_{\mathrm{MLP}}(\mathrm{H}_{\mathrm{LN}}(X')) + X', \tag{18}$$

where X and X'' are the input and output of the STL. $\mathrm{H}_{\mathrm{MLP}}(\cdot)$ and $\mathrm{H}_{\mathrm{LN}}(\cdot)$ denote the multilayer perceptron and the layer normalisation layer. Windows multi-head self-attention (W-MSA) and shifted windows multi-head self-attention (SW-MSA) $\mathrm{H}_{\mathrm{(S)W-MSA}}(\cdot)$ are alternatingly applied in consecutive STLs.

Spatial constraints are added in the Swin transformer layer compared to the original transformers. Fig. 3(B) shows the W-MSA and the SW-MSA compared with the original MSA. Original MSA performs self-attention in the whole image space. Although the information of the entire picture is involved in each attention calculation, it aggravates computational costs and redundant connections. The computational complexity for the original MSA is as follows:

$$\Omega(\mathrm{H}_{\mathrm{MSA}}) = 4HWC^2 + 2(HW)^2C. \tag{19}$$

In Swin transformer layers, a $\mathbb{R}^{H\times W\times C}$ feature map are divided into $\frac{HW}{M^2}$ non-overlapped windows with the size of $M^2 \times C$. (S)W-MSA is calculated in each window, instead of the whole image space. The computational complexity for (S)W-MSA is as follows:

$$\Omega(\mathrm{H}_{\mathrm{(S)W-MSA}}) = 4HWC^2 + 2M^2HWC, \tag{20}$$

which is significantly reduced compared to the original MSA. However, if the separation of windows is fixed between STLs, the network will lose the link between different windows. Normal windows and shifted windows are alternatingly utilised in consecutive STLs to enable information communication from different windows.

(S)W-MSA for each non-overlap window X can be expressed by

$$Q = XP_Q, \quad K = XP_K, \quad V = XP_V, \tag{21}$$

where the P_Q, P_K, P_V are shared projection matrices over all the windows. The query Q, key K, value V and learnable relative position encoding B ($\mathbb{R}^{M^2\times d}$) are used in the calculation of the self-attention mechanism in a local window, which can be expressed by

$$\mathrm{Attention}(Q,K,V) = \mathrm{SoftMax}\left(QK^T/\sqrt{d} + B\right)V. \tag{22}$$

Such self-attention mechanism calculations are performed for h times and concatenated for (S)W-MSA. The pseudo-code of STL and (S)W-MSA are shown in Algorithm 1 and Algorithm 2.

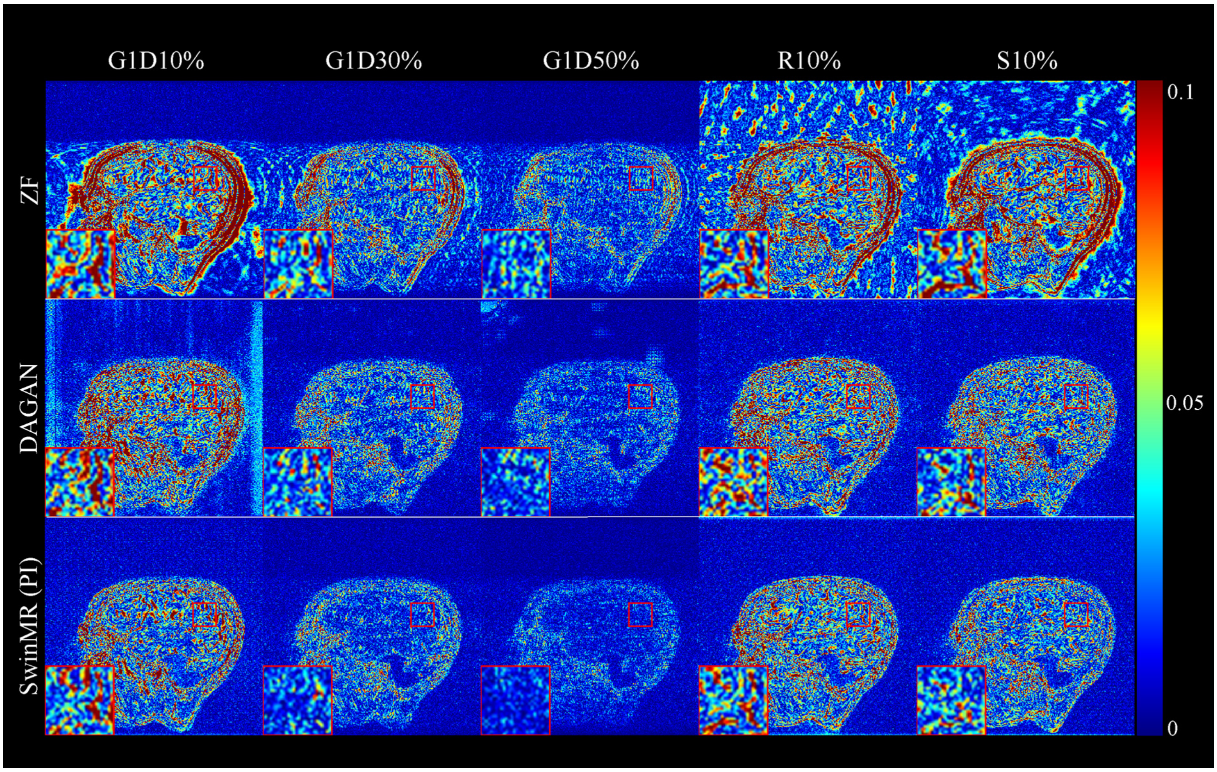


Fig. 8. Absolute differences of standardised pixel intensities (10×) of the experiment on different masks. Five undersampling trajectories including Gaussian 1D 10% (G1D10%), Gaussian 1D 30% (G1D30%), Gaussian 1D 50% (G1D50%), radial 10% (R10%) and spiral 10% (S10%) were applied in this experiment. Row 1: Absolute differences between undersampled zero-filled MR images (ZF) using different masks and ground truth MR images (GT); Row 2: Absolute differences between reconstructed MR images by DAGAN and GT; Row 3: Absolute differences between reconstructed MR images by SwinMR (PI) and GT.

Algorithm 1: Swin transformer layer (STL).

```
Input: $X, j, W_s, N_h$
    # $X$: Feature maps;
    # $X$.shape: ($C$, $H$, $W$); $C$: embedding channel; $H$: height; $W$: width;
    # $j$: index of STL (from 0); $W_s$: size of window; $N_h$: number of heads.

    $N_w = HW/W_s^2$    # Calculate the number of windows.

    $X_{\text{tmp}} \leftarrow X$    # For residual connection.
    $X \leftarrow \text{LN}(X)$    # Layer normalisation 1.

    # Shifting operation used in even STL.
    if $j\%2 \neq 0$ then
    $X \leftarrow \text{cyclic\_shift}(X)$
    end if
    # Split feature maps into non-overlapping windows.
    $X_{\text{win}} \leftarrow \text{window\_partition}(X)$    # $X_{\text{win}}$.shape: ($N_wN_h$, $C/N_h$, $W_s$, $W_s$)
    # Multi-head self-attention
    $X_{\text{win}} \leftarrow \text{MSA}(X_{\text{win}})$    # $X_{\text{win}}$.shape: ($N_wN_h$, $C/N_h$, $W_s$, $W_s$)
    # Recover feature maps from windows
    $X \leftarrow \text{reverse\_window}(X_{\text{win}})$    # $X$.shape: ($C$, $H$, $W$)
    # Corresponding shifting reversing operation used in even STL.
    if $j\%2 \neq 0$ then
    $X \leftarrow \text{reverse\_shift}(X)$
    end if
    $X \leftarrow X + X_{\text{tmp}}$    # Residual connection 1.

    $X_{\text{tmp}} \leftarrow X$    # For residual connection.
    $X \leftarrow \text{LN}(X)$    # Layer normalisation 2.
    $X \leftarrow \text{MLP}(X)$    # Multi-layer perceptron.
    $X \leftarrow X + X_{\text{tmp}}$    # Residual connection 2.

Output: $X$
```

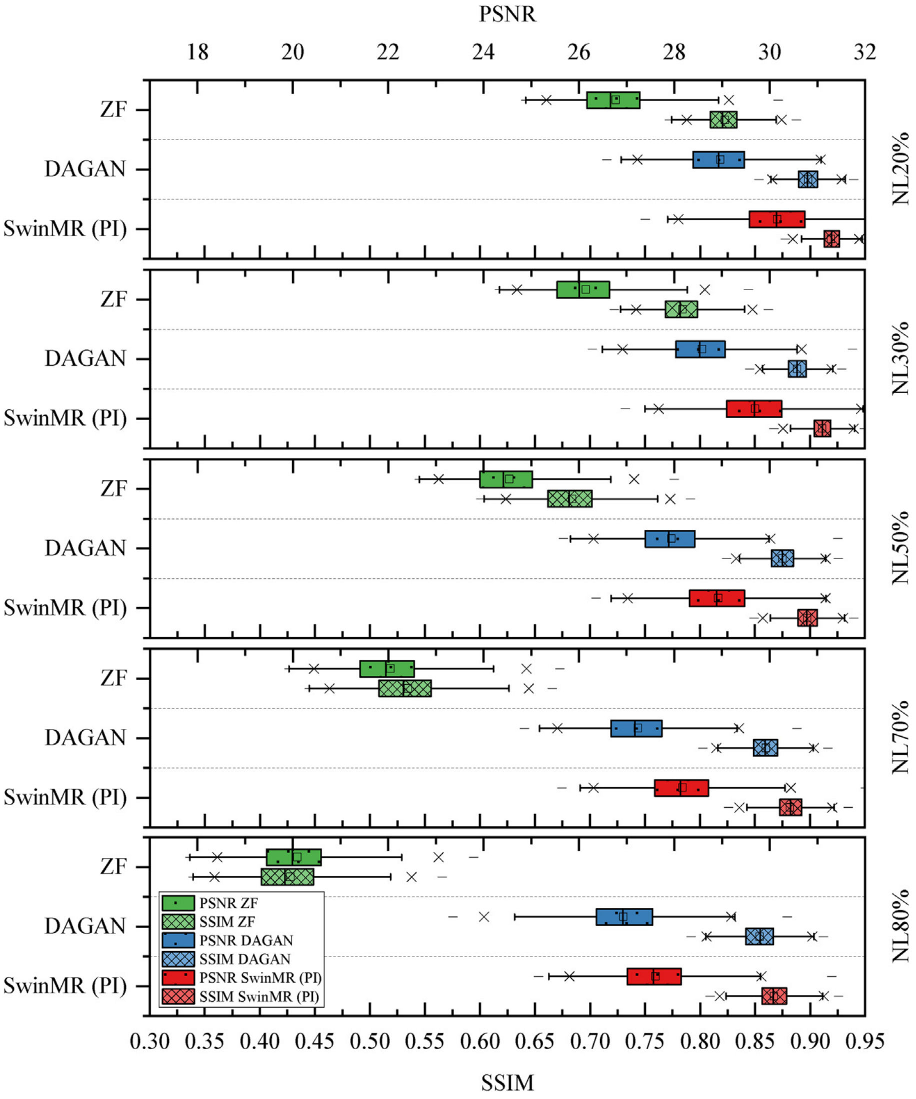


Fig. 9. Peak signal-to-noise ratio (PSNR) and Structural similarity index (SSIM) of the experiment on different noise using Gaussian 1D 30% mask. Five noise levels (NL20%, NL30%, NL50%, NL70% and NL80%) were tested in this experiment. (Box range: interquartile range; ×:1% and 99% confidence interval; –: maximum and minimum; □: mean; |: median.) The SwinMR (PI) outperforms the DAGAN under different noise levels with significantly higher PSNR, SSIM ($p < 0.05$ by paired t-Test).

Algorithm 2: (Shifted) windows multi-head self-attention.

```
Input: $X$
    # $X$: windows for multi-head self-attention operation;
    # $X$.shape: $(N_w N_h, C/N_h, W_s, W_s)$;
    # $N_w$: number of windows; $N_h$: number of heads;
    # $C$: embedding channel; $W_s$: size of window.

    # Calculate the query, key and value.
    $Q \leftarrow \mathrm{Linear}_q(X)$
    $K \leftarrow \mathrm{Linear}_k(X)$
    $V \leftarrow \mathrm{Linear}_v(X)$

    # Calculate the relative position bias.
    $B \leftarrow \mathrm{get\_relative\_position}(X)$

    # Calculate the attention result.
    $\mathrm{attn\_map} \leftarrow \mathrm{dot}(Q, K.\mathrm{transpose})/\sqrt{C}$
    $\mathrm{attn\_map} \leftarrow \mathrm{SoftMax}(\mathrm{attn\_map} + B)$
    $\mathrm{attn} \leftarrow \mathrm{dot}(\mathrm{attn\_map}, V)$
    $\mathrm{attn} \leftarrow \mathrm{Linear}(\mathrm{attn})$
Output: attn
```

Table 3
Fréchet inception distance (FID) of the experiment on different noise using Gaussian 1D 30% mask. Five noise levels (NL20%, NL30%, NL50%, NL70% and NL80%) were applied in this experiment.

Noise Level	SwinMR (PI)	DAGAN	ZF
NL20%	**16.07**	66.60	156.77
NL30%	**16.44**	71.50	168.61
NL50%	**24.29**	75.32	203.39
NL70%	**30.65**	85.32	251.15
NL80%	**33.79**	80.97	282.40

2.3.5. Loss function

A novel multi-channel loss using the sensitivity maps was introduced for better reconstruction quality and more textures and details. Charbonnier loss [65] was utilised for the pixel-wise loss and the frequency loss since it is more robust and able to handle the outliers better. The total loss $\mathscr{L}_{\mathrm{TOTAL}}(\theta)$ consists of the pixel-wise Charbonnier loss $\mathscr{L}_{\mathrm{pixel}}(\theta)$, the frequency Charbonnier loss $\mathscr{L}_{\mathrm{freq}}(\theta)$ and perceptual loss $\mathscr{L}_{\mathrm{VGG}}(\theta)$. The pixel-wise Charbonnier loss can be expressed by

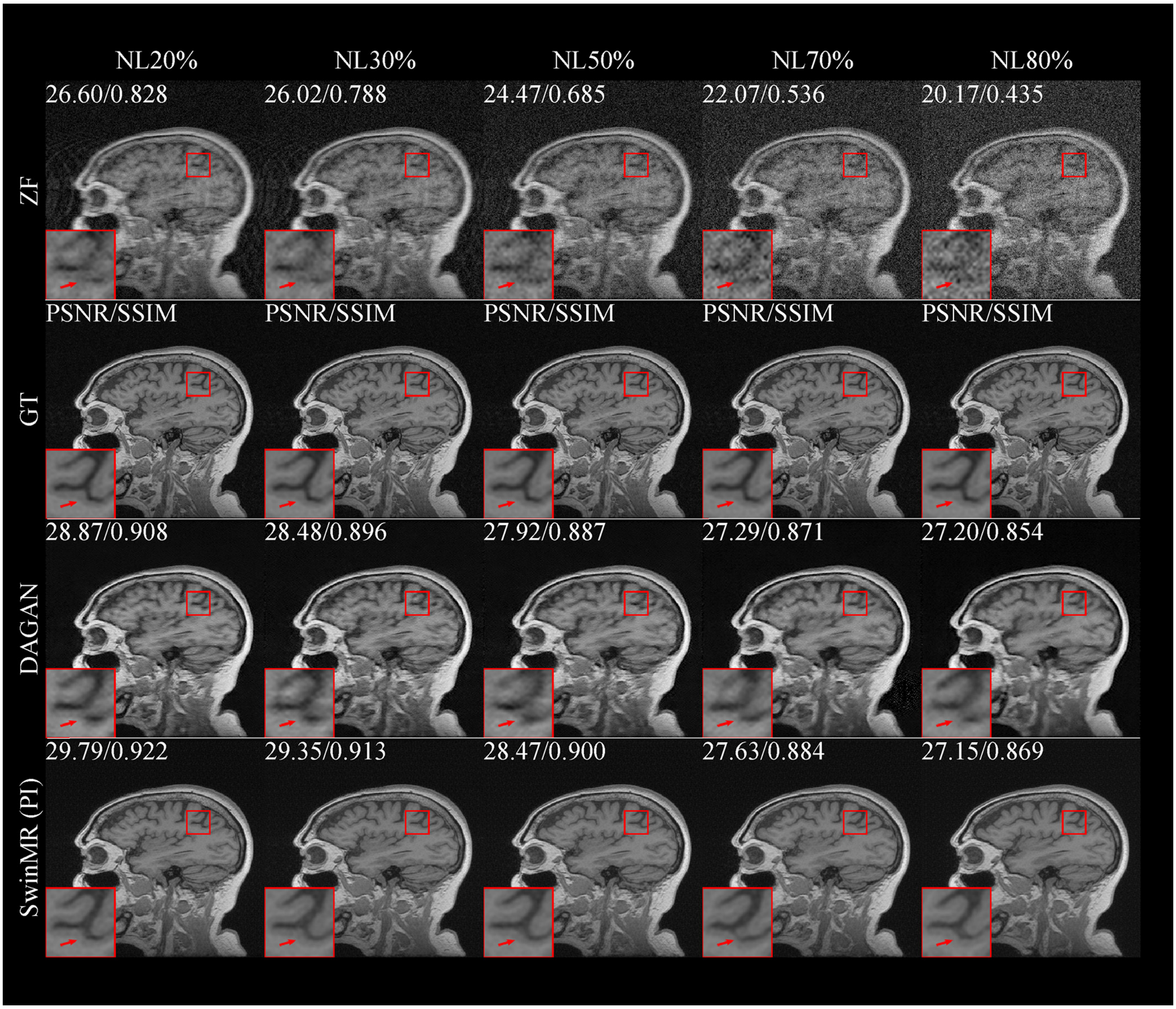


Fig. 10. Samples of the experiment on different noise using Gaussian 1D 30% mask. Five noise levels (NL20%, NL30%, NL50%, NL70% and NL80%) were tested in this experiment. Row 1: Undersampled zero-filled MR images (ZF) with different noise levels; Row 2: Ground truth MR images (GT); Row 3: Reconstructed MR images by DAGAN; Row 4: Reconstructed MR images by SwinMR (PI). The Peak signal-to-noise ratio (PSNR) and Structural similarity index (SSIM) of reconstructed and ZF images are shown in the top-left corner.

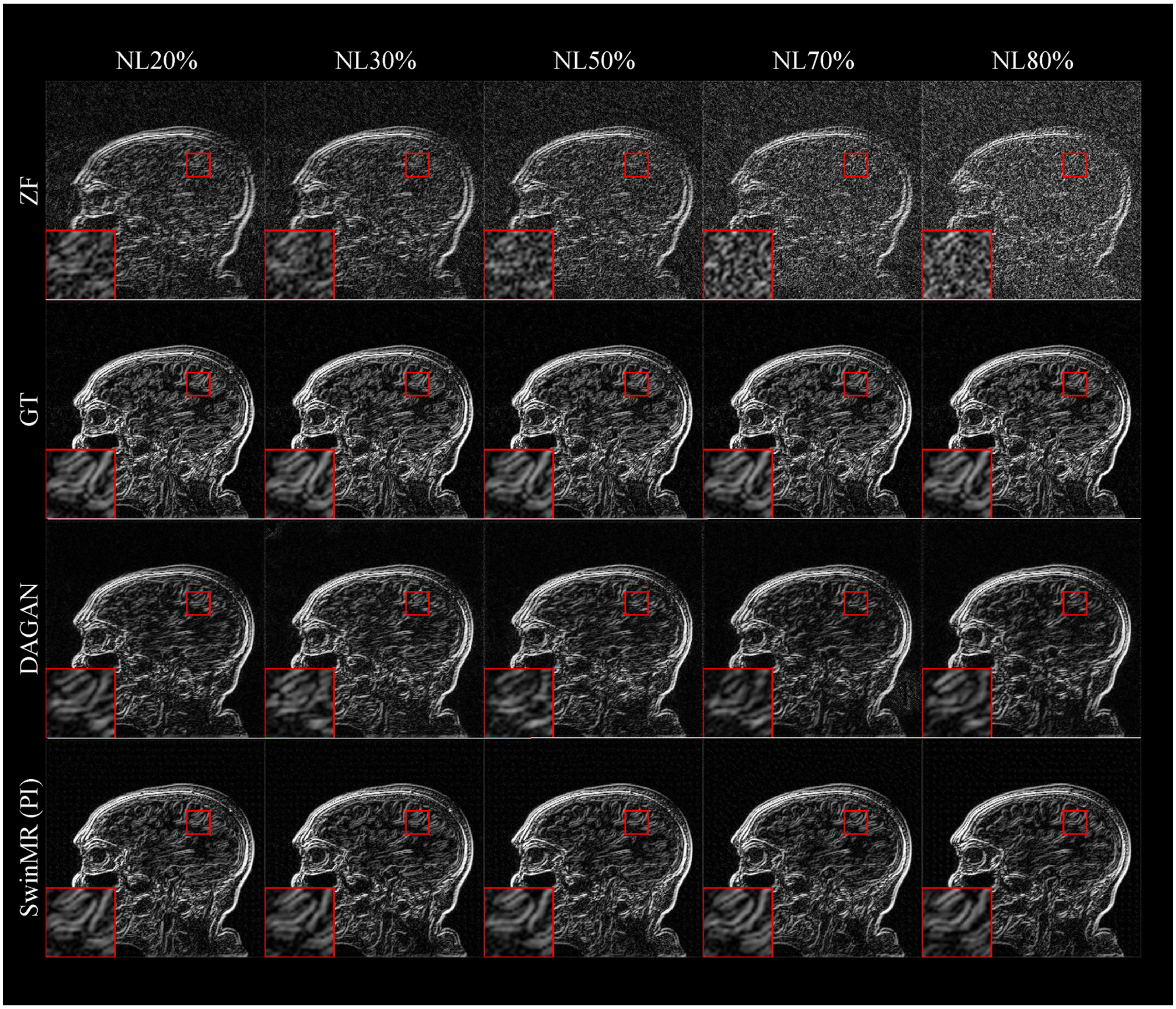


Fig. 11. Edge information of the experiment on different noise using Gaussian 1D 30% mask. Five noise levels (NL20%, NL30%, NL50%, NL70% and NL80%) were tested in this experiment. Row 1: Edge information of undersampled zero-filled MR images (ZF) with different noise levels; Row 2: Edge information of ground truth images (GT); Row 3: Edge information of reconstructed MR images by DAGAN; Row 4: Edge information of reconstructed MR images by SwinMR (PI). The edge information was extracted by the Sobel operator.

$$\min_{\theta} \mathscr{L}_{\text{pixel}}(\theta) = \frac{1}{S}\sum_{q=1}^{S}\sqrt{||x^q - \mathscr{S}^q\hat{x}_u||_2^2 + \epsilon^2}, \tag{23}$$

where ϵ is a constant which is set to 10^{-9} empirically and $\mathscr{S}^q$ is the sensitivity map of q^{th} coil (S colis in total). The frequency Charbonnier loss can be expressed by

$$\min_{\theta} \mathscr{L}_{\text{freq}}(\theta) = \frac{1}{S}\sum_{q=1}^{S}\sqrt{||y^q - \mathscr{F}\mathscr{S}^q\hat{x}_u||_2^2 + \epsilon^2}. \tag{24}$$

The perceptual VGG loss can be expressed by

$$\min_{\theta} \mathscr{L}_{\text{VGG}}(\theta) = ||f_{\text{VGG}}(x) - f_{\text{VGG}}(\hat{x}_u)||_1, \tag{25}$$

where $f_{\text{VGG}}(\cdot)$ denotes the VGG network, and $||\cdot||_1$ denotes the l_1 norm. The utilisation of $\mathscr{L}_{\text{VGG}}$ is able to optimise the perceptual quality of reconstructed results.

The total loss can be expressed by

$$\mathscr{L}_{\text{TOTAL}}(\theta) = \alpha\mathscr{L}_{\text{pixel}}(\theta) + \beta\mathscr{L}_{\text{freq}}(\theta) + \gamma\mathscr{L}_{\text{VGG}}(\theta), \tag{26}$$

where α, β and γ are coefficients controlling the balance of each term in the loss function.

3. Experiments and results

3.1. Datasets

In this work, the Calgary Campinas multi-channel (CC) dataset[1] [66] and the Multi-modal Brain Tumour Segmentation Challenge 2017 (BraTS17)[2] [67–69] dataset were used for the experiment sections.

The available data of the CC dataset contains 67 cases of three-dimensional (3D), 12-channel (117 scans), T1-weighted, gradient-recalled echo, 1 mm isotropic sagittal acquisitions. Acquisition parameters were TR/TE/TI = 6.3 ms/2.6 ms/650 ms (93 scans)

[1] https://sites.google.com/view/calgary-campinas-dataset/mr-reconstruction-challenge.

[2] https://www.med.upenn.edu/sbia/brats2017/data.html.

and TR/TE/TI = 7.4 ms/3.1 ms/400 ms (74 scans), with 170 to 180 contiguous 1.0-mm slices and a field of view of 256 mm × 218 mm. The original CC dataset provides a hybrid (x, k_y, k_z, C) structure (x: read-out direction; y: phase-encoding direction; z: slice-encoding direction; C: channels), where inverse Fourier transform is performed in the read-out direction. These 3D hybrid data were uniformly zero-filled in the phase-encoding direction to 256×256, and turned into 3D image space volumes by 2D iFFT on the $k_y - k_z$ plane. We randomly chose 40 cases for training, 7 cases for validation and 20 cases for testing, according to the ratio of 6:1:3 approximately. In each case, we chose 100 2D slices near the centre along the read-out direction (sagittal view).

For the BraTS17 dataset, we applied the brain data with reference segmentation results (280 3D volumes in BraTS17 official training dataset), including both higher and lower grade glioma. These multi-modal scans contain native T1-weighted (T1), T1-contrast enhanced (T1CE), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR) data. These 280 3D brain data were divided into training, validation and testing set (235, 20, and 30 cases respectively), and cropped to $152 \times 192 \times 144$ volumes (slice, height and width, respectively). For each case, we used 100 slices near the centre in the training stage to avoid invalid data, i.e., slices that are totally dark or with little information, for training.

3.2. Implementation detail

The proposed SwinMR was implemented using PyTorch, trained on two NVIDIA RTX 3090 GPUs with 24 GB GPU memory, and tested on an NVIDIA RTX 3090 GPU or an Intel Core i9-10980XE CPU. We set the RSTB number, the STL number, the window size number and the attention head number to 6, 6, 8 and 6 respectively, which are the default setting in the original SwinIR [45]. The patch number and channel number were empirically set to 96 and 180, according to our ablation studies. For the parameter in the loss function, α, β, γ were set to 15, 0.1 and 0.0025 to balance each term, according to our ablation studies. Our proposed SwinMR was trained for 100,000 steps using Adam optimiser. The initial learning rate was set to 2×10^{-4} and decayed by 0.5 every 10,000 steps from the 50,000th step. Random flip and rotation were applied for data augmentation.

We used SwinMR (PI) to denote the proposed model trained with multi-channel data and sensitivity maps, and SwinMR (nPI) to indicate the proposed model trained with single-channel data without sensitivity maps.

3.3. Evaluation methods

Structural similarity index (SSIM), Peak signal-to-noise ratio (PSNR) and Fréchet inception distance (FID) [70] were utilised for evaluation. SSIM quantifies the structural similarity between two images based on luminance, contrast, and structures. PSNR is the ratio between maximum signal power and noise power, which measures the fidelity of the representation. Both metrics are based on simple and shallow functions, and direct comparisons between images, which are not necessary for the visual quality for human observers [71]. FID is calculated by computing the Fréchet distance between two multivariate Gaussians, which measures the similarity between two sets of images. FID correlates well with visual quality for human observers, and a lower FID indicates more perceptual results.

Both Intersection over Union (IoU) and Dice scores were applied to measure the segmentation quality in the brain tumour segmentation experiment.

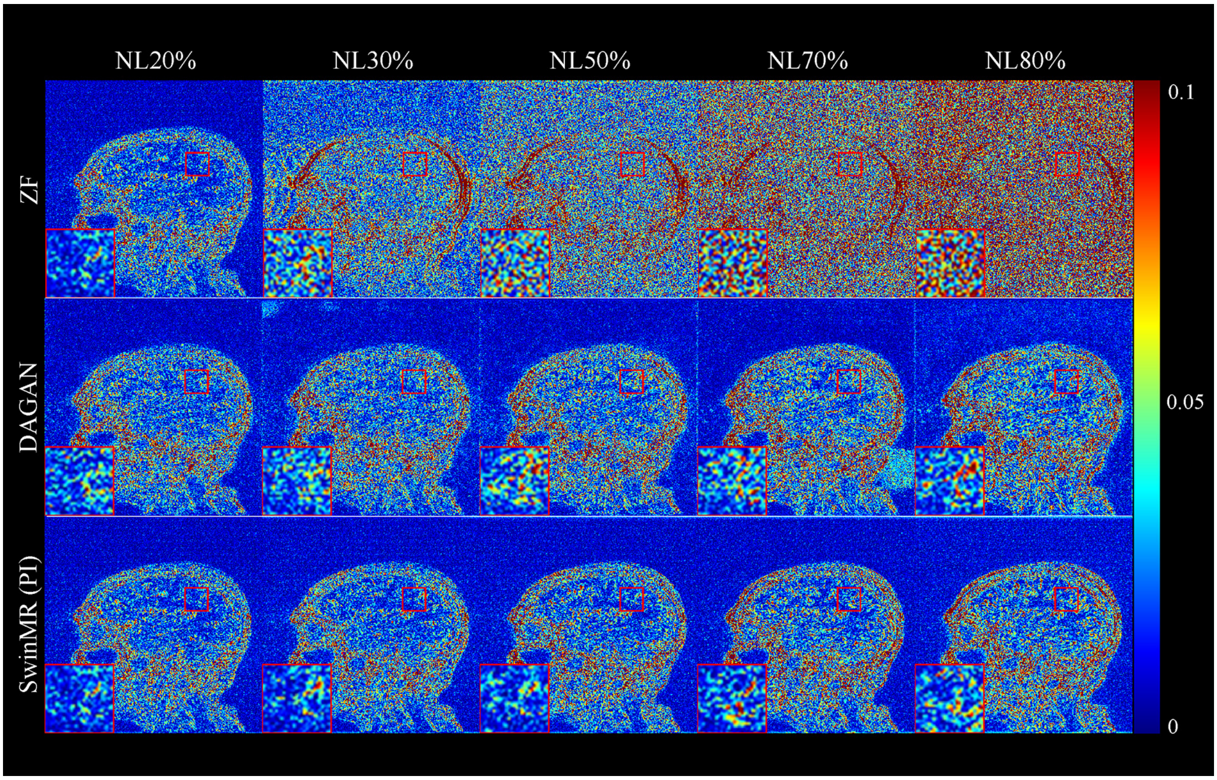


Fig. 12. Absolute differences of standardised pixel intensities (10×) of the experiment on different using Gaussian 1D 30% mask noise. Five noise levels (NL20%, NL30%, NL50%, NL70% and NL80%) were tested in this experiment. Row 1: Absolute differences between undersampled zero-filled MR images (ZF) with different noise levels and ground truth MR images (GT); Row 2: Absolute differences between reconstructed MR images by DAGAN and GT; Row 3: Absolute differences between reconstructed MR images by SwinMR (PI) and GT.

The number of parameters (#PARAMs) and Multiply-Accumulate Operations (MACs) were applied to measure the model size and the computational cost. MACs were calculated using a $1 \times 1 \times 256 \times 256$ array as input (Batch × Channel × Height × Width).

3.4. Comparisons with other methods

In this experimental study, we compared our proposed SwinMR (nPI and PI) with other benchmarked MR reconstruction methods, including Deep ADMM Net [23], U-Net [62], DAGAN [37], PIDDGAN [56], as well as ground truth MR images (GT) and undersampled zero-filled MR images (ZF) using Gaussian 1D 30% mask. Among them, PIDDGAN and SwinMR (PI) were parallel imaging-coupled, i.e., trained with multi-channel MR images. This experiment was conducted using the CC dataset.

The quantitative result of comparisons is shown in Table 1. Our proposed SwinMR (nPI) achieved the highest SSIM and PSNR, and SwinMR (PI) achieved the best FID score. The inference time in Table 1 indicates the average time for one inference measured by ten times inferences in average in an Intel Core i9-10980XE CPU or an NVIDIA RTX 3090 GPU. The computational cost of SwinMR was higher than other CNN-based models. SwinMR has a larger computational cost (MACs) than other CNN-based and GAN-based methods, but with a smaller model size (#PARAMs).

Fig. 4 shows the reconstructed MR images, edge information extracted by Sobel operator and absolute differences of standardised pixel intensities (10×) between reconstructed MR images and GT MR images from top to button respectively. The proposed SwinMR shows superiority to other methods in terms of overall reconstruction quality and edge information.

3.5. Experiments on masks

This experimental study aimed to evaluate the performance of SwinMR using different undersampling trajectories. Three 1D Cartesian undersampling trajectories including Gaussian 1D 10% (G1D10%), Gaussian 1D 30% (G1D30%) and Gaussian 1D 50% (G1D50%), as well as two 2D non-Cartesian undersampling trajectories including radial 10% (R10%) and spiral 10% (S10%) were applied in this experiment. This experiment compared the SSIM, PSNR and FID of SwinMR (PI), DAGAN and ZF, and was conducted using the CC dataset.

The quantitative results of the experiment on masks are shown in Fig. 5 and Table 2. The sample of reconstructed images, edge information and absolute differences of standardised pixel intensities (10×) between reconstructed images and GT images are shown in Figs. 6–8 respectively. According to the results, the proposed SwinMR achieved a higher reconstruction quality compared

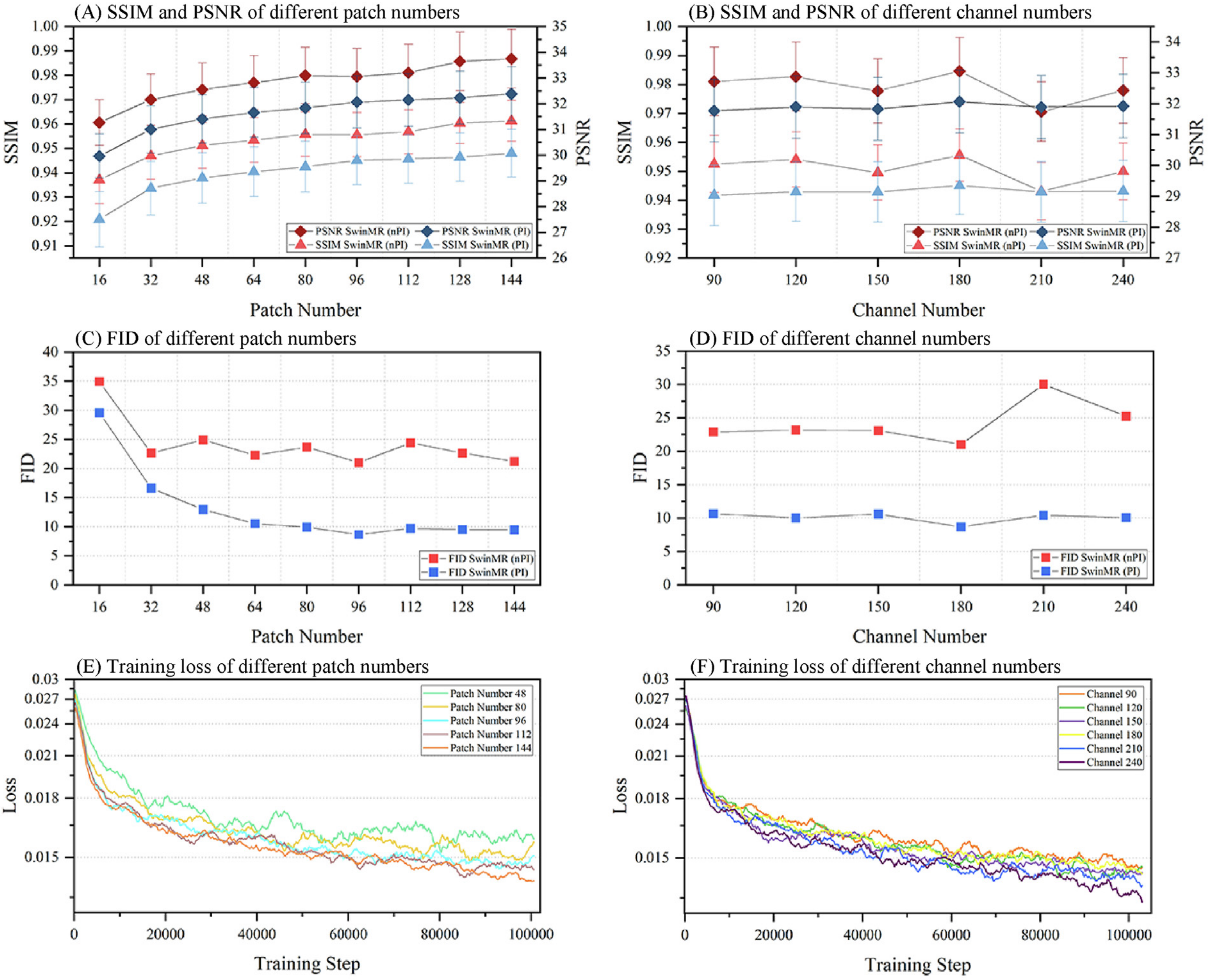


Fig. 13. Structural similarity index (SSIM), Peak signal-to-noise ratio (PSNR), Fréchet inception distance (FID) and training loss of ablation experiments of the patch number and channel number. (A), (C) and (E) are the SSIM/PSNR, FID and training loss of the ablation experiment of the patch number. (B), (D) and (F) are the SSIM/PSNR, FID and training loss of the ablation experiment of the channel number.

to DAGAN using different undersampling trajectories, especially when the mask of low undersampling rate (10%) was applied.

3.6. Experiments on noise

This experimental study aimed to evaluate the robustness of SwinMR under the influence of noise. The noise in MRI is imposed on the *k*-space and that could follow a Gaussian distribution [72]. In our experiments, different noise levels (NL20%, NL30%, NL50%, NL70% and NL80%) were tested after undersampling (Gaussian 1D 30% mask) in *k*-space. The noise level is defined as:

$$\mathrm{NL} = \frac{N'}{S' + N'}, \tag{27}$$

where N' and S' denote the power of noise and signal, respectively. This experiment compared the SSIM, PSNR and FID of SwinMR (PI), DAGAN and ZF, and was conducted using the CC dataset.

The quantitative results of the noise experiments are shown in Fig. 9 and Table 3. The sample of reconstructed images, edge information and absolute differences of standardised pixel intensities (10×) between reconstructed images and GT images are shown in Figs. 10–12, respectively.

According to the results, under the interruption of noise, SwinMR maintains better reconstruction quality compared to DAGAN. The quality improvement becomes more clear when under a high noise level.

3.7. Ablation experiments on the patch number and channel number

The patch number H (or W) and the channel number C decide the input size of STL in the SwinMR. Ablation studies for different patch numbers and channel numbers were conducted to study the impression of them on the reconstruction results.

Figs. 13 (A) and 13 (C) show the SSIM, PSNR and FID of SwinMR with different patch numbers. Fig. 13 (E) shows the loss function of SwinMR in the training process. Fig. 14 displays the sample of reconstructed images of SwinMR with different patch numbers.

Figs. 13 (B) and 13 (D) show the SSIM, PSNR and FID of SwinMR with different channel numbers. Fig. 13 (F) shows the loss function of SwinMR in the training process. Fig. 15 displays the sample of reconstructed images of SwinMR with the different channel numbers.

For the patch number, from Figs. 13 (A) and 13 (C), the results demonstrate that reconstruction quality becomes better as the patch number grows. According to Fig. 13 (E), the training loss converges faster and lower as the patch number grows. However, the growing patch number aggravates the computational cost. Empirically, we applied patch number 96 for training.

For the channel number, from Figs. 13 (B) and 13 (D), the results did not resemble the trend presented in the ablation experiment on patch number. There were no significant differences for the three metrics (SSIM, PSNR and FID) as the channel number changed. According to Fig. 13 (F), the training loss converges faster and lower as the channel number grows. Empirically, we applied a channel number of 180 for training.

For the comparison of multi-channel data (PI) and single-channel data (nPI), SwinMR (PI) tend to have a better (lower) FID, but worse (lower) SSIM/PSNR than SwinMR (nPI).

3.8. Ablation experiments on the loss function

This ablation study aimed to discover the effect of each term in the loss function. According to Eq. (26), the loss function of SwinMR consists of pixel-wise loss, frequency loss and perceptual loss. Four experiments were performed in this ablation study: (1) PFP: **P**ixel-wise, **F**requency and **P**erceptual loss; (2) PP: **P**ixel-

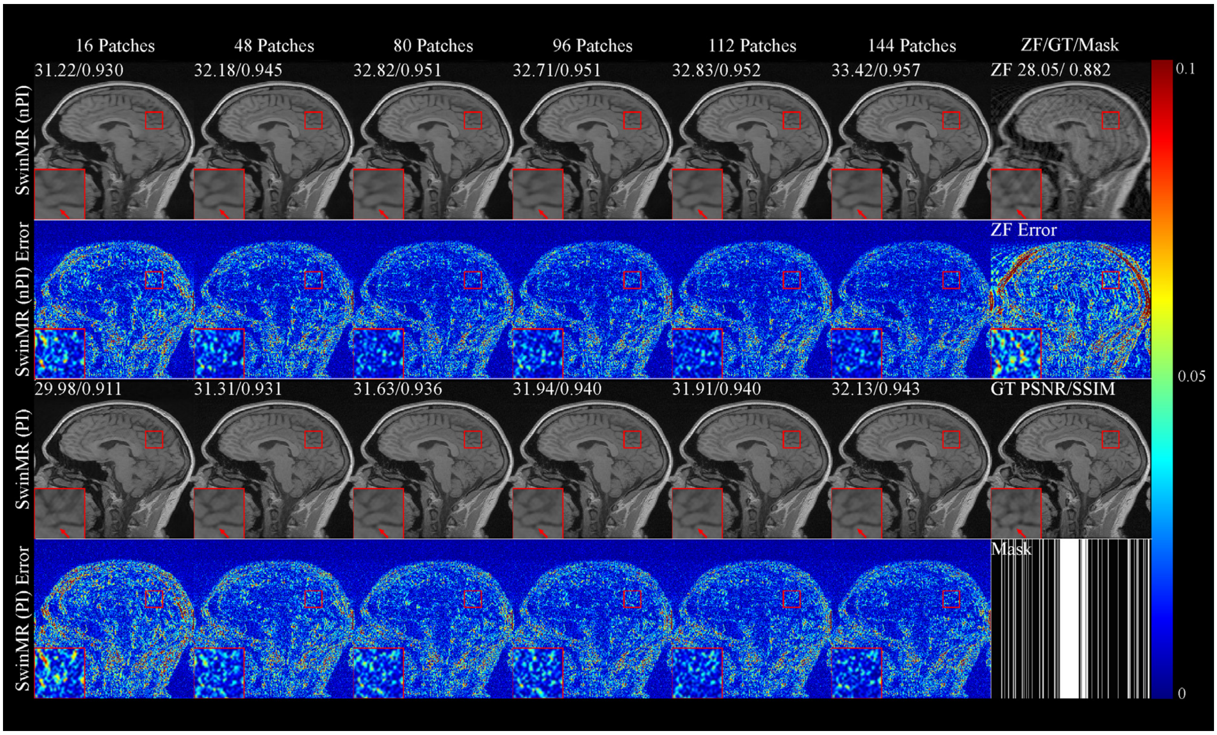


Fig. 14. Samples of the ablation experiment on the patch number using Gaussian 1D 30% mask. Row 1: Reconstructed MR images by SwinMR (nPI) with different patch numbers and zero-filled MR images (ZF); Row 2: Absolute differences (10×) between reconstructed MR images by SwinMR (nPI) and ground truth MR images (GT), and absolute differences (10×) between ZF and GT; Row 3: Reconstructed MR images by SwinMR (PI) with the different patch number and GT; Row 4: Absolute differences (10×) between reconstructed MR images by SwinMR (PI) and GT, and the Gaussian 1D 30% mask.

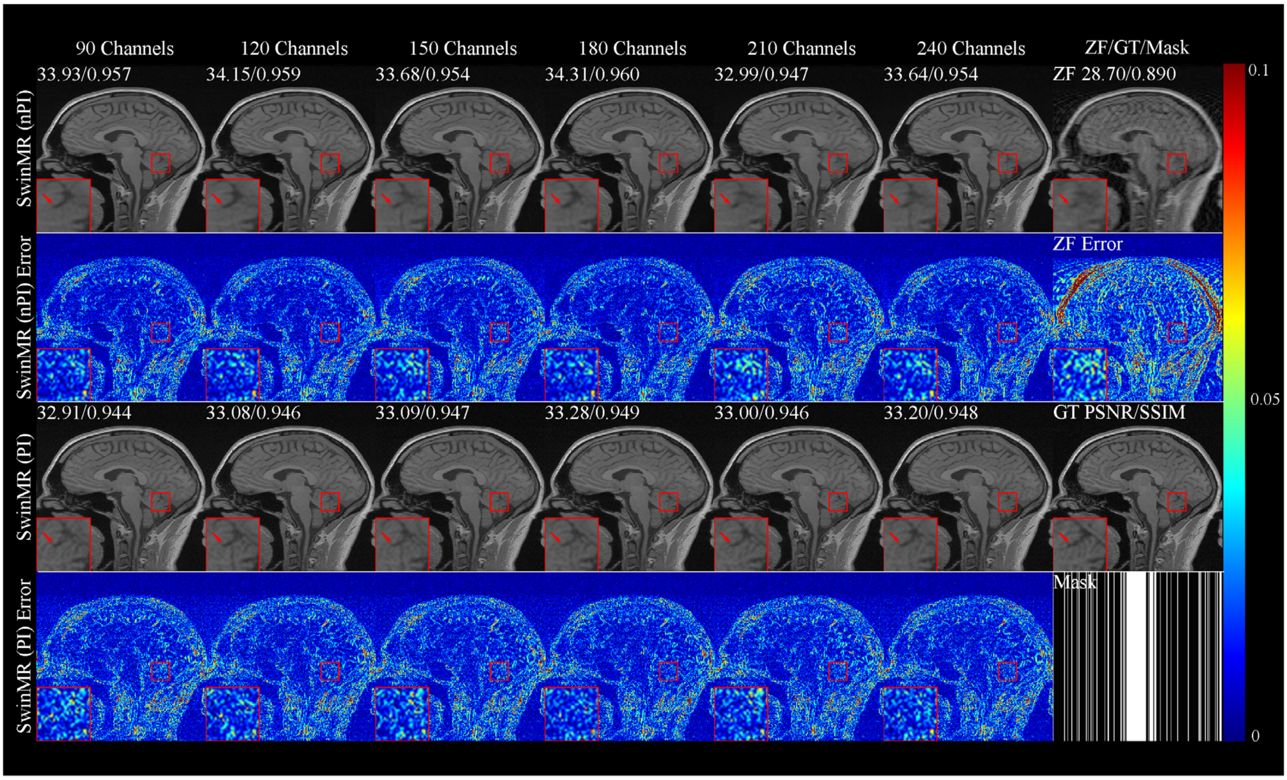


Fig. 15. Samples of the ablation experiment on the channel number using Gaussian 1D 30% mask. Row 1: Reconstructed MR images by SwinMR (nPI) with the different channel numbers and zero-filled MR images (ZF); Row 2: Absolute differences (10×) between reconstructed MR images by SwinMR (nPI) and ground truth MR images (GT), and absolute differences (10×) between ZF and GT; Row 3: Reconstructed MR images by SwinMR (PI) with the different channel number and GT; Row 4: Absolute differences (10×) between reconstructed MR images by SwinMR (PI) and GT, and the Gaussian 1D 30% mask.

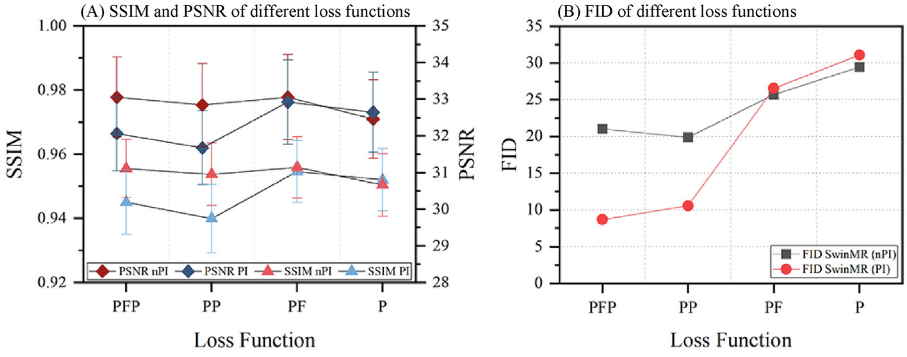


Fig. 16. Structural similarity index (SSIM), Peak signal-to-noise ratio (PSNR) and Fréchet inception distance (FID) of the ablation experiment on the loss function using Gaussian 1D 30% mask. PFP: pixel-wise, frequency and perceptual loss; PP: pixel-wise and perceptual loss; PF: pixel-wise and frequency loss; P: only pixel-wise loss.

wise and **P**erceptual loss; (3) PF: **P**ixel-wise and **F**requency loss; (4) P: only **P**ixel-wise loss.

Fig. 16 shows the SSIM, PSNR and FID of SwinMR trained with different loss functions. Fig. 17 displays the samples of reconstructed images of SwinMR trained with different loss functions.

According to Fig. 16, for SwinMR (PI), the utilisation of frequency loss tends to improve SSIM/PSNR and decreases the FID (PFP vs PP; PF vs P). For SwinMR (nPI), the utilisation of frequency loss leads to improvement only on SSIM and PSNR, but scarcely on FID. In most cases, the utilisation of the frequency loss has a positive impact on reconstruction quality metrics – both SSIM/PSNR and FID.

For SwinMR (PI), the utilisation of perceptual loss tends to slightly decrease SSIM and PSNR, but substantially decreases the FID (PFP vs PF; PP vs P). For SwinMR (nPI), the utilisation of perceptual loss tends to achieve a better FID but scarcely change SSIM and PSNR (PFP vs PF; PP vs P). In most cases, the utilisation of the perceptual loss has a positive impact on FID, but a negative impact on SSIM/PSNR when using multi-channel data.

3.9. Downstream task experiments: brain segmentation experiments on BraTS17 dataset

In this experiment, we performed a downstream task using a reconstructed MR image, in order to measure the reconstruction quality. Specifically, we chose an open-access multi-modalities brain tumour segmentation network[3] [73] for the downstream task

[3] https://github.com/Mehrdad-Noori/Brain-Tumor-Segmentation.

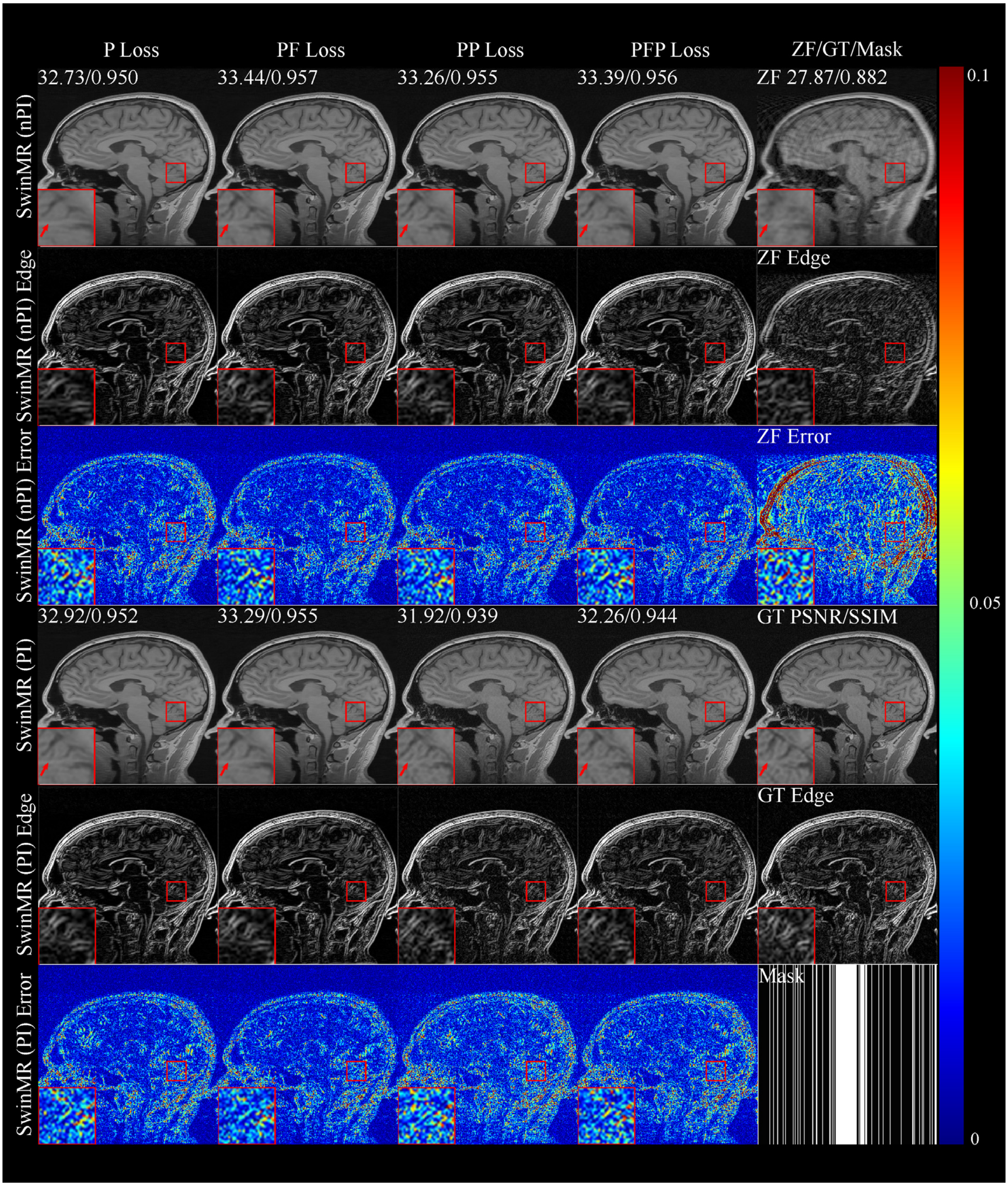


Fig. 17. Samples of the ablation experiment on the loss function using Gaussian 1D 30% mask. PFP: pixel-wise, frequency and perceptual loss; PP: pixel-wise and perceptual loss; PF: pixel-wise and frequency loss; P: only pixel-wise loss. Row 1: Reconstructed MR images by SwinMR (nPI) and zero-filled MR images (ZF); Row 2: Edge information of reconstructed MR images by SwinMR (nPI) and edge information of ZF; Row 3: Absolute differences (10×) between reconstructed MR images by SwinMR (nPI) and ground truth MR images (GT), and absolute differences (10×) between ZF and GT; Row 4: Reconstructed MR images by SwinMR (PI) and GT; Row 5: Edge information of reconstructed MR images by SwinMR (PI) and edge information of GT; Row 6: Absolute differences (10×) between reconstructed MR images by SwinMR (PI) and GT, and the Gaussian 1D 30% mask.

experiments. This segmentation network adopted a U-Net [62] based architecture with the utilisation of residual blocks and strided convolution downsampling compared to the vanilla U-Net. In addition, this segmentation network also employed the Squeeze-and-Excitation Block [74] on concatenated multi-level features for channel attention mechanism.

The segmentation network was trained on the BraTS17 dataset (four modalities are required including FLAIR, T1, T1CE and T2). Then, we trained four SwinMR weights using BraTS17 FLAIR, T1, T1CE and T2 data, respectively. After that, segmentation tasks were conducted on GT MR images, SwinMR reconstructed MR images and ZF MR images directly using the pre-trained segmentation net-

Table 4
Quantitative results of reconstructed images by SwinMR (Recon) and zero-filled images (ZF) on BraTS17 dataset (mean (std)). PSNR: Peak signal-to-noise ratio; SSIM: Structural similarity index; FID: Fréchet inception distance. G1D10%: Gaussian 1D 10% mask; G1D30%: Gaussian 1D 30% mask.

Mask	Metrics	Recon			
		FLAIR	T1	T1CE	T2
G1D10%	PSNR	30.07 (1.99)	33.80 (2.30)	33.80 (1.84)	32.20 (1.81)
	SSIM	0.751 (0.043)	0.760 (0.046)	0.797 (0.049)	0.745 (0.039)
	FID	38.02	32.97	31.46	21.84
G1D30%	PSNR	37.97 (2.42)	41.08 (3.36)	42.29 (2.12)	38.37 (2.02)
	SSIM	0.942 (0.013)	0.953 (0.012)	0.953 (0.015)	0.937 (0.016)
	FID	5.94	4.80	4.39	8.95

Mask	Metrics	ZF			
		FLAIR	T1	T1CE	T2
G1D10%	PSNR	23.87 (1.64)	25.92 (1.48)	25.92 (1.70)	23.92 (1.79)
	SSIM	0.388 (0.070)	0.414 (0.061)	0.414 (0.068)	0.431 (0.057)
	FID	225.70	234.52	227.51	219.09
G1D30%	PSNR	28.74 (1.78)	28.82 (1.60)	30.60 (1.82)	29.46 (2.01)
	SSIM	0.597 (0.046)	0.602 (0.051)	0.602 (0.051)	0.632 (0.038)
	FID	91.18	100.98	106.28	85.49

work. Ideally, the segmentation score of reconstructed images and GT images should be as closer as possible.

Table 4 shows the result of SwinMR trained with BraTS17 FLAIR, T1, T1CE and T2 respectively. Fig. 18 displays the samples of the reconstruction of different modalities. Tables 5 and 6 show the IoU and Dice score of the segmentation task. Fig. 19 displays the sample of the segmentation task.

According to Tables 5 and 6, the IoU and Dice score of reconstructed MR images are improved compared with ZF MR images and much closer to the score of GT MR images. According to the Mann–Whitney Test, the IoU and Dice score distributions of the reconstructed MR images using the Gaussian 1D 30% mask are not significantly different from the distributions of the GT MR images ($p > 0.05$).

4. Discussion

In this work, a novel Swin transformer based model, i.e., SwinMR, for fast MRI reconstruction has been proposed. Most existing deep learning based image restoration methods, including MRI reconstruction approaches, are based on CNNs. The convolution is a very effective feature extractor but lacks long-range dependency. The receptive field of CNNs is limited by the size of the kernel and the depth of the network. To tackle this problem, researchers have developed transformers based image restoration methods that have been originally used for solving NLP tasks. The core of the transformer is MSA, which has global sensitivity. In MSA operation, each patch can link with any other patches in the whole image space but also aggravates the computational burden.

However, we have believed that in MRI reconstruction, the MSA, which is operated in the whole image space, is redundant and not necessary. It is not difficult to understand that in NLP tasks the first and the last words may have a strong connection in a sentence. However, this may not be applicable in CV tasks. Visual elements (e.g., pixels) in CV tasks can vary substantially in scale unlike language elements (e.g., word tokens) in NLP tasks [36]. Since in most cases, for example, the top-left corner patch has no relationship with the bottom-right corner patch within an image. Moreover, for MRI reconstruction, the biggest difficulty is the recovery of detailed information and texture information. Focusing too much on global information and ignoring the detailed (local) information may make the image smoother and lose more details. The utilisation of a Swin transformer can achieve a trade-off for CV tasks. In Swin transformer, operations are conducted in shifted windows instead of the whole images. It has a larger receptive field compared to CNNs but is not overly concerned with global information. This is the reason why we have developed a Swin transformer for MRI reconstruction.

To evaluate our proposed methods, several comparison experiments and ablation studies have been conducted. In this study, we have compared our proposed SwinMR with benchmark MRI reconstruction methods. The results in Table 1 have demonstrated that our SwinMR has achieved the highest SSIM/PSNR and lowest FID compared to CNN-based and GAN-based models. From Fig. 4, we have shown clearly that our SwinMR has obtained better reconstruction quality, especially in the zoom-in area, where the details of the cerebellum have been well-preserved.

In this study, we have also compared SwinMR (PI) that has been trained with multi-channel brain data with SwinMR (nPI) that has been trained with single-channel brain data. The results have led to a similar conclusion in our previous study [56], where FID of the model trained with multi-channel data has been better compared to the model trained with single-channel data, and the SSIM/PSNR has shown the opposite (i.e., SSIM/PSNR: nPI > PI; FID: PI < nPI). This phenomenon can also be observed in the subsequent ablation experiments. From Fig. 4, we can find that the reconstructed images of SwinMR (PI) have shown more details and texture information, but the reconstructed images of SwinMR (nPI) have shown smoother.

The experimental results have demonstrated that the three metrics that compared PI and nPI gave different answers. We have speculated that this might be due to the different principles of these metrics. PSNR is a classic metric based on per-pixel comparisons, which are not able to reflect the structure information for images. SSIM is a perceptual metric that measures structure similarity. However, both of them are based on simple and shallow functions and direct comparisons between images, which is insufficient to account for many nuances of human perception [71]. For FID, the comparison is based on perception and performed on two sets of images. Images are mapped to high-dimension representations by a pre-trained InceptionV3 network, which is well-related to human visual perception. The SwinMR (PI) reconstructed images have demonstrated more details and texture information. Even though these details and texture information may not be so *accurate*, they make the reconstructed images more *visually similar* with

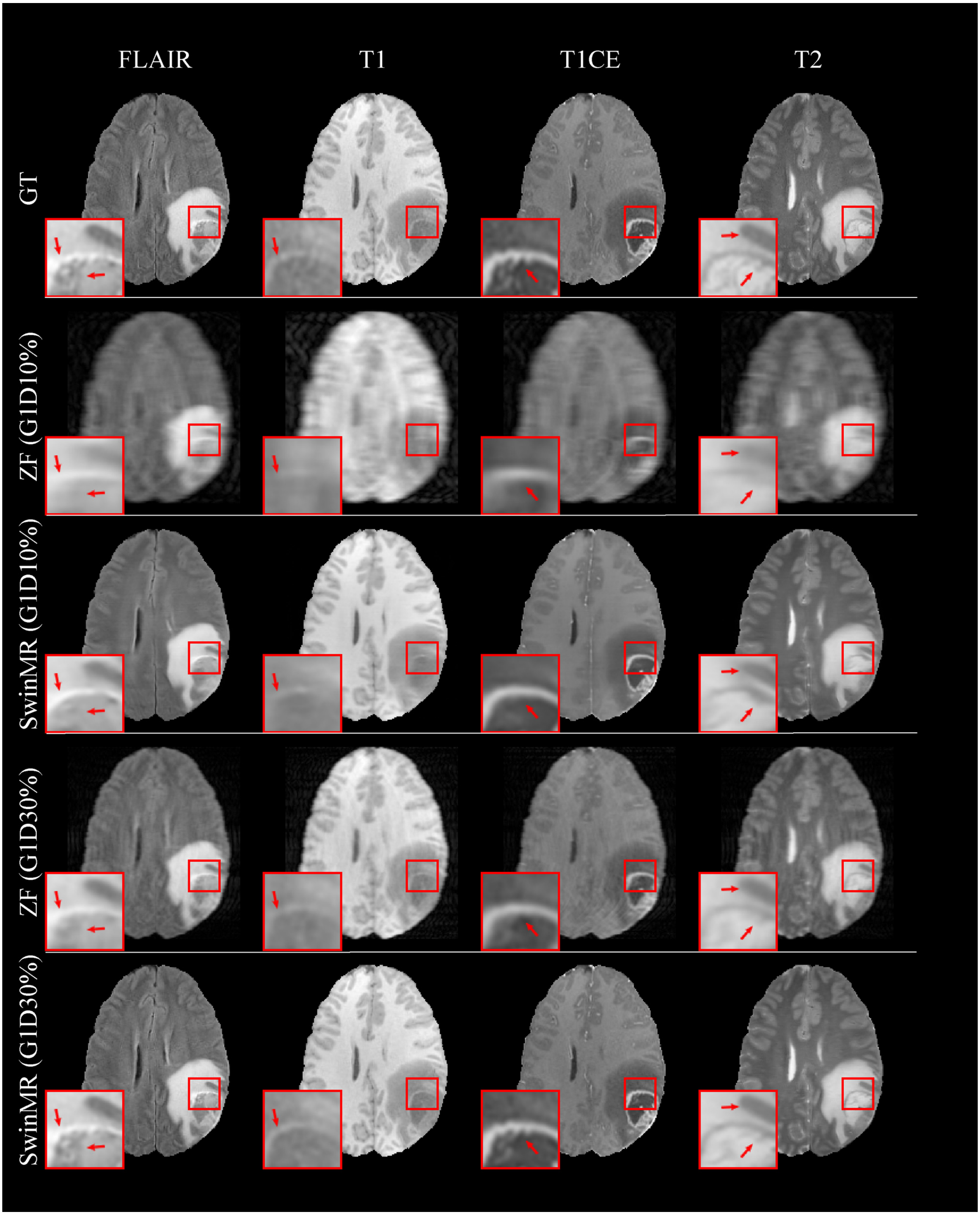


Fig. 18. Samples of reconstruction results for SwinMR on BraTS17 dataset including FLAIR, T1, T1CE and T2 MR images. Row 1: Ground truth MR images (GT); Row 2: Zero-filled MR images (ZF) undersampled by Gaussian 1D 10% mask (G1D10%); Row 3: Reconstructed MR images undersampled by G1D10%; Row 4: ZF undersampled by Gaussian 1D 30% mask (G1D30%); Row 5: Reconstructed MR images undersampled by G1D30%.

the ground truth images. However, the SwinMR (nPI) reconstructed images have shown smoother in pixel-wise scale, at the cost of less detail and texture information. Therefore, SwinMR (PI) have tended to have better FID and worse SSIM/PSNR compared to SwinMR (nPI), due to the principle differences of the evaluation methods.

Table 5
Intersection over union (IoU) of the segmentation experiment (median/mean [Q1,Q3]). *: $p < 0.05$; **: $p < 0.01$ (compared with GT by Mann–Whitney Test). GT: ground truth MR images; Recon: reconstructed MR images by SwinMR; ZF: undersampled zero-filled MR images. G1D10%: Gaussian 1D 10% mask; G1D30%: Gaussian 1D 30% mask. WT: Whole tumour; TC: Enhancing tumour; ET: Tumour core.

IoU		GT	Recon	ZF
G1D10%	WT	0.930/0.924 [0.900,0.954]	0.898/0.899 [0.868,0.940]**	0.838/0.836 [0.795,0.881]**
	TC	0.821/0.771 [0.726,0.903]	0.758/0.722 [0.661,0.890]**	0.617/0.539 [0.393,0.733]**
	ET	0.772/0.735 [0.625,0.889]	0.740/0.652 [0.471,0.846]**	0.570/0.527 [0.336,0.694]**
G1D30%	WT	0.930/0.924 [0.900,0.954]	0.924/0.921 [0.895,0.953]	0.897/0.897 [0.862,0.945]**
	TC	0.821/0.771 [0.726,0.903]	0.811/0.766 [0.719,0.904]	0.763/0.728 [0.669,0.895]**
	ET	0.772/0.735 [0.625,0.889]	0.770/0.725 [0.616,0.883]	0.748/0.697 [0.573,0.859]**

Table 6
Dice score of the segmentation experiment (median/mean [Q1,Q3]). *: $p < 0.05$; **: $p < 0.01$ (compared with GT by Mann–Whitney Test). GT: ground truth MR images; Recon: reconstructed MR images by SwinMR; ZF: undersampled zero-filled MR images. G1D10%: Gaussian 1D 10% mask; G1D30%: Gaussian 1D 30% mask. WT: Whole tumour; TC: Enhancing tumour; ET: Tumour core.

Dice		GT	Recon	ZF
G1D10%	WT	0.968/0.965 [0.952,0.981]	0.950/0.950 [0.933,0.974]**	0.916/0.914 [0.892,0.940]**
	TC	0.904/0.857 [0.845,0.951]	0.863/0.819 [0.800,0.944]**	0.767/0.653 [0.566,0.847]**
	ET	0.874/0.835 [0.777,0.941]	0.852/0.766 [0.640,0.917]**	0.725/0.665 [0.503,0.820]**
G1D30%	WT	0.968/0.965 [0.952,0.981]	0.964/0.963 [0.948,0.980]	0.949/0.949 [0.930,0.975]**
	TC	0.904/0.857 [0.845,0.951]	0.897/0.854 [0.838,0.951]	0.868/0.826 [0.803,0.947]**
	ET	0.874/0.835 [0.777,0.941]	0.871/0.827 [0.765,0.939]	0.857/0.808 [0.729,0.925]**

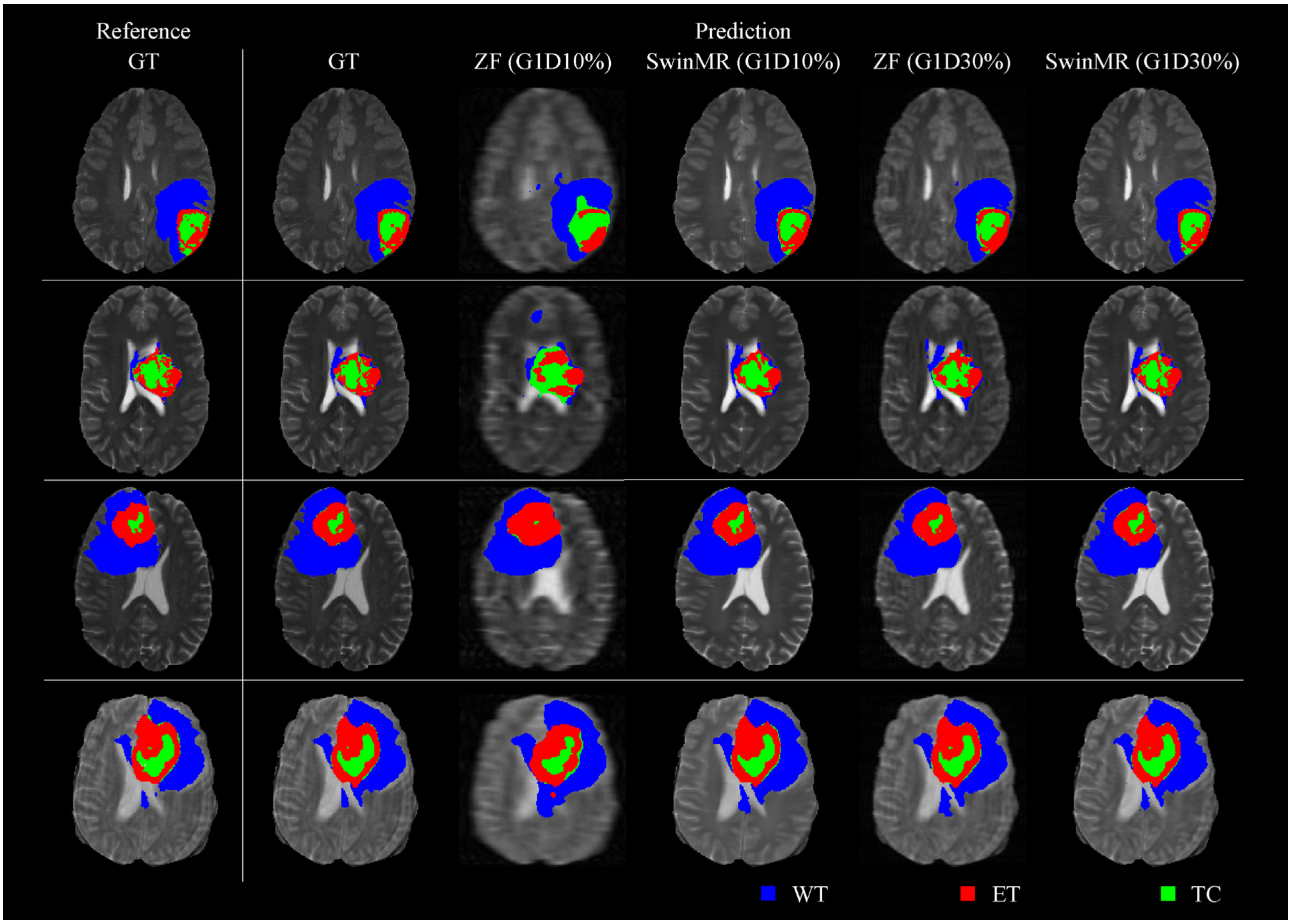


Fig. 19. Samples of segmentation results for SwinMR on the BraTS17 dataset. Col 1: Segmentation reference; Col 2: Segmentation prediction using GT images; Col 3: Segmentation prediction using zero-filled MR images (ZF) undersampled by Gaussian 1D 10% mask (G1D10%); Col 4: Segmentation prediction using reconstructed MR images undersampled by G1D10%; Col 5: Segmentation prediction using ZF undersampled by Gaussian 1D 30% mask (G1D30%); Col 6: Segmentation prediction using reconstructed MR images undersampled by G1D30%. Blue area: Whole tumour (WT); Red area: Enhancing tumour (ET); Green area: Tumour core (TC).

From Table 1, we can find a common problem of transformer-based methods, which is the higher computational cost compared to other CNN-based and GAN-based methods. Eq. (20) have shown that the computational complexity is proportional to the *HW* of the input of (S)W-MSA. The time shown in Table 1 has been the inference time, where the original height and weight have been treated as *H* and *W* (256 × 256 here). For training, randomly cropping have been applied to ease the long processing time.

Experiments using different undersampling masks with various noise levels have demonstrated that our proposed method SwinMR have shown superiority to DAGAN in all the tests. The evaluation metrics change as expected when the condition changes (different masks and noise levels).

Ablation studies on the patch number and the channel number have demonstrated that reconstruction quality has been improved as the patch number has been increased and has gradually been saturated, according to Fig. 13(A) and (C). However, according to Eq. (20), the computational complexity also has been increased as the patch number has been increased. As a trade-off, we have set the patch number to 96. Beyond our expectations, the changing of channel number has not been positively correlated with the evaluation metrics in this experiment, according to Fig. 13(B) and (D). We have assumed that the evaluation metrics have saturated in the range of channel number in this experiment. Empirically, we have set the channel number to 180 according to the default setting of SwinIR.

Ablation studies on different loss functions have been conducted. As expected, the utilisation of the pixel-wise loss and the frequency loss has mainly constrained the fidelity of reconstruction, and the utilisation of perceptual VGG loss has focused on perception, which has been well-related to the human visual system. Therefore, the utilisation of frequency loss has had a positive impact on SSIM and PSNR, which has been more sensitive to the fidelity of reconstruction. The utilisation of perceptual loss has had a positive impact on FID, which has been based on perception.

There are still some limitations of our work. First, in the (S)W-MSA operation, the size of windows is fixed. Inspired by GoogleNet, multi-scale windows could be incorporated and results from different scales could be merged in the (S)W-MSA. Second, the heavy computational cost is still an obstacle to the development of transformers. The improvement that transformers bring is at the sacrifice of increased computational cost. A lightweight transformer model could be a potential future research direction.

5. Conclusion

In this work, we have developed the SwinMR, a novel parallel imaging coupled Swin transformer-based model for fast multi-channel MRI reconstruction. The proposed method has outperformed other benchmark CNN-based and GAN-based MRI reconstruction methods. It has also shown excellent robustness using different undersampling trajectories with various noises.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the UK Research and Innovation Future Leaders Fellowship [MR/V023799/1], in part by the Medical Research Council [MC/PC/21013], in part by the European Research Council Innovative Medicines Initiative [DRAGON, H2020-JTI-IMI2 101005122], in part by the AI for Health Imaging Award [CHAIMELEON, H2020-SC1-FA-DTS-2019-1 952172], in part by the British Heart Foundation [Project Number: TG/18/5/34111, PG/16/78/32402], in part by the NVIDIA Academic Hardware Grant Program, in part by the Project of Shenzhen International Cooperation Foundation [GJHZ20180926165402083], in part by the Basque Government through the ELKARTEK funding program [KK-2020/00049], and in part by the consolidated research group MATHMODE [IT1294-19].

References

[1] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M.J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, M. Parente, K.J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, N. Yakubova, J. Pinkerton, D. Wang, E. Owens, C.L. Zitnick, M.P. Recht, D.K. Sodickson, Y.W. Lui, FastMRI: An open dataset and benchmarks for accelerated MRI, arXiv e-prints (2018) arXiv:1811.08839.

[2] M.K. Stehling, R. Turner, P. Mansfield, Echo-planar imaging: Magnetic resonance imaging in a fraction of a second, Science 254 (5028) (1991) 43–50.

[3] J. Hennig, A. Nauerth, H. Friedburg, RARE imaging: A fast imaging method for clinical MR, Magn. Reson. Med. 3 (6) (1986) 823–833, https://doi.org/10.1002/mrm.1910030602.

[4] M. Blaimer, F. Breuer, M. Mueller, R.M. Heidemann, M.A. Griswold, P.M. Jakob, SMASH, SENSE, PILS, GRAPPA: How to choose the optimal method, Top. Magn. Reson. Imaging 15 (4) (2004) 223–236.

[5] D.K. Sodickson, W.J. Manning, Simultaneous acquisition of spatial harmonics (SMASH): Fast imaging with radiofrequency coil arrays, Magn. Reson. Med. 38 (4) (1997) 591–603.

[6] K.P. Pruessmann, M. Weiger, M.B. Scheidegger, P. Boesiger, SENSE: Sensitivity encoding for fast MRI, Magnetic Resonance in Medicine: An Official Journal of the International Society for, Magn. Reson. Med. 42 (5) (1999) 952–962.

[7] M.A. Griswold, P.M. Jakob, R.M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, A. Haase, Generalized autocalibrating partially parallel acquisitions (GRAPPA), Magn. Reson. Med. 47 (2002) 1202–1210, https://doi.org/10.1002/mrm.10171.

[8] D. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (2006) 1289–1306, https://doi.org/10.1109/TIT.2006.871582.

[9] K.T. Block, M. Uecker, J. Frahm, Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint, Magn. Resonance Med. 57 (6) (2007) 1086–1098.

[10] M. Beladgham, I.B. Hacene, A. Taleb-Ahmed, M. Khélif, MRI images compression using curvelets transforms, in: AIP Conference Proceedings, Vol. 1019, American Institute of Physics, 2008, pp. 249–253.

[11] Z. Zhu, K. Wahid, P. Babyn, R. Yang, Compressed sensing-based MRI reconstruction using complex double-density dual-tree DWT, J. Biomed. Imaging (2013 (2013).), https://doi.org/10.1155/2013/907501.

[12] S. Ravishankar, Y. Bresler, MR image reconstruction from highly undersampled k-space data by dictionary learning, IEEE Trans. Med. Imaging 30 (5) (2011) 1028–1041, https://doi.org/10.1109/TMI.2010.2090538.

[13] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, X. Liu, An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips, IEEE Trans. Nanotechnol. 18 (2019) 819–829, https://doi.org/10.1109/TNANO.2019.2932271.

[14] N. Zeng, H. Li, Y. Peng, A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease, Neural Comput. Appl. (2021).

[15] P. Wu, H. Li, N. Zeng, F. Li, FMD-Yolo: An efficient face mask detection method for COVID-19 prevention and control in public, Image Vis. Comput. 117 (2022) 104341.

[16] Y. Chen, C.-B. Schönlieb, P. Liò, T. Leiner, P.L. Dragotti, G. Wang, D. Rueckert, D. Firmin, G. Yang, AI-based reconstruction for fast MRI-a systematic review and meta-analysis, Proceedings of the IEEE 110 (2) (2022) 224–245.

[17] M. Bakator, D. Radosav, Deep learning and medical diagnosis: A review of literature, Multimodal Technol. Interaction 2 (3) (2018).

[18] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[20] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[21] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 295–307, https://doi.org/10.1109/TPAMI.2015.2439281.

[22] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, D. Liang, Accelerating magnetic resonance imaging via deep learning, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016, pp. 514–517, https://doi.org/10.1109/ISBI.2016.7493320.

[23] Y. Yang, J. Sun, H. Li, Z. Xu, Deep ADMM-Net for compressive sensing MRI, in: Advances in Neural Information Processing Systems, Vol. 29, Curran Associates Inc, 2016.

[24] J. Schlemper, J. Caballero, J.V. Hajnal, A.N. Price, D. Rueckert, A deep cascade of convolutional neural networks for dynamic MR image reconstruction, IEEE Trans. Med. Imaging 37 (2018) 491–503, https://doi.org/10.1109/TMI.2017.2760978.

[25] B. Zhu, J.Z. Liu, S.F. Cauley, B.R. Rosen, M.S. Rosen, Image reconstruction by domain-transform manifold learning, Nature 555 (2018) 487–492, https://doi.org/10.1038/nature25988.

[26] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, Vol. 27, Curran Associates Inc, 2014.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[28] A.P. Parikh, O. Täckström, D. Das, J. Uszkoreit, A Decomposable Attention Model for Natural Language Inference, arXiv e-prints (2016) arXiv:1606.01933.

[29] J. Cheng, L. Dong, M. Lapata, Long Short-Term Memory-Networks for Machine Reading, arXiv e-prints (2016) arXiv:1601.06733.

[30] C. Matsoukas, J. Fredin Haslum, M. Söderberg, K. Smith, Is it time to replace CNNs with transformers for medical images?, arXiv e-prints (2021) arXiv:2108.09038.

[31] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, in: Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 4055–4064.

[32] T. Salimans, A. Karpathy, X. Chen, D.P. Kingma, PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications, arXiv e-prints (2017) arXiv:1701.05517.

[33] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Science China Technological Sciences (2020) 1–26.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 213–229.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv e-prints (2020) arXiv:2010.11929.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, arXiv e-prints (2021) arXiv:2103.14030.

[37] G. Yang, S. Yu, H. Dong, G. Slabaugh, P.L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, D. Firmin, DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction, IEEE Trans. Med. Imaging 37 (2018) 1310–1321, https://doi.org/10.1109/TMI.2017.2785879.

[38] H.-C. Shin, A. Ihsani, S. Mandava, S. Turuvekere Sreenivas, C. Forster, J. Cha, A. Disease Neuroimaging Initiative, GANBERT: Generative adversarial networks with bidirectional encoder representations from transformers for MRI to PET synthesis, arXiv e-prints (2020) arXiv:2008.04393.

[39] X. Zhang, X. He, J. Guo, N. Ettehadi, N. Aw, D. Semanek, J. Posner, A. Laine, Y. Wang, PTNet: A high-resolution infant MRI synthesizer based on transformer, arXiv e-prints (2021) arXiv:2105.13993.

[40] O. Dalmaz, M. Yurt, T. Çukur, ResViT: Residual vision transformers for multi-modal medical image synthesis, arXiv e-prints (2021) arXiv:2106.16031.

[41] Y. Korkmaz, M. Yurt, S.U.H. Dar, T. Cukur, Deep MRI reconstruction with generative vision transformers, in: Machine Learning for Medical Image Reconstruction, Springer International Publishing, Cham, 2021, pp. 54–64.

[42] Y. Korkmaz, S.U. Dar, M. Yurt, M. Özbey, T. Çukur, Unsupervised MRI reconstruction via zero-shot learned adversarial transformers, IEEE Trans. Med. Imaging (2022), https://doi.org/10.1109/TMI.2022.3147426, 1–1.

[43] C.-M. Feng, Y. Yan, H. Fu, L. Chen, Y. Xu, Task transformer network for joint MRI reconstruction and super-resolution, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 307–317.

[44] C.-M. Feng, Y. Yan, G. Chen, H. Fu, Y. Xu, L. Shao, Accelerated multi-modal MR imaging with transformers, arXiv e-prints (2021) arXiv:2106.14248.

[45] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, SwinIR: Image restoration using swin transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Workshops, 2021, pp. 1833–1844.

[46] X. Qu, D. Guo, B. Ning, Y. Hou, Y. Lin, S. Cai, Z. Chen, Undersampled MRI reconstruction with patch-based directional wavelets, Magnetic resonance imaging 30 (7) (2012) 964–977.

[47] T. Wu, D.Z. Wang, Z. Jin, J. Zhang, Solving constrained TV2L1-L2 MRI signal reconstruction via an efficient alternating direction method of multipliers, Numerical Mathematics: Theory, Methods and Applications 10 (4) (2017) 895–912.

[48] S. Ravishankar, Y. Bresler, MR image reconstruction from highly undersampled k-space data by dictionary learning, IEEE Trans. Med. Imaging 30 (5) (2010) 1028–1041.

[49] M. Lustig, D. Donoho, J.M. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging, Magnetic Resonance in Medicine: An Official Journal of the International Society for, Magn. Reson. Med. 58 (6) (2007) 1182–1195.

[50] J. Yang, Y. Zhang, W. Yin, A fast alternating direction method for TVL1-L2 signal reconstruction from partial fourier data, IEEE J. Sel. Top. Signal Process. 4 (2) (2010) 288–297.

[51] L. Wang, K. Lu, P. Liu, Compressed sensing of a remote sensing image based on the priors of the reference image, IEEE Geosci. Remote Sens. Lett. 12 (4) (2014) 736–740.

[52] J.-F. Cai, J.K. Choi, K. Wei, Data driven tight frame for compressed sensing MRI reconstruction via off-the-grid regularization, SIAM J. Imag. Sci. 13 (3) (2020) 1272–1301.

[53] G. Yang, J. Lv, Y. Chen, J. Huang, J. Zhu, Generative Adversarial Network Powered Fast Magnetic Resonance Imaging—Comparative Study and New Perspectives, Springer International Publishing, Cham, 2022, pp. 305–339.

[54] J. Lv, J. Zhu, G. Yang, Which GAN? A comparative study of generative adversarial network-based fast MRI reconstruction, Philos. Trans. R. Soc. A 379 (2021) 20200203, https://doi.org/10.1098/rsta.2020.0203.

[55] R. Shaul, I. David, O. Shitrit, T.R. Raviv, Subsampled brain MRI reconstruction by generative adversarial neural networks, Med. Image Anal. 65 (2020), https://doi.org/10.1016/j.media.2020.101747 101747.

[56] J. Huang, W. Ding, J. Lv, J. Yang, H. Dong, J. Del Ser, J. Xia, T. Ren, S. Wong, G. Yang, Edge-enhanced dual discriminator generative adversarial network for fast MRI with parallel imaging using multi-view information, Appl. Intell. (2021), https://doi.org/10.1007/s10489-021-03092-w.

[57] T.M. Quan, T. Nguyen-Duc, W.-K. Jeong, Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss, IEEE Trans. Med. Imaging 37 (2018) 1488–1497, https://doi.org/10.1109/TMI.2018.2820120.

[58] Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, Y. Zhao, Structure and illumination constrained gan for medical image enhancement, IEEE Trans. Med. Imaging (2021), https://doi.org/10.1109/TMI.2021.3101937, 1–1.

[59] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 214–223.

[60] Y. Guo, C. Wang, H. Zhang, G. Yang, Deep attentive wasserstein generative adversarial networks for MRI reconstruction with recurrent context-awareness, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham, 2020, pp. 167–177.

[61] M. Jiang, M. Zhi, L. Wei, X. Yang, J. Zhang, Y. Li, P. Wang, J. Huang, G. Yang, FA-GAN: Fused attentive generative adversarial networks for MRI image super-resolution, Comput. Med. Imaging Graph. 92 (2021), https://doi.org/10.1016/j.compmedimag.2021.101969 101969.

[62] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[63] J. Lv, G. Li, X. Tong, W. Chen, J. Huang, C. Wang, G. Yang, Transfer learning enhanced generative adversarial networks for multi-channel MRI reconstruction, Comput. Biol. Med. 134 (2021), https://doi.org/10.1016/j.compbiomed.2021.104504 104504.

[64] M. Uecker, P. Lai, M.J. Murphy, P. Virtue, M. Elad, J.M. Pauly, S.S. Vasanawala, M. Lustig, ESPIRiT-an eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA, Magn. Reson. Med. 71 (3) (2014) 990–1001.

[65] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Fast and accurate image Super-Resolution with deep laplacian pyramid networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2019) 2599–2613, https://doi.org/10.1109/TPAMI.2018.2865304.

[66] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, R. Lotufo, An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement, NeuroImage 170 (2018) 482–494, segmenting the Brain. doi: 10.1016/j.neuroimage.2017.08.021.

[67] 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 Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[68] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, J.B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, Scientific Data 4 (1) (2017) 1–13.

[69] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. Takeshi Shinohara, C. Berger, S.M. Ha, M. Rozycki, M. Prastawa, E. Alberts, J. Lipkova, J. Freymann, J. Kirby, M. Bilello, H. Fathallah-Shaykh, R. Wiest, J. Kirschke, B. Wiestler, R. Colen, A. Kotrotsou, P. Lamontagne, D. Marcus, M. Milchenko, A. Nazeri, M.-A. Weber, A. Mahajan, U. Baid, E. Gerstner, D. Kwon, G. Acharya, M. Agarwal, M. Alam, A. Albiol, A. Albiol, F.J. Albiol, V. Alex, N. Allinson, P.H.A. Amorim, A. Amrutkar, G. Anand, S. Andermatt, T. Arbel, P. Arbelaez, A. Avery, M. Azmat, B. Pranjal, W. Bai, S. Banerjee, B. Barth, T. Batchelder, K. Batmanghelich, E. Battistella, A. Beers, M. Belyaev, M. Bendszus, E. Benson, J. Bernal, H. Nagaraja Bharath, G. Biros, S. Bisdas, J. Brown, M. Cabezas, S. Cao, J. M. Cardoso, E.N. Carver, A. Casamitjana, L. Silvana Castillo, M. Catà, P. Cattin, A. Cerigues, V.S. Chagas, S. Chandra, Y.-J. Chang, S. Chang, K. Chang, J. Chazalon, S. Chen, W. Chen, J.W. Chen, Z. Chen, K. Cheng, A.R. Choudhury, R. Chylla, A.

Clérigues, S. Colleman, R. German Rodriguez Colmeiro, M. Combalia, A. Costa, X. Cui, Z. Dai, L. Dai, L.A. Daza, E. Deutsch, C. Ding, C. Dong, S. Dong, W. Dudzik, Z. Eaton-Rosen, G. Egan, G. Escudero, T. Estienne, R. Everson, J. Fabrizio, Y. Fan, L. Fang, X. Feng, E. Ferrante, L. Fidon, M. Fischer, A.P. French, N. Fridman, H. Fu, D. Fuentes, Y. Gao, E. Gates, D. Gering, A. Gholami, W. Gierke, B. Glocker, M. Gong, S. González-Villá, T. Grosges, Y. Guan, S. Guo, S. Gupta, W.-S. Han, I.S. Han, K. Harmuth, H. He, A. Hernández-Sabaté, E. Herrmann, N. Himthani, W. Hsu, C. Hsu, X. Hu, X. Hu, Y. Hu, Y. Hu, R. Hua, T.-Y. Huang, W. Huang, S. Van Huffel, Q. Huo, V. HV, K.M. Iftekharuddin, F. Isensee, M. Islam, A.S. Jackson, S.R. Jambawalikar, A. Jesson, W. Jian, P. Jin, V.J.M. Jose, A. Jungo, B. Kainz, K. Kamnitsas, P.-Y. Kao, A. Karnawat, T. Kellermeier, A. Kermi, K. Keutzer, M. Tarek Khadir, M. Khened, P. Kickingereder, G. Kim, N. King, H. Knapp, U. Knecht, L. Kohli, D. Kong, X. Kong, S. Koppers, A. Kori, G. Krishnamurthi, E. Krivov, P. Kumar, K. Kushibar, D. Lachinov, T. Lambrou, J. Lee, C. Lee, Y. Lee, M. Lee, S. Lefkovits, L. Lefkovits, J. Levitt, T. Li, H. Li, W. Li, H. Li, X. Li, Y. Li, H. Li, Z. Li, X. Li, Z. Li, X. Li, W. Li, Z.-S. Lin, F. Lin, P. Lio, C. Liu, B. Liu, X. Liu, M. Liu, J. Liu, L. Liu, X. Llado, M. Moreno Lopez, P. Ribalta Lorenzo, Z. Lu, L. Luo, Z. Luo, J. Ma, K. Ma, T. Mackie, A. Madabushi, I. Mahmoudi, K.H. Maier-Hein, P. Maji, C. Mammen, A. Mang, B.S. Manjunath, M. Marcinkiewicz, S. McDonagh, S. McKenna, R. McKinley, M. Mehl, S. Mehta, R. Mehta, R. Meier, C. Meinel, D. Merhof, C. Meyer, R. Miller, S. Mitra, A. Moiyadi, D. Molina-Garcia, M.A.B. Monteiro, G. Mrukwa, A. Myronenko, J. Nalepa, T. Ngo, D. Nie, H. Ning, C. Niu, N.K. Nuechterlein, E. Oermann, A. Oliveira, D.D.C. Oliveira, A. Oliver, A.F.I. Osman, Y.-N. Ou, S. Ourselin, N. Paragios, M.S. Park, B. Paschke, J.G. Pauloski, K. Pawar, N. Pawlowski, L. Pei, S. Peng, S.M. Pereira, J. Perez-Beteta, V.M. Perez-Garcia, S. Pezold, B. Pham, A. Phophalia, G. Piella, G.N. Pillai, M. Piraud, M. Pisov, A. Popli, M.P. Pound, R. Pourreza, P. Prasanna, V. Prkovska, T.P. Pridmore, S. Puch, É. Puybareau, B. Qian, X. Qiao, M. Rajchl, S. Rane, M. Rebsamen, H. Ren, X. Ren, K. Revanuru, M. Rezaei, O. Rippel, L.C. Rivera, C. Robert, B. Rosen, D. Rueckert, M. Safwan, M. Salem, J. Salvi, I. Sanchez, I. Sánchez, H.M. Santos, E. Sartor, D. Schellingerhout, K. Scheufele, M.R. Scott, A.A. Scussel, S. Sedlar, J.P. Serrano-Rubio, N.J. Shah, N. Shah, M. Shaikh, B.U. Shankar, Z. Shboul, H. Shen, D. Shen, L. Shen, H. Shen, V. Shenoy, F. Shi, H.E. Shin, H. Shu, D. Sima, M. Sinclair, O. Smedby, J.M. Snyder, M. Soltaninejad, G. Song, M. Soni, J. Stawiaski, S. Subramanian, L. Sun, R. Sun, J. Sun, K. Sun, Y. Sun, G. Sun, S. Sun, Y.R. Suter, L. Szilagyi, S. Talbar, D. Tao, D. Tao, Z. Teng, S. Thakur, M.H. Thakur, S. Tharakan, P. Tiwari, G. Tochon, T. Tran, Y.M. Tsai, K.-L. Tseng, T.A. Tuan, V. Turlapov, N. Tustison, M. Vakalopoulou, S. Valverde, R. Vanguri, E. Vasiliev, J. Ventura, L. Vera, T. Vercauteren, C.A. Verrastro, L. Vidyaratne, V. Vilaplana, A. Vivekanandan, G. Wang, Q. Wang, C.J. Wang, W. Wang, D. Wang, R. Wang, Y. Wang, C. Wang, G. Wang, N. Wen, X. Wen, L. Weninger, W. Wick, S. Wu, Q. Wu, Y. Wu, Y. Xia, Y. Xu, X. Xu, P. Xu, T.-L. Yang, X. Yang, H.-Y. Yang, J. Yang, H. Yang, G. Yang, H. Yao, X. Ye, C. Yin, B. Young-Moxon, J. Yu, X. Yue, S. Zhang, A. Zhang, K. Zhang, X. Zhang, L. Zhang, X. Zhang, Y. Zhang, L. Zhang, J. Zhang, X. Zhang, T. Zhang, S. Zhao, Y. Zhao, X. Zhao, L. Zhao, Y. Zheng, L. Zhong, C. Zhou, X. Zhou, F. Zhou, H. Zhu, J. Zhu, Y. Zhuge, W. Zong, J. Kalpathy-Cramer, K. Farahani, C. Davatzikos, K. van Leemput, B. Menze, Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv e-prints (2018) arXiv:1811.02629.

[70] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30 (2017).

[71] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[72] M.S. Hansen, P. Kellman, Image reconstruction: An overview for clinicians, J. Magn. Reson. Imaging 41 (3) (2015) 573–585, https://doi.org/10.1002/jmri.24687.

[73] M. Noori, A. Bahri, K. Mohammadi, Attention-guided version of 2D UNet for automatic brain tumor segmentation, in: 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), 2019, pp. 269–275, https://doi.org/10.1109/ICCKE48569.2019.8964956.

[74] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

Jiahao Huang is a Ph.D. student at National Heart and Lung Institute, Imperial College London. He received his B.S. degree in Optoelectronics Information Science and Engineering from Beijing Institute of Technology in 2021. His research interests include computer vision, machine learning, and medical image processing and analysis.

Yingying Fang is a Postdoc Research Assistant in the National Heart and Lung Institute of Imperial College London. She obtained her Ph.D. degree from Hong Kong Baptist University in 2020. Her research interests include image restoration, optimisation and deep learning. Her research is currently focused on the medical image processing for lung disease diagnosis and prognosis.

Yinzhe Wu is an undergraduate student at the Department of Bioengineering of Imperial College London, and also an undergraduate research student under the supervision of Dr. Guang Yang at National Heart and Lung Institute of Imperial College London.

Huanjun Wu is an undergraduate student studying Molecular Bioengineering at the Department of Bioengineering, Imperial College London. She has joined Dr. Guang Yang's research group at National Heart and Lung Institute of Imperial College London since June 2021 as a UROP student, currently doing research on MRI reconstruction.

Zhifan Gao is an associate professor at School of Biomedical Engineering, Sun Yat-sen University (SYSU). Before joining SYSU, he was a postdoc at Schulich School of Medicine & Dentistry, Western University (UWO), Canada.

Yang Li is a Professor with the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing, China. His research area is involved system identification and modeling for complex nonlinear processes, signal processing and data modeling, image processing, and machine learning.

Javier Del Ser is a principal researcher in data analytics and optimization at TECNALIA (Spain), and a professor at the University of the Basque Country (UPV/EHU).

Jun Xia is a professor at the Department of Radiology, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University Health Science Center.

Dr Guang Yang is a Future Leaders Fellow (Tenured Senior Research Fellow) in the National Heart and Lung Institute at Imperial College London. He is also an Honorary Senior Lecturer in the School of Biomedical Engineering & Imaging Sciences at King's College London. His research group is interested in developing novel and translational techniques for imaging and biomedical data analysis. His group focuses on the research and development on data-driven fast imaging, data harmonisation, image segmentation, image synthesis, federated learning, explainable AI etc. He is currently working on a wide range of clinical applications in cardiovascular disease, lung disease and oncology. Read more information about Yang's Lab at: https://www.yanglab.fyi/.