A 3D Conditional Diffusion Model for Image Quality Transfer - An Application to Low-Field MRI

Seunghoi Kim[1,2*] **Henry F. J. Tregidgo**[1,2] **Ahmed K. Eldaly**[1,3] **Matteo Figini**[1,3]

Daniel C. Alexander[1,3]

[1]Centre for Medical Image Computing, University College London, London, UK
[2]Department of Medical Physics and Biomedical Engineering, University College London, London, UK
[3]Department of Computer Science, University College London, London, UK

$\{seunghoi.kim.17, h.tregidgo, a.karam, m.figini, d.alexander\}$ @$ucl.ac.uk$

Abstract

Low-field (LF) MRI scanners (<1T) are still prevalent in settings with limited resources or unreliable power supply. However, they often yield images with lower spatial resolution and contrast than high-field (HF) scanners. This quality disparity can result in inaccurate clinician interpretations. Image Quality Transfer (IQT) has been developed to enhance the quality of images by learning a mapping function between low and high-quality images. Existing IQT models often fail to restore high-frequency features, leading to blurry output. In this paper, we propose a 3D conditional diffusion model to improve 3D volumetric data, specifically LF MR images. Additionally, we incorporate a cross-batch mechanism into the self-attention and padding of our network, ensuring broader contextual awareness even under small 3D patches. Experiments on the publicly available Human Connectome Project (HCP) dataset for IQT and brain parcellation demonstrate that our model outperforms existing methods both quantitatively and qualitatively. The code is publicly available at `https://github.com/edshkim98/DiffusionIQT`.

1 Introduction

Magnetic Resonance Imaging (MRI) offers detailed, non-invasive medical diagnostics of human anatomy. However, the quality can vary across different scanners, mainly due to the strength of the magnetic field. In resource-limited settings, MRI with lower-strength fields is still common, leading to reduced tissue contrast and signal-to-noise ratio (SNR). To mitigate these challenges, practitioners often acquire images with thick slices, resulting in a reduction in spatial resolution.

Alexander et al. (2017) have introduced a machine learning technique, called image quality transfer (IQT), which aims to enhance the resolution and contrast of low-quality clinical data using rich information in high-quality images. IQT learns a mapping between low and high-quality paired data. Early frameworks (Alexander et al., 2017; Blumberg et al., 2019; Tanno et al., 2020) demonstrated the effectiveness in enhancing the spatial resolution and information content of diffusion MRI. Subsequent studies (Lin et al., 2023; Ronan and Alexander, 2022) proposed decimation simulators to generate low-field MR images to train deep neural networks, aiming to improve contrast and spatial resolution. More recently, to tackle a domain-shift problem in IQT, Iglesias et al. (2023) developed a

*Corresponding author

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

domain and resolution agnostic model by mapping any input to high-resolution T1-weighted MR images. Despite the progress in IQT, significant limitations persist. Current models, especially with large downsampling factors (e.g. x8), tend to produce blurred outputs. Moreover, while patch-based methods have shown satisfactory performance in IQT, as shown by Huang et al. (2018), they tend to create artifacts stemming from limited contextual information and error propagation from the boundaries of patches.

Recently, Generative Adversarial Networks (GANs) have gained popularity in image synthesis tasks, including super-resolution (Ledig et al., 2017; Wang et al., 2018; Zhang et al., 2021), and segmentation (Kim and Alexander, 2021). Despite their ability to generate realistic images, GANs are difficult to train due to problems, such as mode collapse and training instability. As an alternative to GANs, diffusion models have recently been developed. They iteratively denoise an image, enabling them to generate high-quality images that surpass GANs in Fréchet inception distance (FID) for image synthesis, as shown in Dhariwal and Nichol (2021). However, most studies have focused on unconditional settings (Ho et al., 2020; Song et al., 2021), and natural/medical 2D images (Li et al., 2022; Wu et al., 2022). There have been few studies in 3D domains, especially for volumetric medical data (Chung et al., 2023; Bieder et al., 2023). Chung et al. (2023) adopted diffusion models to solve inverse problems in 3D medical images, but their work was limited to constructing 3D volumes from 2D slices. Bieder et al. (2023) proposed PatchDDPM for 3D medical image segmentation, utilizing patches during training to reduce memory requirements. However, their approach is constrained by the need for very large patches (e.g., 128^3), which entail significant computational costs.

In this work, we extend the 2D unconditional diffusion model by Kingma et al. (2021) to a 3D conditional model, abbreviated as DiffusionIQT, to enhance 3D volumetric data, such as low-field MR images. To the best of our knowledge, this is the first work to apply a diffusion model for 3D medical image enhancement. Our proposed network is a 3D neural network, featuring an encoder equipped with transformer and convolution blocks, to capture local and global information. The decoder uses channel-shuffle and convolution blocks to restore fine-detail textures through efficient upsampling. Additionally, we introduce a novel cross-batch mechanism, which aims to share information across patches in a mini-batch through self-attention and padding. This mechanism provides non-local information, leading to improved global consistency and less artifact. Evaluation on the HCP dataset (Sotiropoulos et al., 2013) for IQT and brain parcellation demonstrates superior performance compared to other existing methods. This highlights the significance of the proposed approach for medical image enhancement.

2 Methods

2.1 Diffusion Process

Our diffusion process is inspired by Kingma et al. (2021), and is composed of a forward process and a reverse process. Assume that we have a clean image x, whose noisy version at an arbitrary time step t is x_t, with $0 \leq t \leq 1$. In the forward process, Gaussian noise is gradually added to a high-quality image until it becomes an isotropic Gaussian noise at $t = 1$, where the noise scale is modulated with the cosine scheduler by Nichol and Dhariwal (2021).

In the reverse process, the model gradually denoises a high-quality image starting from an isotropic Gaussian noise ($t = 1$) until we get a clean high-quality image x. However, without an additional prior, it is difficult to control the generation process and predict a target image faithfully. To mitigate this, we add the corresponding low-quality image at each time step as a condition to our network as an input during the reverse process. The variational lower bound (vlb) derivation (Ho et al., 2020) identifies multiple parameterizations of diffusion models, including ε (noise), x-parametrization, and ν (interpolation between ε and x) by Salimans and Ho (2022). Empirically, we found that x-parameterization shows superior performance over the others. It comes from the easiness of predicting a deterministic parameter x (target image) compared to the other parametrizations that are non-deterministic. Furthermore, this increases the capacity to model complex transitions, requiring fewer time steps to sample high-quality images. If we denote $\hat{x}_\theta$ as a neural network, it takes a noisy image x_t with a condition, low-quality image x_c, as inputs to predict a target image x. Hence, the loss function L can then be simply constructed as a difference between x and the output of $\hat{x}_\theta$:

$$L = \mathbb{E}_{x,t,\varepsilon}[||x - \hat{x}_\theta(\alpha_t x + \sigma_t \varepsilon, x_c, t)||]. \tag{1}$$

where α_t and σ_t are strictly positive scalar-valued functions of t, determined by a pre-defined noise scheduler, and ε is a Gaussian noise.

2.2 Network Architecture

As shown in Figure 1, our novel 3D network, DiffusionIQT, is composed of an encoder and a decoder. It takes concatenated low-field and noisy images as input to predict a target image at each time step. The time embedding is also conditioned on each residual block, in the same manner as Dhariwal and Nichol (2022). The encoder consists of transformers, a 3D extension of Shen et al. (2021), and convolution blocks to introduce long-range dependencies and extract local features. A skip connection is added after the transformer block for an effective fusion of local and global features, and enables vital local spatial information to be preserved, especially crucial in IQT.

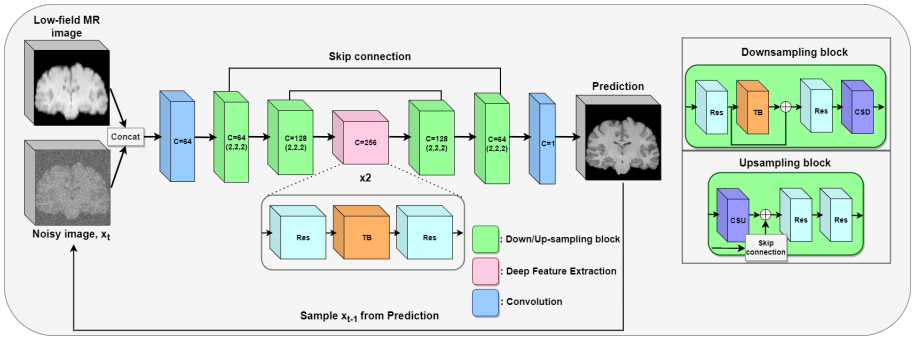

Figure 1: Proposed network architecture (C: number of filters, TB: Transformers block, Res: ResNet block, CSD: Channel-shuffle for downsampling, CSU: Channel-shuffle for upsampling). The numbers in parentheses denote downsampling factor in each dimension.

With a small input patch size, however, this tends to create boundary artifacts and limits the effectiveness of self-attention. To address these limitations, we introduce a cross-batch mechanism, designed to share information between patches within a mini-batch for self-attention and padding. The mini-batch of 3D patches is constructed from patches that are adjacent and non-overlapping to each other. The key and value in self-attention are computed including the adjacent patches rather than a single patch, mimicking cross-attention. This allows computing attention beyond each patch, enabling features to fuse across the patches. Additionally, we pad patch boundaries with actual values from the adjacent patches before each convolution. These approaches enable each small patch to access broader contextual information, which aids in preserving structural consistency and preventing boundary artifact. Please refer to Appendix A.1 for more details. A deep feature extraction module is also added at the bottleneck. This extracts more abstract and high-level features from the input, and it serves to extract a richer representation that compensates for any loss of spatial information.

The decoder consists of an upsampling operation followed by residual blocks. For both downsampling and upsampling feature maps within the network, we employ a 3D channel-shuffle operation, an extension of Shi et al. (2016). This method is parameter-free and makes spatial information to be preserved more faithfully than conventional pooling operations. In contrast to the encoder, the decoder does not incorporate transformers. Empirically, we found that transformers can actually deteriorate the model's performance. This may be due to the fact that the long-range dependencies are not pivotal in the decoder, when it comes to reconstructing local fine details and generating high-quality images.

3 Experiments

3.1 Data

An evaluation was done using the HCP dataset (Sotiropoulos et al., 2013). Synthetic low-field MR volumes akin to those obtained using a real scanner were generated using the decimation simulator by Lin et al. (2023), with a downsampling factor of 8 in the slice direction, mimicking a 4.2mm/1.4mm

Table 1: Quantitative comparison results for IQT and brain parcellation. An upward arrow indicates that a higher value is better, and vice versa.

-	Interpolation	3D U-Net	3D ResU-Net	3D ESRGAN	DiffusionIQT
PSNR (↑)	22.4±0.32	26.2±1.21	26.1±1.45	29.7±1.23	**33.7**±0.80
MSSIM (↑)	0.805±0.06	0.907±0.03	0.871±0.07	0.940±0.02	**0.968**±0.02
LPIPS (↓)	0.293 ±0.018	0.232±0.02	0.181±0.02	0.134±0.01	**0.094**±0.02
Seg mIoU (↑)	0.488±0.05	0.670±0.01	0.751±0.008	0.749±0.01	**0.858**±0.01
Num. Params (↓)	-	48.8M	37.2M	28.7M	**15.2M**

slice thickness/gap. Of 80 subjects curated randomly, 48 were used for training, 12 for validation, and 20 for testing.

3.2 Results

Table 1 presents quantitative comparisons between DiffusionIQT and the four baselines: Interpolation (Alexander et al., 2017), 3D U-Net (Lin et al., 2023), ResU-Net (Lin et al., 2023), and 3D extension of ESRGAN (Wang et al., 2018) in IQT and brain parcellation tasks. Brain parcellation was performed using a pre-trained network from Henschel et al. (2020). The model was trained using MR images that have same resolution as our evaluated data, which predicts 95 different classes and their statistics in each brain volume. To compare the models quantitatively, we used peak signal-to-noise ratio (PSNR), multi-scale structural similarity index measure (MSSIM), and learned perceptual image patch similarity (LPIPS) for IQT, and mean intersection over union (mIoU) for brain parcellation by using the predicted classes for each voxel. The results demonstrate that DiffusionIQT surpasses all other models by large margins in both tasks. In particular, considerable improvements are observed in the PSNR and LPIPS metrics. This suggests that DiffusionIQT predicts overall voxel intensities and restores high-frequency textures more accurately than the other models. In brain parcellation, our approach outperformed other models by more than 0.1 in mIoU. We can also observe that our model surpasses the performance of other models while having the smallest number of parameters — nearly half the size of 3D ESRGAN. This clearly underscores the effectiveness of the diffusion process and our proposed network architecture.

Figure 2 offers qualitative comparisons between the models, illustrating the ability of DiffusionIQT to restore high-frequency textures and accurately delineate tissue structures. As emphasized in the figure, DiffusionIQT more precisely restored the morphology of the sulcus and the lateral ventricle volume compared to other models. While ESRGAN also seemed effective in restoring high-frequency textures, a visible distortion in the size of the restored tissues was observed.

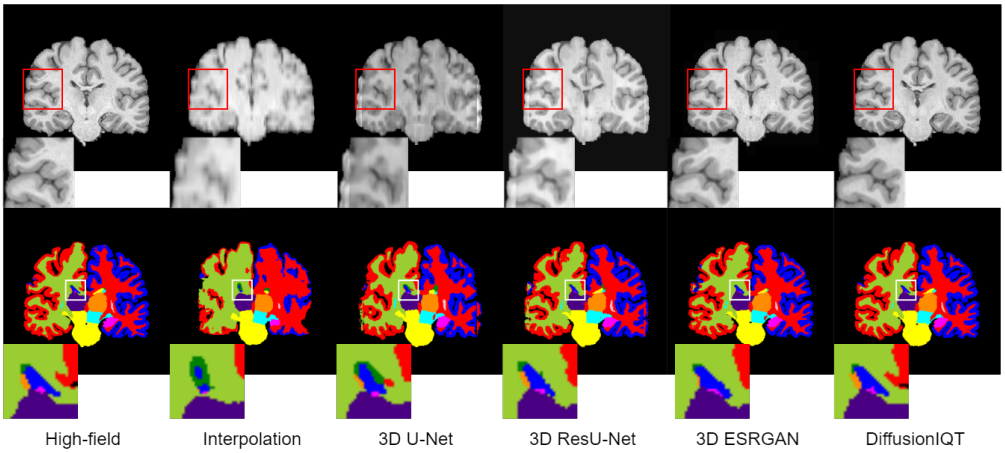

Figure 2: Qualitative comparison on IQT (top) and brain parcellation (bottom)

Table 2: Comparative analysis of the proposed modules (x: non-utilized, o: utilized)

Deep feature	Cross-batch	PSNR	MSSIM	LPIPS	Num. Params
x	x	26.4±1.41	0.848±0.06	0.188±0.01	**9.9M**
o	x	31.5±1.13	0.948±0.04	0.116±0.01	13.6M
o	o	**33.7**±0.80	**0.968**±0.02	**0.094**±0.02	15.2M

3.3 Module Component Analysis

In Table 2, we present ablation studies to demonstrate the effectiveness of our proposed modules. The results indicate that the addition of each module significantly enhances overall performance across various image quality metrics. Notably, the deep feature extraction module contributes the most substantial improvement, yielding more than a 10% increase in all three metrics. While the deep feature extraction module shows much larger gains compared to the cross-batch mechanism, the latter is designed to reduce patch artifacts and enhance structural and contrast consistency, which are subtleties not fully captured by current metrics. Moreover, we achieved this improvement with only a marginal increase in the number of parameters relative to the deep feature extraction module.

4 Discussion

Previous work in IQT demonstrated the effective performance with modest downsampling factors. However, these methods often failed to restore high-frequency textures, resulting in blurry predictions. In order to address this limitation, we introduced a 3D conditional diffusion model for IQT to enhance low-field MRI. Our DiffusionIQT is a pioneering work to harness a generative model and 3D vision transformers in tackling IQT challenges. We also proposed cross-batch mechanisms, sidestepping the limitations of learning from small patches. The proposed approach was evaluated using a low-field MRI application in order to recover contrast enhanced images as similar to high field scanners. Comparative experiments using the HCP dataset for IQT and brain parcellation showed that our model consistently outperformed the baseline models, particularly excelling in restoring high-frequency textures and tissue volumes accurately.

Despite the superior performance in IQT, our work is limited due to the use of synthetic MR images and only healthy subjects. This leaves a future exploration of our model's generalizability to out-of-distribution data and real low-field MR images.

Nevertheless, we can conclude that leveraging diffusion models signals immense potential in IQT, possibly revolutionizing the medical imaging domain in various clinical scenarios, where diagnostic precision and enhanced image quality are essential.

5 Acknowledgement

The authors gratefully acknowledge the helpful input and guidance of the NIHR UCLH Biomedical Research Centre. This work is supported by the EPSRC-funded UCL Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) EP/S021930/1.

References

Daniel C. Alexander, Zikic Darko, Aurobrata Ghosh, Ryutaro Tanno, Viktor Wottschel, Jiaying Zhang, Enrico Kaden, Tim B. Dyrby, Stamatios N. Sotiropoulos, Hui Zhang, and Criminisi Antonio. Image quality transfer and applications in diffusion mri. *NeuroImage*, 152:283–298, 2017.

Florentin Bieder, Julia Wolleb, Alicia Durrer, Robin Sandkuehler, and Philippe Cattin. Diffusion models for memory-efficient processing of 3d medical images. In *Medical Imaging with Deep Learning (MIDL)*, 2023.

Stefano B. Blumberg, Marco Palombo, Can Son Khoo, Chantal M. W. Tax, Ryutaro Tanno, and Daniel C. Alexander. Multi-stage prediction networks for data harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

Hyungjin Chung, Dohoon Ryu, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Solving 3d inverse problems using pre-trained 2d diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021.

Prafulla Dhariwal and Alex Nichol. Photorealistic text-to-image diffusion models with deep language understanding. In *36th Conference on Neural Information Processing Systems*. Publisher, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219: 117012, 2020. ISSN 1053-8119.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Bohao Huang, Daniel Reichman, Leslie M. Collins, Kyle Bradbury, and Jordan M. Malof. Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations. *Arxiv*, 2018.

Juan E. Iglesias, Benjamin Billot, Yaël Balbastre, Colin Magdamo, Steven E. Arnold, Sudeshna Das, Brian L. Edlow, Daniel C. Alexander, Polina Golland, and Bruce Fischl. Synthsr: A public ai tool to turn heterogeneous clinical brain scans into high-resolution t1-weighted images for 3d morphometry. *Science Advances*, 9(5):eadd3607, 2023. doi: 10.1126/sciadv.add3607.

Seunghoi Kim and Daniel C. Alexander. Agcn: Adversarial graph convolutional network for 3d point cloud segmentation. In *British Machine Vision Conference*, 2021.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, 2021.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–114, 07 2017.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.01.029.

Hongxiang Lin, Matteo Figini, Felice D'Arco, Godwin Ogbole, Ryutaro Tanno, Stefano B. Blumberg, Lisa Ronan, Biobele J. Brown, David W. Carmichael, Ikeoluwa Lagunju, Judith Helen Cross, Delmiro Fernandez-Reyes, and Daniel C. Alexander. Low-field magnetic resonance image enhancement via stochastic image quality transfer. *Medical Image Analysis*, 87:102807, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102807.

Diganta Misra. Mish: A self regularized non-monotonic activation function. In *British Machine Vision Conference*, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171, 18–24 Jul 2021.

Lisa Ronan and Daniel C. Alexander. Refining synthetic training data improves image quality transfer for ultra-low-field structural brain mri. In *International Society for Magnetic Resonance in Medicine (ISMRM)*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *IEEE/CVF Winter Conference (WACV)*, 2021.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 06 2016. doi: 10.1109/CVPR.2016.207.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Stamatios N. Sotiropoulos, Saâd Jbabdi, Junqian Xu, Jesper L. R. Andersson, Steen Moeller, Edward J. Auerbach, Matthew F. Glasser, Moisés Hernández, Guillermo Sapiro, Mark Jenkinson, David A. Feinberg, Essa Yacoub, Christophe Lenglet, David C. Van Essen, Kâmil Uğurbil, and Timothy Edward John Behrens. Advances in diffusion mri acquisition and processing in the human connectome project. *NeuroImage*, 80:125–143, 2013.

Ryutaro Tanno, Daniel E. Worrall, Enrico Kaden, and Daniel C. Alexander. Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion mri. *NeuroImage*, 225, 2020.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV WorkShop*, page 63–79, 2018. ISBN 978-3-030-11020-8. doi: 10.1007/978-3-030-11021-5_5.

Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022.

Kuan Zhang, Haoji Hu, Kenneth A. Philbrick, Gian Marco Conte, Joseph D. Sobek, Pouria Rouzrokh, and Bradley James Erickson. Soup-gan: Super-resolution mri using generative adversarial networks. *Tomography*, 8:905 – 919, 2021.

A Appendix

A.1 Network Architecture Details

In this section, we detail our proposed network architecture, especially the transformers block.

Our transformers are inspired by Shen et al. (2021). We extend their work to process 3D volumetric data and apply several modifications to effectively capture spatial features. In particular, instead of the traditional method of tokenizing input into non-overlapping patches and then flattening them into 1D, we maintain the 3D shape by tokenizing patches using a 3D depth-wise convolution with both a kernel size and a stride equivalent to the transformer's patch size. Furthermore, we utilize additional depth-wise convolutions for the computation of the query, key, and value.

As shown in Figure 3, the self-attention relies on our novel cross-batch mechanism. Unlike conventional vision transformers (Dosovitskiy et al., 2021), we first perform a dot-product between the key and value. As proposed by Shen et al. (2021), this approach allows for a more efficient operation, as the complexity does not increase with the number of patches. Furthermore, we derive the key and value from adjacent patches, which offers broader contextual information and maximizes the efficacy of self-attention.

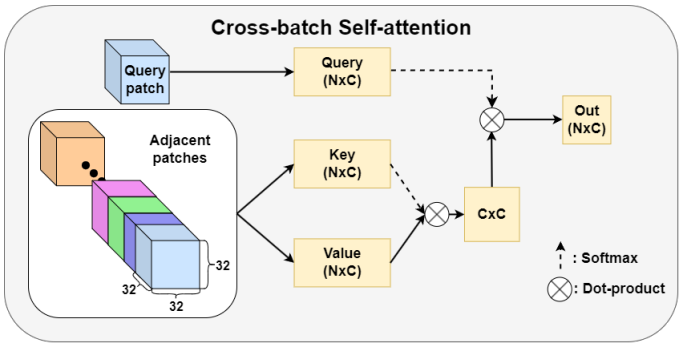

Figure 3: Cross-batch mechanism for self-attention. N and C denote the number of patches for self-attention and the number of channels, respectively.

Network Architecture

- Input size: $32 \times 32 \times 32$
- Learning rate: $1e^{-4}$
- Number of filters: [64,128,256]
- Loss function: Mean Squared Error (MSE)
- Activation function: Mish (Misra, 2020)
- Skip-connection scale: $\frac{1}{\sqrt{2}}$
- Number of multi-head self-attention (MHSA): 8
- Embedding size (C): 512
- Number of patches (N): 216

Diffusion Process

- Number of time steps: 20
- Parametrization method: x (target image)
- Noise scheduler: Cosine (Nichol and Dhariwal, 2021)

A.2 Diffusion Process in DiffusionIQT

In this section, we provide mathematical detail of our diffusion process inspired from Kingma et al. (2021).

A.2.1 Forward Process

As illustrated in Section 2.1, our diffusion process is mainly composed of a forward process and a reverse process. In the forward process, it involves gradually adding noise to a clean image x to transform it into a noisy image x_t, where t is an arbitrary time step value $0 \leq t \leq 1$. Formally, the probability distribution of x_t given any arbitrary previous time step x_s is defined as:

$$\begin{aligned} q(x_t|x_s) &= \mathcal{N}(\alpha_{t|s}x_s, \sigma^2_{t|s}I) \\ where\ \alpha_{t|s} &= \alpha_t/\alpha_s, and\ \sigma^2_{t|s} = \sigma^2_t - \alpha^2_{t|s}\sigma^2_s, \end{aligned} \tag{2}$$

where α_t and σ^2_t denote the diffusion coefficient and noise level at time step t, respectively. These parameters are strictly positives and determined by cosine scheduler (Nichol and Dhariwal, 2021). Kingma et al. (2021) chose variance preserving diffusion process, by setting $\alpha_t = \sqrt{1-\sigma^2_t}$. To sample an arbitrary noisy image x_t conditioned on x, we sample using $x_t = \alpha_t x + \sigma_t \varepsilon$ where ε is a Gaussian noise.

A.2.2 Reverse Process

In the reverse process, our goal is to iteratively remove the noise from the isotropic Gaussian noise we sampled at time $t = 1$, until we obtain a clean image, which is achieved by maximizing the likelihood of $p(x)$. For a finite number of T time steps, by defining $s(i) = (i-1)/T$ and $t(i) = i/T$, the data x can be represented through our generative model as:

$$p(x) = \int p(x_1)p(x|x_0, x_c) \prod_{i=1}^{T} p(x_{s(i)}|x_{t(i)}, x_c) dx_{0:1}, \tag{3}$$

where $p(x_1) = \mathcal{N}(x_1; 0, I)$ and x_c is a conditioned low-quality MR image. To maximize the likelihood, the neural network is trained to approximate $q(x_s|x_t, x)$ with $p_\theta(x_s|x_t, x_c)$, where θ is learnable parameters. Given that both $q(x_t|x)$ and $q(x_s|x)$ are Gaussian distributions:

$$q(x_s|x_t, x) = \frac{q(x_s, x_t|x)}{q(x_t|x)} = \frac{q(x_s|x)q(x_t|x_s)}{q(x_t|x)}, \tag{4}$$

we leverage the conjugate prior property in Bayes' theorem, which ensures that $q(x_s|x_t, x)$ is a Gaussian. Consequently, we model $p_\theta(x_s|x_t, x_c) \sim \mathcal{N}(x_s; \hat{\mu}_\theta(x_t; t, x_c), \sigma^2_t I)$, where $\hat{\mu}_\theta$ represents predicted μ by a neural network.

Being a Gaussian, this formulation aligns with the derivation of the variational lower bound as defined in (Ho et al., 2020) for the continuous-time diffusion process. This simplifies our learning objective to computing the mean-squared error as follows:

$$\begin{aligned} L_t &= \mathbb{E}_{x,t}[||\mu(x_t, x; t) - \hat{\mu}_\theta(x_t, x_\theta; t, x_c)||^2] \\ &= \mathbb{E}_{x,t}[||x - \hat{x}_\theta||^2] \\ &= \mathbb{E}_{\varepsilon}[||\varepsilon - \hat{\varepsilon}_\theta||^2]. \end{aligned} \tag{5}$$

Following (Kingma et al., 2021), the model is parameterized to predict the noise, building upon the approach in (Ho et al., 2020). However, in our approach, we parameterize the model to predict the target image x at each time step, which has demonstrated faster convergence in our experiments.

During training, we randomly sample a time step t from $0 \leq t < 1$, to sample a noisy image x_t. Then, we predict a target image using our neural network $\hat{x}_\theta$ given x_t, x_c and t. This prediction is then compared with the target image using the reconstruction loss (e.g. L1 or L2) as shown in Equation 5. Once the training is finished, during sampling, given a finite T, we discretize time uniformly into T timesteps (segments) and scale them into $[0, 1]$. We start from $t = 1$ by sampling isotropic Gaussian noise. After predicting the target image x at time step t given the LF image, we then sample our prediction $\hat{x}$ to obtain the noisy image x_{t-1}. This process is iterated until $t = 0$. Algorithm 1 and

2 show the complete training procedure with the simplified objective and sampling procedure after training.

Algorithm 1 Training

repeat
Sample $t \sim \text{Uniform}(0, 1)$
$x_t = \alpha_t \cdot x + \sigma_t \cdot \varepsilon$
Take gradient descent on $\nabla_\theta ||x - \hat{x}_\theta(x_t, x_c, t)||^2$
until converged

Algorithm 2 Inference (Sampling)

Scale the timesteps from the range $[0, T]$ to the normalized range $[0, 1]$
$x_1 \sim \mathbf{N}(0, I)$
for $t = 1, ..., 0$ **do**
    $\varepsilon \sim \mathcal{N}(0, I)$ if $t > 0$, else $\varepsilon = 0$
    $x_{t-1} = \alpha_{t-1} \hat{x}_\theta(x_t, x_c, t) + \sigma_{t-1}\varepsilon$
    $x_t = x_{t-1}$
end for