A New Chapter for Medical Image Generation: The Stable Diffusion Method

Loc X.Nguyen, Pyae Sone Aung, Huy Q.Le, Seong-Bae Park, Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University, 171-04, Republic of Korea
Email:{xuanloc088, pyaesoneaung, quanghuy69, sbpark71, cshong}@khu.ac.kr

2023 International Conference on Information Networking (ICOIN) | 978-1-6654-6268-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICOIN56518.2023.10049010

***Abstract*—Data collecting and sharing have been widely accepted and adopted to improve the performance of deep learning models in almost every field. Nevertheless, in the medical field, sharing the data of patients can raise several critical issues, such as privacy and security or even legal issues. Synthetic medical images have been proposed to overcome such challenges; these synthetic images are generated by learning the distribution of realistic medical images but completely different from them so that they can be shared and used across different medical institutions. Currently, the diffusion model (DM) has gained lots of attention due to its potential to generate realistic and high-resolution images, particularly outperforming generative adversarial networks (GANs) in many applications. The DM defines state of the art for various computer vision tasks such as image inpainting, class-conditional image synthesis, and others. However, the diffusion model is time and power consumption due to its large size. Therefore, this paper proposes a lightweight DM to synthesize the medical image; we use computer tomography (CT) scans for SARS-CoV-2 (Covid-19) as the training dataset. Then we do extensive simulations to show the performance of the proposed diffusion model in medical image generation, and then we explain the key component of the model.**

***Index Terms*—Medical Image Generation, Diffusion Model, U-Net architecture, CT scan of Covid-19.**

I. Introduction

WITH the rapid development in medical technologies, the radiologist can precisely diagnose the diseases of patients based on their medical imaging such as MRI, CT, PET, and others without kind of images invasive test [1]. However, to achieve such performances, the radiologist may need years of experience for many reasons, one of which is the unavailable access to large amounts of training images for radiologists. Medical image data cannot be shared among medical institutions, they not only contain privacy information but also identifiers such as social security number, age, and occupation [2]. Medical image generation is the way around this issue, in which they use a deep learning model to generate the synthesis images based on the real one [3], [4]. With the high quality and availability of synthetic data, radiologists can improve the accuracy in diagnosing, while researchers can understand image perception and innovate new techniques in medical imaging. Therefore, the need for a generative model that can capture the data distribution of the medical image is essential. In addition, the generative model has to have the ability to deal with the limited amount of training data.

In the computer vision field, researchers have proposed different kinds of generative models, in which GANs have been the most popular [5]. However, the training process of GANs is difficult to optimize and easily falls into collapse mode [6]. On the other hand, the diffusion model has gained lots of attention from researchers due to its potential and ability to beat GANs in image generation [7]. The diffusion model is likelihood-based, so it can estimate the distribution of data and also sample quality better than other methods. However, the computation cost of the diffusion model is extremely high and raises a critical problem for researchers. It is even more challenging when the training data is limited in size.

In this paper, our goal is to synthesize the medical image using a lightweight diffusion model while maintaining the image quality. Our model will be trained on a public dataset; the SARS-CoV-2 CT scan dataset, with only over two thousand images. Our contributions can be summarized as follows:

- Firstly, we evaluate the performance of the diffusion model in image generation for the medical dataset. This is the first paper to consider this approach.
- Secondly, we propose a lightweight diffusion model with U-Net architecture to reduce the amount of computation so that it fits with our dataset.
- Thirdly, we point out the essential components for a diffusion model and how to boost its performance with a simple technique effectively.
- Finally, we conduct extensive simulations to analyze the performance of the proposed diffusion model, and also to prove that the diffusion model can be utilized in generating high-quality images for the medical field.

The remaining part of the paper will be organized as follows. First, we will outline some related works in the next section. Then, we will briefly formulate the problem model and then propose our method in section III. Finally, numerical results are demonstrated in Section IV, and based on that we conclude our paper in Section V.

II. Related works

A. Diffusion Model

The diffusion model was first introduced by *Jascha Sohl-Dickstein* in [8], at the time, its performance was not impressive to researchers. Recently, authors in [7] have demonstrated

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068 and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A4A1018607) *Dr. CS Hong is the corresponding author.

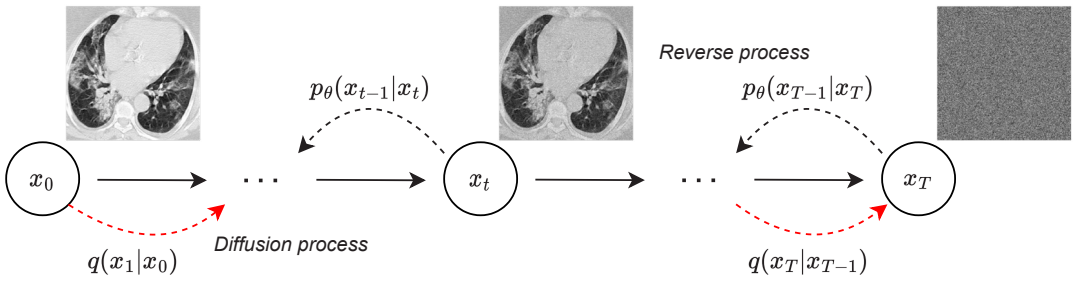

Fig. 1. The diffusion model

its potential in image generation by changing its loss function, which achieves more attention. The diffusion model's idea is to add noise into the image following a Markov chain and then keep doing this step by step. After T timesteps, the original image becomes complete noise, and then the noise-adding process is stopped, and it is called the diffusion process. From the complete noise image, a deep neural network is used to reverse the diffusion process, subtracting the noise from the image gradually through T timesteps, which is called the reverse process.

B. Medical Image Generation

Most of the works related to medical image generation use GANs as their generative model [1], [2], [9], [10]. The training process of GANs is very difficult to optimize due to many problems: mode collapse (in which the model only generates a particular image) and failure to converge. In [1], GANs was utilized to generate and manipulate the synthesis image for medical image studies. While authors in [2] adopted another variation of GANs, called PGGAN, to promote medical image analysis, particularly in gastric X-ray images. Both [9] and [10] tried to improve the realistic of the generated images so they can be useful in medical imaging tasks. Up to date, there is only one work [11] that considers the diffusion model in the medical field, but the 4D-MRI dataset used in that paper has a very large size and is also high quality, which consumes expensive computation.

III. Diffusion Mechanism and Proposed Model

As aforementioned in Fig. 1, the diffusion model includes two stages: the diffusion process (forward process) and the reverse process. The forward process can be modeled by a Markov chain, which converts the image distribution into a well manner distribution by adding noise step by step, while reverse process intends to recover the image distribution from a Gaussian noise distribution.

A. The forward process

With a given image x_0, its data distribution will be denoted as $q(x_0)$, and the noise-adding phrase has to follow a standard method; in this paper, we use the most popular method-Gaussian noise. Therefore, this Gaussian diffusion process can be mathematically modeled as follows:

$$q(x_1|x_0) := \mathcal{N}(x_1; \sqrt{1-\beta_1}x_0, \beta_1 \boldsymbol{I}), \tag{1}$$

where β_1 is the variance at timestep one, and $\boldsymbol{I}$ is the identity matrix; normally the value of this variance is small and increases throughout timesteps. At a starting point, we can estimate the data distribution after T steps diffusion with the following equation:

$$q(x_T|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

After reasonable timesteps, the given image is going to be complete noise, as shown in Fig. 1.

B. The reverse process

After the forward process, we want to recover the original image from the noisy image through reverse. To do this, a deep neural network has been proposed to estimate the mean and variance of each diffusion step and then subtract them from the noisy image, which can be described as follows

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{3}$$

where $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ are estimated mean and variance value of added noise at timestep t. In the work [7], they found that it is better for the network to estimate one value while fixing the other, and also indicated fix $\Sigma_\theta(x_t, t) = \beta_t \boldsymbol{I}$ is good option.

The outstanding property of the diffusion model is that we can sample x_t at any arbitrary timesteps as follows:

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{4}$$

where $\bar{\alpha}_t$ can be calculated by $\Sigma_{s=1}^{t}\alpha_s$ and $\alpha_t := 1 - \beta_t$. We can express the above equation in an easier way as

$$x_t := \sqrt{\bar{\alpha}_t}x_0 + (1-\bar{\alpha}_t)\boldsymbol{I}. \tag{5}$$

From experiments, reparameterizing the $\boldsymbol{I}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ can give the model a performance boost. For reverse process, we also take advantage of this property and try to approximate the original image from timestep t as follows:

$$x_0 := \frac{(x_t - (1-\bar{\alpha}_t)\boldsymbol{\epsilon})}{\sqrt{\bar{\alpha}_t}} \tag{6}$$

where $\boldsymbol{\epsilon}$ is the noise added to x_0. Therefore, at this step, our network only needs to estimate $\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}_\theta(x_t, t)$ instead of the mean value.

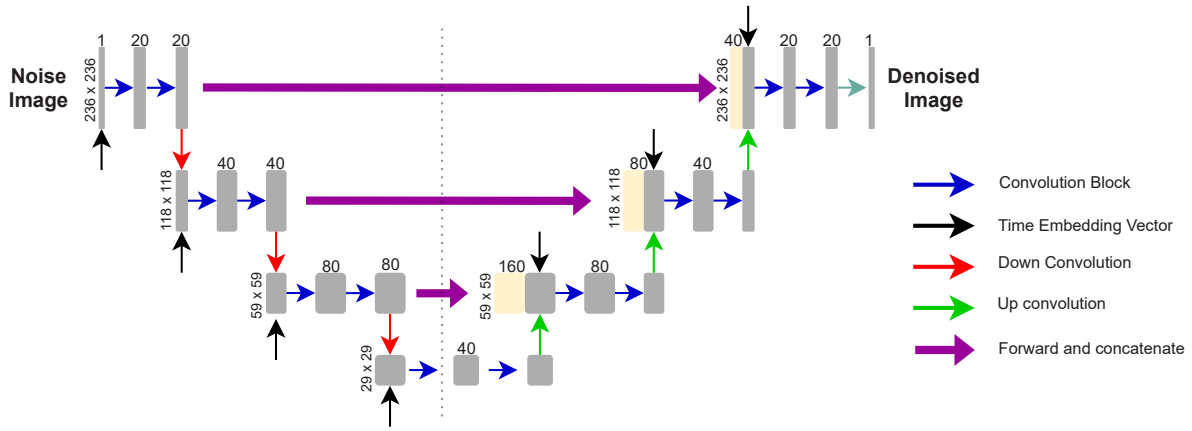

Fig. 2. Proposed network architecture with multiple time embedding

C. Customized U-net with Time Embedding

Based on the mechanism of the reverse process explained in the above subsection, our deep neural network implicitly needs to do two operations: estimate the noise distribution of the input image, and subtract the noise from the input image with the estimated statistics. With this procedure in mind, we recognize the U-NET architecture [12] can be a potential candidate. However, we still have lots of work to guarantee the model's performance. First, we will discuss the reasons why do we choose this U-Net architecture for the neural network model and then how we embed the timestep information into it.

1) Customized U-Net: As we can see in the architecture of our customized U-Net in Fig. 2, we divide the network into two portions: the first half and the second half. The first half of the network is used to map the input image into its latent space using a sequence of convolution blocks and down convolutions operation. By doing this mapping procedure, we can detect abnormal (noise) in the input image. Based on that latent space, the second half of the network works as a decoder, and its goal is to generate a representative image through multiple convolution blocks and up-convolution operations. Besides that, the feature maps from the first half are forwarded and concatenated (skip connection) with the corresponding feature maps in the second half. This mechanism is the key to the success of this diffusion model because it affects the network performance the most. We prove the above claim in the experiment result. With the feature maps of the input image and the representative image, the neural networks can accurately subtract the noise from the input image throughout T timesteps.

2) Time Embedding: We re-use the above network for every single timestep, and also with different timesteps, the noise value that needs to be estimated is different. Therefore, we have to embed the timestep information into the network so that we can guide it in the proper direction. If we only embed the timestep vector at the beginning of the network, it rarely affects the outcome of the network. We have done intensive simulations to find a good way to inject the timestep information into the network. Among them, the best way is to embed it into all the feature maps that result from down convolution or up convolution, as shown in Fig.2. This phenomenon can be easily understood as we frequently remind the network at which timesteps we are in. Therefore, the network can precisely estimate the noise value and return a better-denoised image.

D. Loss Function

In this paper, we focus on designing a lightweight architecture for the reverse process instead of improving the model's performance. Therefore, we re-use the simple loss function in [7], which is denoted as follows:

$$L_{\text{simple}} = E_{t,x_0,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(x_t, t)||^2\right]. \tag{7}$$

This simple loss function is a reweighted form of the variational lower bound (VLB) and achieves better sample quality.

IV. Performance Evaluation

A. Medical Dataset

We test the performance of our proposed architecture with the SARS-CoV-2 CT scan dataset, which is publicly available on Kaggle [13]. It contains nearly 2500 images of CT - Scan images from real patients, half of which are positive for COVID. Nevertheless, our goal is to verify the ability of the diffusion model in the medical field, where a large amount of data is not always feasible, so we will treat scan images from both COVID positive and negative the same. These images vary in size, so we need to resize them to the same size to guarantee the stability of the training process, particularly in this case, 236 x 236.

B. Evaluation metric

A number of evaluation metrics for the generated images have been proposed by researchers; among them, the Frechet Inception Distance (FID) score metric is the most popular [14]. Researchers have used it to determine the state of the art of the generative models, and it is an improved version of the Inception Score (IS). The FID score not only evaluates the quality of the synthesis image but also compares its statistics

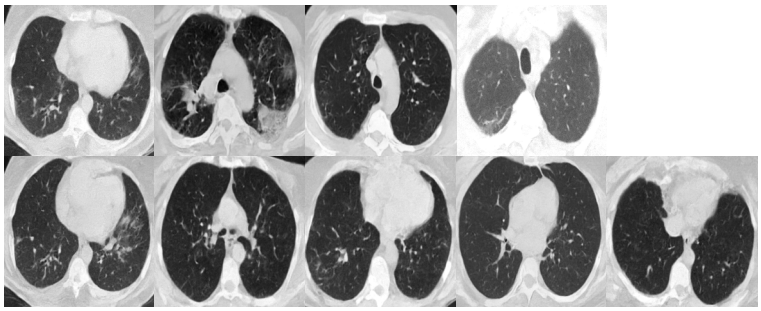

Fig. 3. Generated images by one-time embedded and proposed time embedded models

TABLE I
FID SCORE

Model	FID
Encoder-decoder architecture	548.73
U-Net with one-time embedded	73.234
U-Net with four-time embedded	76.294
Proposed U-Net	***45.627***

with the original image statistics. The lower this value, the better the synthesis images. We evaluate the performance of our proposed model with the benchmark schemes below.

Encoder-decoder architecture: We can implement this network architecture easily by eliminating all the forward and concatenate from our network. With this baseline, we can prove our claim about the importance of forward and concatenate operations.

U-net with one-time embedded: In this baseline, we only embed the time vector into our network one time, at the beginning. U-net with four-time embedded is implemented by injecting the time vector into the first half of the network. From these, we gain valuable insight into how the time-embedding vector affects the model performance.

As shown in Table. 1, our proposed model achieves the lowest FID score. In addition, there are two key ideas we can take away from this table. First, we can recognize the performance of the encoder-decoder architecture performs poorly with a high FID score; the generated images by this model are just organized noise. Therefore, the first takeaway is that the forward and concatenate operations are the key to boosting the performance of the diffusion model. Secondly, the time embedding vector should be carefully injected into the network. Otherwise, it degrades the performance, as shown in one-time and four-time embedded cases.

In Fig. 3 we demonstrate the generated image from two cases. The first row contains images from one time embedded case, and we can say the generated samples are acceptable. However, the model performance is not stable, in some cases, it even generated a completely white image in the first row. In contrast, our model guarantees the sample quality and provides consistency in image generation, as shown in the second row. None of the samples from our model appears to be completely white.

V. CONCLUSION

We have proposed a lightweight version of the diffusion model to synthesize images in the medical field, where a large number of data training is unavailable. In addition, we have shown the importance of the forward and concatenate feature maps in our proposed model by conducting extensive simulations. Moreover, we also proved that the way we inject the step information into our network affects the model performance. The difference between original images and generated images is insignificant since the FID score is relatively low. Therefore, our proposed lightweight model can be useful for devices with limited computation power.

REFERENCES

[1] Z. Ren, X. Y. Stella, and D. Whitney, "Controllable medical image generation via gan," *Journal of Perceptual Imaging*, vol. 5, pp. 1–15, 2022.

[2] R. Togo, T. Ogawa, and M. Haseyama, "Synthetic gastritis image generation via loss function-based conditional pggan," *IEEE access*, vol. 7, pp. 87 448–87 457, 2019.

[3] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 417–425.

[4] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[8] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[9] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "Gan-based synthetic brain mr image generation," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 734–738.

[10] Y. Jiang, H. Chen, M. Loew, and H. Ko, "Covid-19 ct image synthesis with a conditional generative adversarial network," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 441–452, 2020.

[11] B. Kim and J. C. Ye, "Diffusion deformable model for 4d temporal medical image generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 539–548.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, "Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification," *MedRxiv*, 2020.

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