Generating High-Resolution 3D CT with 12-Bit Depth Using a Diffusion Model with Adjacent Slice and Intensity Calibration Network

Jiheon Jeong[1], Ki Duk Kim[2], Yujin Nam[1], Kyungjin Cho[1], Jiseon Kang[2], Gil-Sun Hong[3], and Namkug Kim[2,3](✉)

[1] Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine, Seoul 05505, South Korea
[2] Department of Convergence Medicine, Asan Medical Center, Seoul 05505, South Korea
namkugkim@gmail.com
[3] Department of Radiology, Asan Medical Center, Seoul 05505, South Korea

Abstract. Since the advent of generative models, deep learning-based methods for generating high-resolution, photorealistic 2D images have made significant successes. However, it is still difficult to create precise 3D image data with 12-bit depth used in clinical settings that capture the anatomy and pathology of CT and MRI scans. Using a score-based diffusion model, we propose a slice-based method that generates 3D images from previous 2D CT slices along the inferior direction. We call this method stochastic differential equations with adjacent slice-based conditional iterative inpainting (ASCII). We also propose an intensity calibration network (IC-Net) that adjusts the among slices intensity mismatch caused by 12-bit depth image generation. As a result, Frechet Inception Distance (FIDs) scores of FID-Ax, FID-Cor and FID-Sag of ASCII(2) with IC-Net were 14.993, 19.188 and 19.698, respectively. Anatomical continuity of the generated 3D image along the inferior direction was evaluated by an expert radiologist with more than 15 years of experience. In the analysis of eight anatomical structures, our method was evaluated to be continuous for seven of the structures.

Keywords: Score-based Diffusion Model · Adjacent Slice-based 3D Generation · Intensity Calibration Network · 12-bit Depth DICOM

1 Introduction

Unlike natural images that are typically processed in 8-bit depth, medical images, including X-ray, CT, and MR images, are processed in 12-bit or 16-bit depth to retain more detailed information. Among medical images, CT images are scaled using a quantitative measurement known as the Hounsfield unit (HU), which ranges from -1024 HU to 3071

J. Jeong and K. D. Kim—Contributed equally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43999-5_35.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 366–375, 2023.
https://doi.org/10.1007/978-3-031-43999-5_35

HU in 12-bit depth. However, in both clinical practice and research, the dynamic range of HU is typically clipped to emphasize the region of interest (ROI). Such clipping of CT images, called windowing, can increase the signal-to-noise ratio (SNR) in the ROI. Therefore, most research on CT images performs windowing as a pre-processing method [1, 2].

Recent advancements in computational resources have enabled the development of 3D deep learning models such as 3D classification and 3D segmentation. 3D models have attracted much attention in the medical domain because they can utilize the 3D integrity of anatomy and pathology. However, the access to 3D medical imaging datasets is severely limited due to the patient privacy. The inaccessibility problem of 3D medical images can be addressed by generating high quality synthetic data. Some researches have shown that data insufficiency or data imbalance can be overcome using a well-trained generative model [3, 4]. However, generating images with intact 3D integrity is very difficult. Moreover, generating high-quality images [5] in the 12-bit depth, which is used in real clinical settings, is even more challenging.

The present study proposes a 2D-based 3D-volume generation method. To preserve the 3D integrity and transfer spatial information across adjacent axial slices, prior slices are utilized to generate each adjacent axial slice. We call this method Adjacent Slice-based Conditional Iterative Inpainting, ASCII. Experiments demonstrated that ASCII could generate 3D volumes with intact 3D integrity. Recently, score-based diffusion models have shown promising results in image generation [6–12], super resolution [13] and other tasks [14–17]. Therefore, ASCII employs a score-based diffusion model to generate images in 12-bit depth. However, since the images were generated in 12-bit depth, errors in the average intensity arose when the images were clipped to the brain parenchymal windowing range. To solve this issue, we propose a trainable intensity-calibration network (IC-Net) that matches the intensity of adjacent slices, which is trained in a self-supervised manner.

2 Related Works

Score-based generative models [7–9] and denoising diffusion probabilistic models (DDPMs) [6, 16] can generate high-fidelity data without an auxiliary network. In contrast, generative adversarial networks (GANs) [18] require a discriminator and variational auto-encoders (VAEs) [19] require a Gaussian encoder. Score-based generative models and diffusion models have two processing steps: a forward process that creates perturbed data with random noise taken from a pre-defined noise distribution in each step, and a backward process that denoises the perturbed data using a score network. The perturbation methods were defined as the stochastic differential equation (SDE) in [8]. A continuous process was defined as $\{x(t)\}_{t=0}^{T}$ with $x(0) \sim p_{data}$ and $x(T) \sim p_T$, where $t \in [0, 1]$ and p_{data}, p_T are the data distribution and prior noise distribution, respectively. The forward process was defined as the following SDE:

$$dx = f(x, t)dt + g(t)dw, \tag{1}$$

where f and g are the coefficients of the drift and diffusion terms in the SDE, respectively, and w induces the Wiener process (i.e., Brownian motion). The backward process was defined as the following reverse-SDE:

$$dx = \left[f(x,t) - g(t)^2 \nabla x log p_t(x)\right]dt + g(t)dw, \tag{2}$$

where w is the backward Wiener process. We define each variance σ_t as a monotonically increasing function. To solve the reverse-SDE given by above equation we train a score network $S_\theta(x,t)$ to estimate the score function of the perturbation kernel $\nabla x log p_t(x_t|x_0)$. Therefore, the objective of the score network is to minimize the following loss function:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \int \lambda(t)\mathbb{E}_{x(0)}\mathbb{E}_{x(t)|x(0)} \| S_\theta(x(t),t) - \nabla x log p_t(x(t)|x(0)) \|_2^2 dt, \tag{3}$$

where $\lambda(t)$ is a coefficient function depending on SDE. When the score network is trained using above equation, we approximate the score network $S_\theta(x,t)$ as $\nabla x log p_t(x_t|x_0)$. The model that generates using the predictor-corrector (PC) sampler, which alternately applies a numerical solver called predictor and Langevin MCMC called corrector, is called a *score-based diffusion model*. We also apply a perturbation kernel $p(x_t|x_0) = N\left(x_t; x_0, \sigma_t^2\right)$ and set the drift and diffusion coefficients of the SDE to $f = 0$ and $g = \sqrt{d\left[\sigma_t^2\right]/dt}$, respectively. We call this variance the exploding SDE (VESDE).

3 Adjacent Slice-Based Conditional Iterative Inpainting, ASCII

3.1 Generating 12-Bit Whole Range CT Image

VESDEs experimented on four noise schedules and two GAN models [20–22] were compared for 12-bit whole range generation. As shown in Fig. 1, setting σ_{min} to 0.01 generated a noisy image (third and fifth column) whereas the generated image was well clarified when σ_{min} was reduced to 0.001 (forth and last column). However, in two GAN models, the important anatomical structures, such as white matter and grey matter, collapsed in the 12-bit generation. The coefficient of variation (CV), which is a measure used to compare variations while eliminating the influence of the mean, measured in the brain parenchyma of the generated images (excluding the bones and air from the windowing) at both noise levels [23]. It is observed that the anatomical structures tended to be more distinguishable when the CV was low. Furthermore, a board-certified radiologist also qualitatively assessed that the images generated with a lower σ_{min} showed a cleaner image. It can be interpreted that reducing σ_{min} lowers CV and therefore improves the image quality.

For quantitative comparison, we randomly generated 1,000 slices by each parameter and measured the CV. As shown Fig. 1, the variance of CV was lowest when σ_{min} and σ_{max} were set to 0.001 and 68, respectively. Also, setting σ_{max} to 1,348 is theoretically plausible according to previous study [9] because we preprocessed CT slices in the range of -1 to 1. Finally, we fixed σ_{min} and σ_{max} to 0.001 and 1,348, respectively in subsequent experiments.

In the case of GAN, as the convolution operator can be described as a high-pass filter [24], GANs trained through the discriminator's gradient are vulnerable to generating

precise details in the low-frequency regions. Therefore, while anatomy such as bone (+1000HU) and air (−500HU) with strong contrast and high SNR are well-generated, anatomy such as parenchyma (20HU–30HU) with low contrast and SNR are difficult to generate accurately. As shown in the first and second column of Fig. 1, we can see that the GAN models seem to generate anatomy with strong contrast and high SNR well but failed to generate the others.

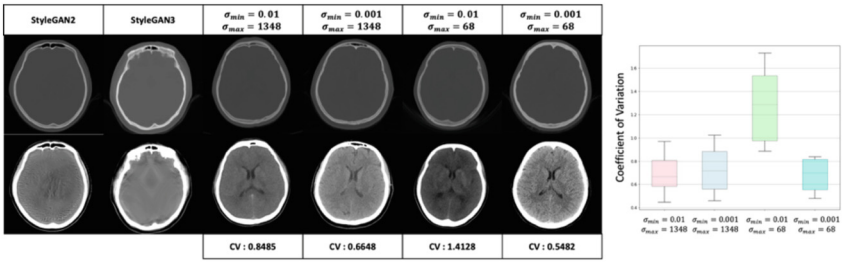

Fig. 1. Qualitative and quantitative generation results according to σ_{min} and σ_{max}

The SDE's scheduling can impact how the score-based diffusion model generates the fine-grained regions of images. To generate high-quality 12-bit images, the diffusion coefficient must be set to distinguish 1HU (0.00049). By setting σ_{max} to 1,348 and σ_{min} to 0.001, the final diffusion coefficient (0.00017) was set to a value lower than 1HU. In other words, setting the diffusion coefficient (0.00155) to a value greater than 1HU can generate noisy images. The detailed description and calculation of diffusion coefficient was in **Supplementary Material.**

3.2 Adjacent Slice-Based Conditional Iterative Inpainting

To generate a 3D volumetric image in a 2D slice-wise manner, a binary mask was used, which was moved along the channel axis. A slice $x^0 = \{-1024HU\}^D$ filled with intensity of air was padded before the first slice x^1 of CT and used as the initial seed. Then, the input of the model was given by $\left[x^t : x^{t+K-1}\right]$, where $t \in [0, N_s - K + 1]$ and N_s and K are the total slice number of CT and the number of contiguous slices, respectively. In addition, we omit augmentation because the model itself might generate augmented images. After training, the first slice was generated through a diffusion process using the initial seed. The generated first slice was then used as a seed to generate the next slice in an autoregressive manner. Subsequently, the next slices were generated through the same process. We call this method *adjacent slice-based conditional iterative inpainting, ASCII.*

Two experiments were conducted. The first experiment preprocessed the CT slices by clipping it with brain windowing $[-10HU, 70HU]$ and normalizing it to $[-1, 1]$ with σ_{min} set to 0.01 and σ_{max} set to 1,348, while the second experiment preprocessed the whole windowing $[-1024HU, 3071HU]$ and normalized it to $[-1, 1]$ with σ_{min} set to 0.001 and σ_{max} set to 1,348.

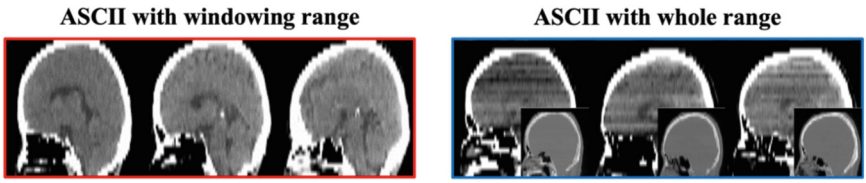

Fig. 2. Results of ASCII with windowing range (red) and 12-bit whole range (blue)

As shown Fig. 2, the white matter and gray matter in the first experiment (brain windowing) could be clearly distinguished with maintained continuity of slices, whereas they were indistinguishable and remained uncalibrated among axial slices in the second experiment (whole range).

3.3 Intensity Calibration Network (IC-Net)

It was noted that the intensity mismatch problem only occurs in whole range generation. To address this issue, we first tried a conventional non-trainable post-processing, such as histogram matching. However, since each slice has different anatomical structure, the histogram of each slice image was fitted to their subtle anatomical variation. Therefore, anatomical regions were collapsed when the intensities of each slice of 3D CT were calibrated using histogram matching. Finally, we propose a solution for this intensity mismatching, a trainable intensity calibration network: IC-Net.

To calibrate the intensity mismatch, we trained the network with a self-supervised manner. First, adjacent two slices from real CT images, x^t, x^{t+1} were clipped using the window of which every brain anatomy HU value can be contained. Second, the intensity of x^{t+1} in ROI is randomly changed and the result is $\hat{x}_c^{t+1} = \left(x^{t+1} - \overline{x^{t+1}}\right) * \mu + \overline{x^{t+1}}$, where $\overline{x^{t+1}}$ and μ are the mean of x^{t+1} and shifting coefficient, respectively. And μ was configured to prevent the collapse of anatomical structures.

Finally, *intensity calibration network,* IC-Net was trained to calibrate the intensity of x^{t+1} to the intensity of x^t. The objective of IC-Net was only to calibrate the intensity of x^t and preserve both the subtle texture and the shape of a generated slice. The IC-Net uses the prior slice to calibrate the intensity of generated slice. The objective function of IC-Net is given by,

$$\mathcal{L}_{IC} = \mathbb{E}_t \mathbb{E}_{\mu \sim \mathcal{U}[-0.7,1.3]}\left[\left| ICNet\left(\hat{x}_\mu^{t+1}, x^t\right) - x^{t+1}\right|\right] \tag{4}$$

As shown in Fig. 3, some important anatomical structures, such as midbrain, pons, medulla oblongata, and cerebellar areas, are blurred and collapsed when histogram matching was used. It can be risky as the outcomes vary depending on the matching seed image. On the other hand, the anatomical structure of the IC-Net matched images did not collapse. Also, the IC-Net does not require to set the matching seed image because it normalizes using the prior adjacent slice.

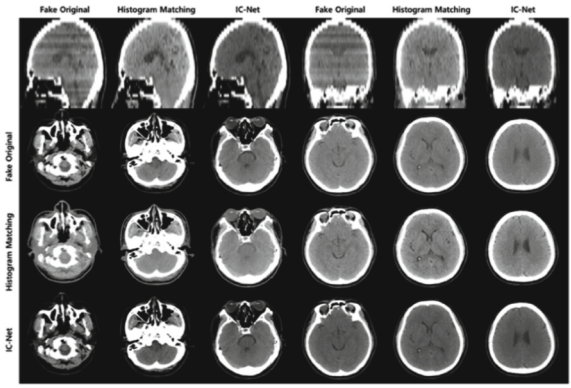

Fig. 3. Post-processing results of ASCII(2) in windowed range of whole range generation.

4 Experiments

4.1 ASCII with IC-Net in 12-Bit Whole Range CT Generation

The description of the dataset and model architecture is available in the **Supplementary Material**. We experimented ASCII on continuous K slices of K = 2 and 3 and called them ASCII(2) and ASCII(3), respectively. We generated a head & neck CT images via ASCII(2) and ASCII(3) with and without IC-Net, and slice-to-3D VAE [25]. Figure 4 demonstrate the example qualitative images. The 3D generated images were shown both in whole range and brain windowing range. The results showed that the both ASCII(2) and ASCII(3) were well calibrated using IC-Net. Also, anatomical continuity and the 3D integrity is preserved while the images were diverse enough. However, there was no significant visual difference between ASCII(2) and ASCII(3). Although the results in whole range appear to be correctly generated all models, the results in brain windowing range showed the differences. The same drawback of convolution operation addressed in the 12-bit generation of GAN based models, which was shown in Fig. 1, was also shown in slice-to-3D VAE.

The quantitative results in whole range are shown in Table 1. The mid-axial slice, mid-sagittal slice, and mid-coronal slice of the generated volumes were used to evaluate the Fréchet Inception Distance (FID) score, which we designated as FID-Ax, FID-Sag, and FID-Cor, respectively. And multi-scales structural similarity index measure (MS-SSIM) and batch-wise squared Maximum Mean Discrepancy (bMMD^2) were also evaluated for quantitative metrics. In general, quantitative results indicate that ASCII(2) performs better than ASCII(3). It's possible that ASCII(3) provides too much information from prior slices, preventing it from generating sufficiently diverse images. Additionally, IC-Net significantly improved generation performance, especially in the windowing range. The FID-Ax of ASCIIs was improved by IC-Net from 15.250 to 14.993 and 18.127 to 16.599 in the whole range, respectively. Also, the performance of FID-Cor and FID-Sag had significantly improved when IC-Net was used. The MS-SSIM showed that ASCIIs can generated it diverse enough.

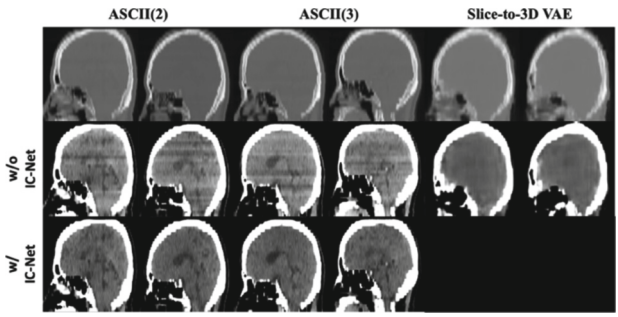

Fig. 4. Qualitative comparison among ASCII(2) and ASCII(3) with/without IC-Net calibration and Slice-to-3D VAE models.

The FID-Ax, FID-Cor, and FID-Sag scores of ASCIIs with IC-Net were improved in windowing range. The FID-Ax of ASCIIs was improved by IC-Net from 15.770 to 14.656 and 20.145 to 15.232 in the windowing range, respectively. On the other hand, ASCIIs without IC-Net had poor performance in the windowing range and this means that even when IC-Net is used, structures do not collapse.

Table 1. Quantitative comparison of ASCII(2) and ASCII(3) with/without IC-Net calibration and Slice-to-3D VAE in whole range and windowing range. Whole range and windowing range are set to [−1024HU, 3071HU] and [−10HU, 70HU], respectively.

	ASCII(2) w/ IC-Net	ASCII(2) w/o IC-Net	ASCII(3) w/ IC-Net	ASCII(3) w/o IC-Net	Slice-to-3D VAE [25]
Whole Range					
FID-Ax	**14.993**	15.250	16.599	18.127	29.137
FID-Cor	19.188	**19.158**	20.930	21.224	28.263
FID-Sag	19.698	**19.631**	21.991	22.311	29.024
MS-SSIM	**0.6271**	0.6275	0.6407	0.6406	0.9058
bMMD2	425704	429120	428045	432665	**311080**
Windowing Range					
FID-Ax	**14.656**	15.770	15.232	20.145	28.682
FID-Cor	**18.920**	19.830	19.996	24.230	28.828
FID-Sag	**18.569**	19.675	19.840	24.511	29.912
MS-SSIM	**0.5287**	0.5384	0.5480	0.5447	0.8609
$bMMD^2$	1975336	**1854921**	2044218	1858850	1894911

4.2 Calibration Robustness of IC-Net on Fixed Value Image Shift

To demonstrate the performance of IC-Net, we conducted experiments with the 7, 14, 21 and 28th slices, which sufficiently contain complex structures to show the calibration performances. The previous slice was used as an input to the IC-Net along with the target slice whose pixel values were to be shifted. And the absolute errors were measured between GT and predicted slice using IC-Net.

As shown in Fig. 5, it worked well for most shifting coefficients. The mean absolute error was measured from 1HU to 2HU when the shifting coefficient was set from 0.7 to 1.1. However, the errors were exploded when shifting coefficient was set to 1.2 or 1.3. It was because the images were collapsed when shifting coefficient increases than 1.2 since the intensity deviates from the ROI range $[-150HU, 150HU]$. Nevertheless, IC-Net can calibrate intensity to some extent even in the collapsed images as shown Fig. 5.

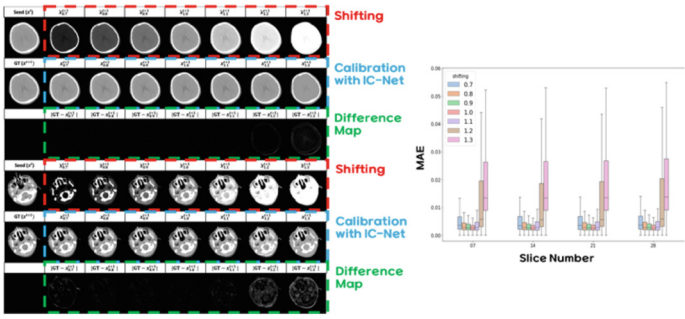

Fig. 5. (Left) IC-Net calibration results for fixed value image shift. Shifted images, IC-Net calibration result, and difference map are presented, respectively. Images were shifted with fixed value from 0.7 to 1.3. (Right) IC-Net calibration results for fixed value image shift. All slices were normalized from $[-150HU, 150HU]$ to $[-1, 1]$.

4.3 Visual Scoring Results of ASCII with IC-Net

Due to the limitations of slice-based methods in maintaining connectivity and 3D integrity, an experienced radiologist with more than 15 years of experience evaluated the images. Seeded with the 13th slices of real CT scans, in which the ventricle appears, ASCII(2) with IC-Net generated a total of 15 slices. Visual scoring shown in Table. 2 on a three-point scale was conducted blindly for 50 real and 50 fake CT scans, focusing on the continuity of eight anatomical structures. Although most of the fake regions were scored similarly to the real ones, the basilar arteries were evaluated with broken continuity. The basilar artery was frequently not generated, because it is a small region. As the model was trained on 5-mm thickness non-contrasted enhanced CT scans, preserving the continuity of the basilar artery is excessively demanding.

Table 2. Visual scoring results of integrity evaluation

Anatomy	Real	Fake
Skull (bone morphology, suture line)	3.00	2.98
Skull base (foramina and fissure)	3.00	2.84
Facial bone	3.00	2.98
Ventricles	3.00	2.92
Brain sulci and fissure	3.00	2.98
Basilar artery	2.92	1.38
Cerebra venous sinus	3.00	3.00
Ascending & descending nerve tract through internal capsule	3.00	3.00

***Note**: Scale 1 – discontinuity, Scale 2 – strained continuity, Scale 3 – well preserved continuity*

5 Conclusion

We proposed a high-performance slice-based 3D generation method (ASCII) and combined it with IC-Net, which is trained in a self-supervised manner without any annotations. In our method, ASCII generates a 3D volume by iterative generation using previous slices and automatically calibrates the intensity mismatch between the previous and next slices using IC-Net. This pipeline is designed to generate high-quality medical image, while preserving 3D integrity and overcoming intensity mismatch caused in 12-bit generation.

ASCII had shown promising results in 12-bit depth whole range and windowing range, which are crucial in medical contexts. The integrity of the generated images was also confirmed in qualitative and quantitative assessment of 3D integrity evaluations by an expert radiologist. Therefore, ASCII can be used in clinical practice, such as anomaly detection in normal images generated from a seed image [26]. In addition, the realistic 3D images generated by ASCII can be used to train deep learning models [3, 4] in medical images, which frequently suffer from data scarcity.

References

1. Gerard, S.E., et al.: CT image segmentation for inflamed and fibrotic lungs using a multi-resolution convolutional neural network. Sci. Rep. **11**(1), 1–12 (2021)
2. Lassau, N., et al.: Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nat. Commun. **12**(1), 1–11 (2021)
3. Frid-Adar, M., et al.: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)
4. Bowles, C., et al.: Gan augmentation: augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)
5. Hong, S., et al.: 3d-stylegan: a style-based generative adversarial network for generative modeling of three-dimensional medical images. In: Deep Generative Models, and Data Augmentation, Labelling, and Imperfections, pp. 24–34. Springer (2021)

6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
7. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Adv. Neural Inform. Process. Syst. **32** (2019)
8. Song, Y., et al.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
9. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Adv. Neural. Inf. Process. Syst. **33**, 12438–12448 (2020)
10. Meng, C., et al.: Sdedit: Image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
11. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. PMLR (2021)
12. Hyvärinen, A., Dayan, P.: Estimation of non-normalized statistical models by score matching. J. Mach. Learn. Res. **6**(4) (2005)
13. Saharia, C., et al.: Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636 (2021)
14. Kong, Z., et al.: Diffwave: a versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
15. Chen, N., et al.: WaveGrad: estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713 (2020)
16. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
17. Mittal, G., et al.: Symbolic music generation with diffusion models. arXiv preprint arXiv:2103.16091 (2021)
18. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inform. Process. Syst. **27** (2014)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Karras, T., et al.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
21. Karras, T., et al.: Training generative adversarial networks with limited data. Adv. Neural. Inf. Process. Syst. **33**, 12104–12114 (2020)
22. Karras, T., et al.: Alias-free generative adversarial networks. Adv. Neural. Inf. Process. Syst. **34**, 852–863 (2021)
23. Brunel, N., Hansel, D.: How noise affects the synchronization properties of recurrent networks of inhibitory neurons. Neural Comput. **18**(5), 1066–1110 (2006)
24. Park, N., Kim, S.: How Do Vision Transformers Work? arXiv preprint arXiv:2202.06709 (2022)
25. Volokitin, A., Erdil, Ertunc, Karani, Neerav, Tezcan, Kerem Can, Chen, Xiaoran, Van Gool, Luc, Konukoglu, Ender: Modelling the distribution of 3D brain MRI using a 2D slice VAE. In: Martel, A.L., Abolmaesumi, Purang, Stoyanov, Danail, Mateus, Diana, Zuluaga, Maria A., Kevin Zhou, S., Racoceanu, Daniel, Joskowicz, Leo (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII, pp. 657–666. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_64
26. Schlegl, T., et al.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019)