Explanation-driven Cyclic Learning for High-Quality Brain MRI Reconstruction from Unknown Degradation

Ning Jiang[1,2], Zhengyong Huang[1,2], and Yao Sui[1,2(✉)]

[1] National Institute of Health Data Science, Peking University, Beijing, China
[2] Institute of Medical Technology, Peking University, Beijing, China
yaosui@pku.edu.cn

Abstract. Spatial resolution, signal-to-noise ratio (SNR), and motion artifacts critically matter in any Magnetic Resonance Imaging (MRI) practices. Unfortunately, it is difficult to achieve a trade-off between these factors. Scans with an increased spatial resolution require prolonged scan times and suffer from drastically reduced SNR. Increased scan time necessarily increases the potential of subject motion. Recently, end-to-end deep learning techniques have emerged as a post-acquisition method to deal with the above issues by reconstructing high-quality MRI images from various sources of degradation, such as motion, noise, and reduced resolution. However, those methods focus on a single known source of degradation, while multiple unknown sources of degradation commonly happen in a single scan. We aimed to develop a new methodology that enables high-quality MRI reconstruction from scans corrupted by a mixture of multiple unknown sources of degradation. We proposed a unified reconstruction framework based on explanation-driven cyclic learning. We designed an interpretation strategy for the neural networks, the Cross-Attention-Gradient (CAG), which generates pixel-level explanations from degraded images to enhance reconstruction with degradation-specific knowledge. We developed a cyclic learning scheme that comprises a front-end classification task and a back-end image reconstruction task, circularly shares knowledge between different tasks and benefits from multi-task learning. We assessed our method on three public datasets, including the real and clean MRI scans from 140 subjects with simulated degradation, and the real and motion-degraded MRI scans from 10 subjects. We identified 5 sources of degradation for the simulated data. Experimental results demonstrated that our approach achieved superior reconstructions in motion correction, SNR improvement, and resolution enhancement, as compared to state-of-the-art methods.

Keywords: MRI · Image reconstruction · Image-to-image translation · Transformer · Multi-task learning.

1 Introduction

Spatial resolution, signal-to-noise ratio (SNR), and motion artifacts severe influence the quality of Magnetic resonance imaging (MRI) data [1]. Efforts has

been made to offer a trade-off between between these factors, in order to achieve high-quality MRI acquisitions [2]. However, according to the MRI physics, we are unable to acquire the data with simultaneously high spatial resolution and SNR. High spatial resolution results in small pixel size, and in turn leads to reduced SNR. High-resolution scans prolong scan time as well, so consequently increase the potential of in-scanner motion. Literature has shown that the data quality is improved with various techniques, such parallel imaging [3], robust k-space sampling [4], and post-acquisition processing [5].

Deep learning has recently emerged as a post-acquisition technique to address the above issues by reconstructing high-quality MRI images from various sources of degradation [6]. Current deep learning-based methods rely on a single task-specific model that focus on a single known source of degradation for the reconstruction. However, the degradation of a single MRI scan is commonly from the mixture of multiple unknown sources. Therefore, a unified framework is desired to allow for high-quality reconstruction in the face of multiple unknown sources of degradation.

Degradation representations are important for the reconstruction. Previous methods extracted degradation-specific information to elaborate complex degradation patterns, such as degraded encoder [7, 8], and degradation prompt [9]. However, it is challenging to determine whether the extracted degradation representations are content-independent because the obtained high-quality images may benefit from a diverse feature space. This was identified and addressed in the latest work, such as reinforcing disease diagnosis [10] and semi-supervised segmentation [11, 12], by using pre-trained classifiers that generate category-specific explanations as guidance. Moreover, literature has recently shown that Transformer is highly effective for low-level vision tasks [13, 14]. The self-attention mechanism provides a straightforward explanation of the input. Zamir et al. [14] proposed a multi-Dconv head transposed attention (MDTA) module, that changes token-level self-attention to pixel-level, to guide the image reconstruction.

Inspired by these successes, we proposed to use an explanation-driven method to isolate the degradation knowledge from the image. We first pre-trained a degradation classifier with a Restormer [14] backbone by distinguishing MRI scans with different sources. The pre-trained classifier is then interpreted by our proposed Cross-Attention-Gradient (CAG) method, which integrates attention maps with gradients to generate pixel-level explanations of degradation sources. Similar to the saliency maps, the explanations quantify the contribution of each pixel to the classification results and represent the degradation-specific knowledge.

With the isolated degradation knowledge, we proposed a cyclic learning framework to circularly share knowledge between a front-end classification task and a back-end image reconstruction task. As illustrated in Fig .1 (a), the front-end generates degradation explanations to guide the back-end reconstruction and evaluates the semantic-level distance between the reconstructed and corresponding clean images. Meanwhile, the back-end reconstructs high-quality images with an encoder-decoder architecture, whose encoder slowly updates the isomorphic front-end classifier by an exponential moving average (EMA) method. Explanations,

losses, and parameter weights circularly flow between the front-end and back-end tasks, facilitating better optimization of the whole framework.

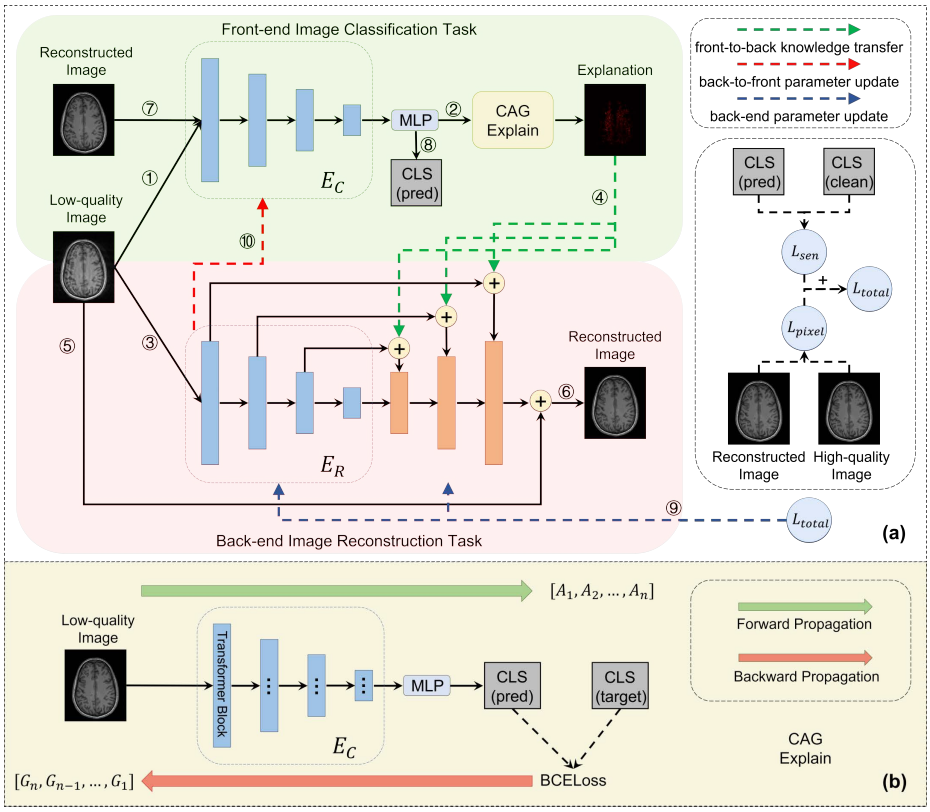

Fig. 1. Illustration of the proposed method. (a) Cyclic learning, circularly shares knowledge between the front-end task and the back-end task through losses, explanations, and weights until convergence. Steps 1 to 6 are executed during both the training and testing phases. In contrast, steps 7 to 10 are only executed during the training phase. (b) Cross-Attention-Gradient method, extracts attention maps and gradients through a forward/backward propagation.

We summarized our main contributions as follows:

(1) We proposed Cross-Attention-Gradient (CAG), an NN interpreting method to quantify the contribution of each pixel to the results of the classification of degradation sources. This approach takes into account both the attention information and the gradient information across layers and heads, providing transformers with pixel-level interpretability, which is suitable for image reconstruction tasks.

(2) We designed a cyclic learning framework to share knowledge between the front-end classification task and the back-end reconstruction task. This achieved

a combination of learning, interpreting, and reinforcing, with each task benefiting from the other.

Comprehensive experiments were conducted on both in vivo and in silico data. The proposed reconstruction framework demonstrated superior performance on three public datasets, as compared to four other competing methods.

2 Methods

We aim to reconstruct high-quality MRI from unknown sources of degradation. Interpreting a pre-trained degradation classifier reveals degradation-specific information that assists the image generator in performing degradation-aware reconstruction. We proposed a cyclic learning framework to share useful information contained in the front-end classification task and the back-end image reconstruction task.

We simulate different degradation sources (motion, low resolution, noise, motion and noise, and motion and low resolution) using a unified forward model that characterizes the process of MRI acquisition: $I_D = DHTI_C + \epsilon$. I_C and I_D denote the clean image and the degraded image; D denotes downsampling; H denotes the sampling operator; T denotes a rigid body transform; ϵ denotes additive noise. The reconstruction process is formulated as finding a solution of the inverse problem of the forward model: $I_D^{'} = f_\theta(D, H, T, I_D)$. The nonlinear function f_θ is accomplished by learning a deep neural network.

2.1 Front-End Image Classification

We pre-train a degradation classifier with low-quality images with various simulated degradations, and image-level labels of degradation sources are used for supervised training of a multi-label multi-classification task. High-quality clean images are also involved in pre-training as a separate category (The pre-training stage is not included in Fig .1).

In the second training stage, the pre-trained classifier E_C is assembled in cyclic learning (Fig .1 (a)). Given I_D and its corresponding degradation label CLS_{target}, a pixel-level degradation explanation is generated as:

$$E = CAG[E_C(I_D);\ CLS_{target}] \tag{1}$$

where CAG denotes the Cross-Attention-Gradient method and E denotes the generated explanation, which is transferred to reinforce the back-end shortcut features with degradation-aware knowledge. CLS_{target} is unknown during testing. We use a pseudo-label, assuming that all test data have a mixture of motion, low resolution, and noise degradation, to generate corresponding explanations.

In addition, given the high-quality image (denoted by $I_D^{'}$) reconstructed in the back-end task conditioned on the generated explanation, E_C is also employed as a semantic discriminator to align the semantics of the reconstructed image $I_D^{'}$ with the clean image I_C:

$$L_{sen} = BCELoss(E_C(I_D^{'}), CLS_{clean}) \tag{2}$$

where CLS_{clean} denotes the degradation label of clean images and L_{sen} denotes the semantic loss, which constrains the image reconstruction in the back-end.

2.2 Back-End Image Reconstruction

The back-end reconstruction network has an encoder-decoder architecture with a Restormer [14] backbone, and its encoder E_R is initialized by the weights from the isomorphic front-end classifier. The reconstructed image $I_D^{'}$ is conducted by:

$$I_D^{'} = I_D + D_R[E_R(I_D);\ E] \tag{3}$$

where E_R and D_R denote the reconstruction encoder and decoder. E denotes the degradation explanation of the input low-quality image $I_D^{'}$ and is downsampled by bilinear interpolation for different skip-connection scales. The back-end task predicts the residual between the clean image and the degraded image instead of predicting the whole degraded image from scratch, making it easier to optimize.

The loss function of our back-end task is formulated as:

$$L_{total} = L_{pixel} + \lambda_{sen} L_{sen} \tag{4}$$

where L_{pixel} denotes the mean absolute error between $I_D^{'}$ and I_D, and λ_{sen} is set to 0.01. L_{total} is used to optimize the reconstruction network at both the pixel and semantic levels to eliminate artefacts and noise, while preserving semantic and content consistency with the clean image.

At the end of each iteration, the front-end classifier will be updated by:

$$\theta_{E_C} = m\theta_{E_C} + (1-m)\theta_{E_R} \tag{5}$$

where θ_{E_C} and θ_{E_R} denote the parameters in E_C and E_R, and m is set to 0.99. The cyclic flow of the explanations, losses, and weights facilitates knowledge sharing between our front-end and back-end tasks, which allows the whole system to converge toward a better optimum.

2.3 Cross-Attention-Gradient Interpreting

The attention maps in Multi-head Self-Attentions (MSAs) [15] can describe the relevance and long-range dependencies of tokens. Previous studies typically use attention maps to visualize and comprehend transformers [16, 17]. Zamir et al. [14] provides pixel-level self-attention, making it possible to directly use attention maps to interpret the network. As illustrated in Fig .1 (b), a low-quality image I_D is fed in the pre-trained classifier E_C, and attention maps $[A_1, A_2, ..., A_n]$ of all the n transformer blocks are extracted through forward propagation. Additionally, we compute gradients $[G_1, G_2, ..., G_n]$ with respect to the input of each layer through a backward propagation, aiming for a more comprehensive explanation of the image degradation pattern. Finally, the cross-layer and cross-head attention-gradient explanation can be computed as:

$$E = \sum_{k=1}^{n} drop_{min} drop_{neg}(AVG_{heads}(A_k G_k))) \tag{6}$$

where $drop_{neg}$ denotes replacing all negative contributions with 0, and $drop_{min}$ denotes replacing the smaller contributions with 0 by a fixed percentage (90% in this paper). The dual-drop strategy eliminates confusion from other pixels and highlights pixels that better describe image degradation.

3 Experiments and Results

3.1 Dataset Description and Data Partition

We employed three public brain MRI datasets to assess our methods, including the Cam-bridge Centre for Ageing and Neuroscience (Cam-Can) [18], UCLA [19], and the Autism Brain Imaging Data Exchange I (ABIDE I) [20]. On the Cam-Can and UCLA datasets, we used 70 T1w scans for in silico experiments with simulated degradation (40 volumes for training, 15 volumes for validation, and 15 volumes for testing). On the ABIDE I dataset, we leveraged 10 T1w scans corrupted by real in-scanner head motion for in vivo experiment. We extracted 90 axial slices from each volume, and normalized each slice to the size of 192 $\times$ 224 pixels (Cam-Can), 176 $\times$ 224 pixels (UCLA and ABIDE I). The pixel intensities were rescaled to [-1, 1]. Every three adjacent slices were combined into a three-channel image as the network input.

3.2 Experimental Settings

Degradation simulation. We simulated different sources and levels of degradation for the T1w scans: single degradation (motion, low-resolution, noise), and mixed degradation (motion and low resolution, motion and noise). A single degradation had different levels: mild, moderate/normal, and severe. We added head motion simulated from the translations and rotations of a random sampling of phase-encoding lines in the frequency domain [21]. We cropped out the low-frequency data in the center of k-space, and zeroed out the peripheral data, to generate low-resolution scans [22]. We added white Gaussian noise to simulate noisy acquisitions. Simulations details are shown in the supplementary material.

Implementation details. The front-end classifier was optimized by Adam, with a batch size of 24 and a learning rate of 1e-5 in the pre-training stage. The whole framework was then optimized by Adam, with a batch size of 16 and a learning rate of 2e-4. We evaluated our method by comparing it to its four competing counterparts, including two general baselines (i.e., SRGAN [23], pix2pix [24]), a strong baseline for image restoration (HINet [25]), and a state-of-the-art method for medical image synthetics (Resvit [26]).

3.3 Experimental Results on Simulated Data

Table 1 shows the quantitative results on the simulated datasets. The results show that our method offered superior performance, as compared to its four competing

peers, in terms of PSNR, SSIM, and RMSE. Fig .2 shows the reconstructed images for representative scans from the test datasets. These qualitative results show that SRGAN and pix2pix were unable to completely remove the motion artifacts or restore the images from the blurry and/or noisy acquisitions. HINet and Resvit achieved better reconstruction results but yielded distortions caused by excessive smoothing. In contrast, our method successfully eliminated the motion artifacts, while in parallel, generated images with noise substantially removed and edges considerably sharpened.

Table 1. Experimental results on simulated data.

Dataset	Cam-Can			UCLA		
	PSNR(dB)	SSIM	RMSE	PSNR(dB)	SSIM	RMSE
SRGAN	26.01±1.40	0.823±0.051	0.101±0.017	25.94±1.42	0.821±0.048	0.102±0.015
pix2pix	29.28±2.13	0.889±0.048	0.071±0.017	29.02±2.35	0.859±0.051	0.073±0.019
HINet	30.26±2.63	0.902±0.049	0.064±0.019	29.75±2.84	0.880±0.050	0.069±0.021
Resvit	30.65±2.74	0.906±0.069	0.062±0.023	29.76±2.38	0.879±0.047	0.067±0.018
Ours	**31.34±2.64**	**0.918±0.041**	**0.057±0.017**	**30.75±3.17**	**0.895±0.047**	**0.061±0.019**

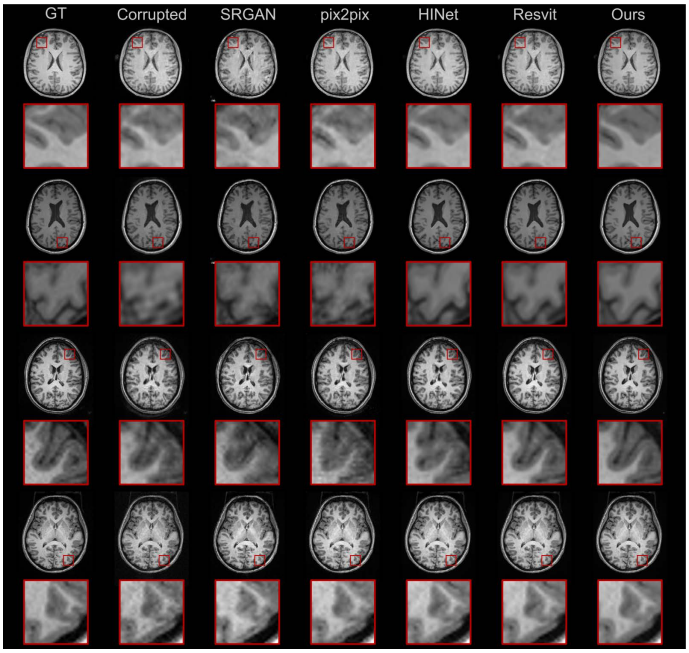

Fig. 2. Qualitative assessments for representative scans from the test datasets. The corrupted images were taken from test sets of Cam-Can and UCLA. The first and third rows show degradation sources: motion and low resolution. The second and fourth rows show degradation sources: motion and noise.

3.4 Experimental Results on Real Motion Data

Table 2 shows the reconstruction results for real acquisitions that are labeled mild/moderate motion [21] from ABIDE I, where no ground-truth was available. The reconstruction model was trained by simulated data from Cam-Can. Three reference-free evaluation metrics (Tenengrad, Entropy, and Grey-level Range) are used to measure the reconstruction quality. Qualitative assessments are shown in the supplementary material due to space limit. The results show that our method achieved better performance in removing motion artefacts, improved image sharpness, and provided superior reconstructions as compared to the four competing methods.

Table 2. Experimental results on real motion data. A smaller Tenengrad value indicates fewer pseudo edges caused by motion artefacts.

Method	Tenengrad/10^2(↓)	Entropy(↑)	Range/10^2(↑)
SRGAN	1.633±2.486	5.418±0.706	1.159±0.974
pix2pix	1.536±2.899	5.582±0.621	1.241±1.044
HINet	1.540±3.099	5.440±0.654	1.305±1.199
Resvit	1.432±2.756	6.019±0.658	1.576±1.215
Ours	**1.272±2.845**	**6.078±0.750**	**1.599±1.399**

3.5 Ablation Study

We performed a thorough ablation study to analyse the contributions of our proposed CAG and cyclic learning to the reconstruction. We started with the vanilla Restormer backbone, and incrementally added cyclic learning (Steps 2 and 4 in Fig .1 have been removed. The pre-trained classifier acted as a semantic-level discriminator but was not utilized for explanation.) and CAG on the framework. Table 3 shows the effectiveness of our proposed CAG and cyclic learning. Visualization of some representative explanations generated by CAG are shown in the supplementary material to illustrate which pixels characterize the degradation pattern and are given more attention during reconstruction.

4 Conclusion

We have developed a new methodology that enables high-quality brain MRI reconstruction from scans corrupted by a mixture of multiple unknown sources of degradation. We have demonstrated the efficacy of our method on both in vivo and in silico data. Experiments have shown that our approach allowed for reconstructing high-quality MRI scans from a variety of unknown sources of degradation that commonly happen in clinical and research MRI studies.

Table 3. Ablation study results on different components of the proposed method.

Dataset	Cam-Can			UCLA		
	PSNR(dB)	SSIM	RMSE	PSNR(dB)	SSIM	RMSE
Restormer	30.40±2.64	0.903±0.046	0.063±0.018	29.81±2.71	0.880±0.048	0.068±0.020
+Cyclic	30.98±2.65	0.914±0.043	0.059±0.017	30.37±2.92	0.888±0.048	0.064±0.019
+Cyclic +CAG(ours)	**31.34±2.64**	**0.918±0.041**	**0.057±0.017**	**30.75±3.17**	**0.895±0.047**	**0.061±0.019**

Acknowledgement. This work was supported by the Faculty Development Award from Peking University under Award No. 71013Y2268 and 73201Y1278.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Afacan, O., Erem, B., Roby, D.P., Roth, N., Roth, A., Prabhu, S.P., Warfield, S.K.: Evaluation of motion and its effect on brain magnetic resonance image quality in children. Pediatric Radiology **46**, 1728–1735 (2016)
2. Plenge, E., Poot, D.H., Bernsen, M., Kotek, G., Houston, G., Wielopolski, P., van der Weerd, L., Niessen, W.J., Meijering, E.: Super-resolution methods in mri: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? Magnetic Resonance in Medicine **68**(6), 1983–1993 (2012)
3. Pruessmann, K.P., Weiger, M., Scheidegger, M.B., Boesiger, P.: SENSE: Sensitivity Encoding for Fast MRI. Magnetic Resonance in Medicine **42**, 952–962 (1999)
4. Pipe, J.: Motion correction with PROPELLER MRI: application to head motion and free-breathing cardiac imaging. Magnetic Resonance in Medicine **42**, 963–969 (1999)
5. Sui, Y., Afacan, O., Jaimes, C., Gholipour, A., Warfield, S.: Gradient-guided isotropic mri reconstruction from anisotropic acquisitions. IEEE Transactions on Computational Imaging **7**, 1240–1253 (2021)
6. Sui, Y., Afacan, O., Jaimes, C., Gholipour, A., Warfield, S.K.: Scan-Specific Generative Neural Network for MRI Super-Resolution Reconstruction. IEEE Transactions on Medical Imaging **41**(6), 1383–1399 (2022)
7. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17452–17462 (2022)
8. Li, D., Zhang, Y., Cheung, K.C., Wang, X., Qin, H., Li, H.: Learning degradation representations for image deblurring. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
9. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.S.: Promptir: Prompting for all-in-one blind image restoration. arXiv preprint arXiv:2306.13090 (2023)
10. Oh, K., Yoon, J.S., Suk, H.I.: Learn-explain-reinforce: counterfactual reasoning and its guidance to reinforce an alzheimer's disease diagnosis model. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4843–4857 (2022)
11. Zhou, Y., Wu, Y., Wang, Z., Wei, B., Lai, M., Shou, J., Fan, Y., Xu, Y.: Cyclic learning: Bridging image-level labels and nuclei instance segmentation. IEEE Transactions on Medical Imaging **42**(10), 3104–3116 (2023)

12. Haque, A., Wang, A., Terzopoulos, D., et al.: Multimix: sparingly-supervised, extreme multitask learning from medical images. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 693–696. IEEE (2021)
13. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1844 (2021)
14. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.H.: Intriguing properties of vision transformers. Advances in Neural Information Processing Systems **34**, 23296–23308 (2021)
17. Park, N., Kim, W., Heo, B., Kim, T., Yun, S.: What do self-supervised vision transformers learn? arXiv preprint arXiv:2305.00729 (2023)
18. Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Henson, R.N., et al.: The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage **144**, 262–269 (2017)
19. Poldrack, R.A., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T., et al.: A phenome-wide examination of neural and cognitive function. Scientific Data **3**(1), 1–12 (2016)
20. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry **19**(6), 659–667 (2014)
21. Duffy, B.A., Zhao, L., Sepehrband, F., Min, J., Wang, D.J., Shi, Y., Toga, A.W., Kim, H., Initiative, A.D.N., et al.: Retrospective motion artifact correction of structural mri images using deep learning improves the quality of cortical surface reconstructions. NeuroImage **230**, 117756 (2021)
22. Lyu, Q., Shan, H., Steber, C., Helis, C., Whitlow, C., Chan, M., Wang, G.: Multi-contrast super-resolution mri through a progressive network. IEEE Transactions on Medical Imaging **39**(9), 2738–2749 (2020)
23. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4681–4690 (2017)
24. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)
25. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 182–192 (2021)
26. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)