



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사 학위논문

Score-based Diffusion model 을 이용한 의료 이  
미지 처리의 강화법

Enhancement of Medical Imaging Processing with Score-based  
Diffusion Model

울산대학교 대학원

의 과학 과

정지현

Score-based Diffusion model 을 이용한 의료 이  
미지 처리의 강화법

지도교수 김 남 국

이 논문을 공학석사 학위 논문으로 제출함

2023 년 06 월

울산대학교 대학원  
의 과 학 과  
정 지 현

정지현의 공학석사학위 논문을 인준함

심사위원    김   남   국            ( 인 )

심사위원    박   지   은            ( 인 )

심사위원    이   준   구            ( 인 )

울 산 대 학 교 대 학 원

2023 년 06 월

## **Abstract**

Generative models can be very useful in the field of medical imaging. These models can be used to address data imbalance issues or to transform to different modalities. Additionally, 3D generation can be applied to clinical research, distribution analysis and more. However, medical images are more complex than natural images, making generation difficult. This means that a lot of effort is needed to create plausible generation and that it is challenging for generative models to excel in the medical image field. Nonetheless, recent advances in diffusion models have made it possible to generate high-quality images, and the use of latent diffusion models has also solved the issue of generation speed. Therefore, this paper proposes experiments on generation using diffusion models, data augmentation through generation, image-to-image transformation, 3D generation, and predicted generation. The results of this study can significantly impact the field of medical imaging by providing more accurate and comprehensive diagnostic tools for medical professionals. The use of diffusion models can also reduce the time and effort required for medical image generation and improve the overall quality of medical images, leading to better treatment outcomes for patients. This paper provides a comprehensive overview of the technical implementation and clinical applications of score-based diffusion models in medical imaging and highlights their potential to revolutionize the field of medical imaging diagnosis.

## 차 례

영문요약	i
그림목차	v
1. Introduction	1
2. Background	4
2.1. GAN	5
2.1.1. Generative Adversarial Networks	5
2.1.2. Conditional GAN	6
2.2. Variational Auto-Encoders	7
2.3. Score-based generative model and diffusion model	7
2.3.1. Noise Conditional Score Networks (NCSNs)	8
2.3.2. Denoising Diffusion Probabilistic Models (DDPMs)	9
2.3.3. Score SDE	11
2.3.4. Consistency generation	12
2.3.5. Controllable generation	13
2.3.6. Latent Diffusion Models (LDMs)	16
3. High-bit depth generation in computed tomography	17
3.1. <i>Dataset and model architecture</i>	17
3.2. <i>Results of GAN</i>	18
3.3. <i>Results of score-based diffusion model</i>	20
4. <i>Data augmentation with isocitrate dehydrogenase type in glioma</i>	23
4.1. <i>Dataset and model architecture</i>	24
4.1.1. Dataset	24
4.1.2. <i>IDH mutation status and image preprocessing</i>	26
4.1.3. <i>Imaging phenotype</i>	27
4.2. <i>Results</i>	29

4.2.1. <i>Evaluation by human readers</i> .....	29
4.2.2. <i>Deep learning-based prediction of IDH type using the real data and nonselective GMA</i> .....	31
4.2.3. <i>Deep learning-based prediction of IDH type using imaging phenotype-based GMA according to tumor size</i> .....	32
4.2.4. <i>Deep learning-based prediction of IDH type using imaging phenotype-based GMA according to tumor size</i> .....	34
5. <i>Image-to-image translation with H&amp;E staining normalization of whole slide imaging</i> .....	37
5.1. <i>Dataset</i> .....	38
5.2. <i>Stain normalization without stain separation</i> .....	38
5.3. <i>Stain normalization with stain separation</i> .....	39
5.4. <i>Performance evaluation criterion</i> .....	40
5.5. <i>Result</i> .....	41
5.5.1. <i>Quantitative and Qualitative Results</i> .....	41
5.5.2. <i>Result of overlapping</i> .....	44
6. <i>3D generation in brain CT</i> .....	46
6.1. <i>Material and Methods</i> .....	47
6.1.1. <i>Adjacent Slice-based Conditional Iterative Inpainting, ASCII</i> .....	47
6.1.2. <i>Result of ASCII</i> .....	48
6.1.3. <i>Result of ASCII in 12-bit whole range</i> .....	49
6.1.4. <i>Intensity Calibration Network</i> .....	49
6.1.5. <i>Dataset and model architecture</i> .....	52
6.2. <i>Experiments</i> .....	53
6.2.1. <i>Results of intensity calibration network</i> .....	53
6.2.2. <i>Results of ASCII with IC-Net in whole range</i> .....	55
6.2.3. <i>Quantitative Evaluation</i> .....	56
6.2.4. <i>Qualitative Evaluation</i> .....	58
7. <i>Post-surgery imaging generation in cephalogram</i> .....	59
7.1. <i>Material and Methods</i> .....	60

7.1.1. <i>Datasets</i> .....	60
7.1.2. <i>Model architecture</i> .....	61
7.2. <i>Surgical Movement Prediction</i> .....	63
7.3. <i>Post-operation imaging generation</i> .....	63
8. <i>Discussion</i> .....	67
9. <i>Conclusion</i> .....	68
<i>Reference</i> .....	68
<i>국문 요약</i> .....	80



## 그림목차

Figure 1	10
Figure 2	24
Figure 3	25
Figure 4	26
Figure 5	27
Figure 6	32
Figure 7	34
Figure 8	41
Figure 9	44
Figure 10	45
Figure 11	47
Figure 12	50
Figure 13	51
Figure 14	53
Figure 15	54
Figure 16	56
Figure 17	57
Figure 18	58
Figure 19	59
Figure 20	60
Figure 21	61
Figure 22	67
Figure 23	69

## ***1. Introduction***

Generative models are highly valued for their potential applications in medical imaging. They are highly effective in analyzing and understanding unlabeled information, which is why they are preferred. In addition to generating images, generative models can be used to synthesize missing disease groups for data augmentation to address data imbalances, or to easily obtain images in different modalities through image-to-image translation. They can also generate 3D images with clinical integrity for research purposes or create normal images that are most like abnormal ones for anomaly detection. Therefore, generative models are widely used and receiving a lot of attention in many research studies.

Generative models are widely used in deep learning. They can be used to create high-fidelity data, improve text-to-image performance or semi-supervised performance, or detect anomalies. There are three main approaches to generative models: likelihood-based models (variational auto-encoder (VAEs) [1-3], autoregressive models [4-6], normalizing flows [7-9]), generative adversarial networks [10-17] (GANs), and diffusion-based models (score-based generative models [18-23] and diffusion models [24-33]). Each of these models has its own advantages and disadvantages. Likelihood-based models estimate the distribution and calculate likelihood to sample, but they may produce low-quality samples. Adversarial models can produce high-quality samples, but training can be unstable due to adversarial learning. Diffusion-based models can also produce high-quality samples, but the sampling process is slow due to the stochastic process. Therefore, it is important to choose the appropriate generative model based on the desired generation purpose.

Medical images are different from natural images because they often contain structures such as diseases or anatomical structures [34, 35]. For example, mammography images can contain irregular patterns of fatty tissue and lobules, while retinal images can contain patterns of the optic disk and blood vessels. These complex structures can be challenging to generate using deep generative model. In addition, conditions such as tumors

or hemorrhages can also have varied patterns that make generation difficult.

The recently introduced score-based diffusion model, which encompasses both score-based generative models and diffusion models, has shown promising performance and potential. It uses Langevin dynamics based on gradients of the perturbation distribution caused by perturbation data for sampling or reverses the noise-added Markov chain called the diffusion process to generate data. In addition, through a paper that interprets these two different sampling methods as a stochastic differential equation, the model has demonstrated performance that can rival that of GANs. As a result, it has not only enabled the generation of high-quality images, but also audio and video, and has shown remarkable performance in multi-modal generation tasks such as text-to-image generation.

Therefore, in this paper, we propose the use of diffusion models for generating medical images and their potential applications. First, we compare the generation ability of diffusion models with that of another generative model, GANs, through the generation of images in multiple modalities. Next, we demonstrate the use of diffusion models for data augmentation. To overcome a problems with limited data sets, we show how data augmentation using generative models can improve the performance of classifiers on imbalanced datasets and present the results when clinical bias is introduced through data augmentation. Third, we propose the use of image-to-image translation using diffusion models for stain normalization. Fourth, we propose a method for generating CT images with 3D integrity and continuity using the generation of adjacent slices. Finally, we propose predicting and generating post-operation lateral cephalogram images by using the diffusion model.

Score-based diffusion models (SDM) [19, 21, 33] have demonstrated significant potential in diverse fields, including image generation and super resolution, and their generative abilities are comparable to those of GANs and VAEs. SDMs have also shown

promising results in colorization and inpainting. Therefore, researchers have explored the application of SDMs in stain normalization.

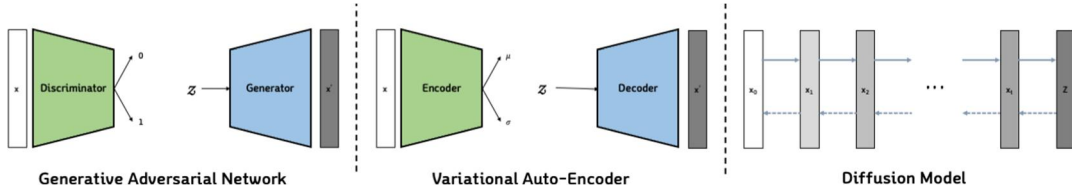
Deep generative models can generate realistic imaging data that can be used as data input for deep learning algorithms, thereby overcoming problems with limited data sets [36]. Data augmentation using generated images can improve the accuracy of deep learning model classifications for rare tumor types and unbalanced classes [37, 38]. Although synthetic image generation is of particular interest for clinical applications in brain tumor imaging because of the inherently small sample sizes in imaging-based genomic and molecular prediction, few studies have evaluated the performance of diagnostic models using generated images [39].

## *2. Background*

GAN [11] is a deep learning model that consists of a generator and a discriminator. The generator generates samples that mimic real data, and the discriminator tries to distinguish between real and fake samples. These two models compete, and as a result, the generator learns to create more realistic samples, while the discriminator learns to become more accurate in its classification. GANs have been successfully used in various fields, including image generation, video generation, and text generation.

VAE [2] is a generative model that consists of an encoder and a decoder. The encoder maps the input data into a latent space, and the decoder generates samples from this latent space. The VAE is trained to maximize the probability of the input data given the latent space, while minimizing the divergence between the latent space distribution and a prior distribution. The VAE can be used for various tasks, including image generation, data compression, and feature extraction.

Diffusion-based models [19, 21, 22, 24, 27, 31-33, 40] are a class of generative models that learn a diffusion process that can transform a simple noise distribution into a complex data distribution. Diffusion models are based on the idea of simulating a random walk through the data distribution by iteratively adding noise to the data. The diffusion process is modeled using a series of steps, where each step consists of a noise injection and a step of a trainable neural network. Diffusion models have been shown to be highly effective in generating high-quality images, and they have several advantages over other generative models, including improved generation speed, better stability during training, and the ability to model complex distributions with high-dimensional data. **Figure 1** provides an overview of the three models.



**Figure 1.** Overview of three types of generative models.

## 2.1. GAN

### 2.1.1. Generative Adversarial Networks

GANs [11] are one of the most important models in the field of generative models. GANs are trained using an adversarial training method of the discriminator, where the discriminator trains to distinguish between real and fake data, while the generator trains to generate data that is real by the discriminator. The adversarial training method is an iterative competition between the discriminator and the generator, with the goal of generating synthetic data that is indistinguishable from real data by the discriminator.

To learn the data distribution  $p(x)$  of a target dataset, GANs set up a min-max game between two neural networks: a generator and a discriminator. With a random noise vector  $z$  sampled from a straightforward prior distribution  $p(z)$ , such as a standard normal or uniform distribution, as input, the generator  $G$  attempts to generate samples  $G(z)$  that show up realistic and resemble the data. A real data sample  $x$  sampled from  $p(x)$  or a fake sample  $G(z)$  generated by  $G$  are sent to the discriminator  $D$ , which attempts to properly identify which is real or fake. The objective function of GAN is given by:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x \sim p(x)} [\log(D(x))] + \mathbb{E}_{z \sim p(z)} [1 - \log(D(G(z)))]$$

And there are various models with different objective functions, including LSGAN [41] that replaces binary cross entropy with mean square error, EBGAN [42] that uses margin loss, WGAN [43] that replaces the loss function with Wasserstein distance, and WGAN-GP [44] that adds a gradient penalty term. There are also techniques that can help GAN training. For

example, progressive growing of image resolution [45], spectral normalization [46] to stabilize the model by fixing its Lipschitz constant, and StyleGAN [15] which adds adaptive instance normalization (AdaIN) [47] with the addition of a style latent.

### 2.1.2. Conditional GAN

By providing labels to the generator and discriminator, it is possible to train the model for conditional generation. The objective function for a given condition  $y$  can be expressed as follows:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x \sim p(x)} [\log(D(x, y))] + \mathbb{E}_{z \sim p(z)} [1 - \log(D(G(z, y), y))]$$

where  $x$  is the input,  $z$  is the random noise,  $G$  is the generator, and  $D$  is the discriminator.

For example, it is possible to generate by providing a label for a class together with noise [48] or perform image-to-image translation using a well-aligned label image [49, 50]. However, in such cases, it is necessary to have well-aligned pairs of condition-image. To overcome this, unpaired image-to-image translation from domain A is proposed. This is achieved by transforming from one domain A to another domain B and then back to domain A, with the addition of mean absolute error (MAE) loss during the reconstruction process. This approach involves training two generators and two discriminators, which is called CycleGAN [51].

The loss function of CycleGAN is as follows:

$$\mathcal{L}_{total}(G_X, G_Y, D_X, D_Y) = \mathcal{L}_{GAN}(G_X, D_Y, X, Y) + \mathcal{L}_{GAN}(G_Y, D_X, Y, X) + \lambda \mathcal{L}_{cycle}(G_X, G_Y),$$

where  $G_X$  and  $G_Y$  are generators that perform transformations from  $X$  to  $Y$  and from  $Y$  to  $X$ , respectively, while  $D_X$  and  $D_Y$  are discriminators that distinguish between real and fake images in  $X$  and  $Y$ , respectively.  $\lambda$  is a weighting factor that controls the cycle-consistency loss. And  $\mathcal{L}_{GAN}(G_X, D_Y, X, Y)$  is an objective function of GAN and  $\mathcal{L}_{cycle}(G_X, G_Y)$  is cycle-consistency loss. Moreover, there are methods to provide conditions by not only simply concatenating them but also normalizing them spatially, which is called spatially adaptive (de)-normalization (SPADE) [52].

## 2.2. Variational Auto-Encoder

VAE [2] is a generative model that learns the underlying distribution of a dataset and generates new samples from it. VAE is a probabilistic approach to autoencoding, which maps an input into a latent space and then reconstructs it. In VAE, the encoder and decoder are modeled as probabilistic neural networks that allow us to model complex distributions. VAE learns a latent representation of the input data by maximizing the evidence lower bound (ELBO) objective. ELBO is a lower bound on the log-likelihood of the data, which is decomposed into two terms: reconstruction loss and KL divergence between the learned latent distribution and a prior distribution.

$$ELBO = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - D_{KL}(q_{\phi}(z|x) || p(z))$$

Here,  $q_{\phi}(z|x)$  is the encoder that approximates the true posterior distribution  $p(z|x)$ ,  $p_{\theta}(x|z)$  is the decoder that models the likelihood of generating the data  $x$  given the latent variable  $z$  and  $D_{KL}(q_{\phi}(z|x) || p(z))$  is the Kullback-Leibler (KL) divergence between the approximate posterior  $q_{\phi}(z|x)$  and the prior distribution  $p(z)$ . The first term encourages the reconstruction accuracy, while the second term encourages the learned distribution of  $z$  to match the prior distribution. The ELBO serves as a lower bound to the log-likelihood of the data, which is intractable to compute directly.

VAE has been shown to be effective in generating realistic images and has been widely used in various fields such as image generation, image inpainting, and data compression. However, VAE suffers from the posterior collapse problem, which occurs when the learned latent distribution becomes uninformative about the input data. This problem can be mitigated by using various regularization techniques such as hierarchical architecture [3], vector quantized [1, 53], and beta-VAE [54].

## 2.3. Score-based generative model and diffusion model

Score-based generative model [19, 21, 22, 33] and diffusion model [24, 27, 31, 32, 40] were



shown remarkable performance in the generative task, such as image generation, super-resolution [55], video generation, inpainting [56], and text-to-image [25, 26, 57]. Two models succeeded to generate high-fidelity data without another auxiliary network, such as a discriminator in a GANs or a gaussian encoder in a variational auto-encoder VAEs. Two models have the forward-backward process; the forward process perturbs data into noise in each step, whereas the backward process generates noise into data using a stochastic process. And the score-based generative model, such as noise conditional score networks (NCSNs) [33], and the diffusion model, such as denoising diffusion probabilistic models (DDPMs) [31], were inspired by Langevin dynamics and thermodynamics, respectively, and a generalized model was proposed using stochastic differential equations (SDEs) in [19].

### 2.3.1. Noise Conditional Score Networks (NCSNs)

NCSN is based on the (*stein*) score [23] of the logarithmic data density, which is the gradient of log density at data  $\nabla_x \log p(x)$ . The model estimates the gradient of log density at data  $\nabla_x \log p(x)$  and the objective function of score matching was defined as following:

$$\mathbb{E}_{x \sim p(x)} [\|s_\theta(x) - \nabla_x \log p_{data}(x)\|_2] \approx \mathbb{E}_{x \sim p(x)} \left[ \text{tr}(\nabla_x s_\theta(x)) + \frac{1}{2} \|s_\theta(x)\|_2^2 \right] + C$$

where  $s_\theta(x)$  is score network. However, the score function of data density  $\nabla_x \log p_{data}(x)$  is generally challenging to calculate and the resource of calculating  $\text{tr}(\nabla_x s_\theta(x))$ , which is the Jacobian of  $s_\theta(x)$ , is high. To avoid intractable objective, denoising score matching [18] and sliced score matching [20] was proposed; the denoising score matching employed the score function of perturbed data density, not data density  $\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x) = -\frac{\tilde{x}-x}{\sigma^2}$ , where  $p_\sigma(\tilde{x}|x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I)$  is a perturbation kernel, and sliced score matching used random projections to approximate  $\text{tr}(\nabla_x s_\theta(x))$  by using  $\mathbf{v}^T \nabla_x s_\theta(x) \mathbf{v}$ , where  $\mathbf{v} \sim p(\mathbf{v})$  is a random vector of simple distribution, e.g., the multivariate standard normal distribution. However, there are several issues to generate a data due to problems related to manifold hypothesis. For

example, the inconsistent approximation of score function  $s_\theta(x)$  in low-density region was caused by low-dimensional manifolds embedded in a high-dimensional space. The authors solved the problem by perturbing the data with several scales of Gaussian noise.

The  $\{\sigma_i\}_{i=1}^T$  is a positive geometric sequence that satisfies  $\gamma = \frac{\sigma_i}{\sigma_{i+1}} > 1$ , for all  $i$ . Then, the objective function of denoising score matching can be written as:

$$\frac{1}{T} \sum_{t=1}^T \lambda(\sigma_t) \mathbb{E}_{p(x)} \mathbb{E}_{x_t \sim p_{\sigma_t}(x_t|x)} \|s_\theta(x_t) - \nabla_{x_t} \log p_{\sigma_t}(x_t|x)\|_2^2$$

where  $\lambda(\sigma_t)$  is a coefficient function depend on  $t$ . After training score network, the generation process was performed by Langevin dynamic, which can generate samples from a density using only score function. Finally, an initial value  $x_0 \sim \pi(x)$  with  $\pi$  being prior distribution, the Langevin dynamic compute the following as:

$$x_t = x_{t-1} + \frac{\epsilon}{2} \nabla_{x_{t-1}} \log p(x_{t-1}) + \sqrt{\epsilon} z_t$$

where  $z_t$  is Gaussian noise. When  $\epsilon \rightarrow 0$  and  $T \rightarrow \infty$ , the distribution of  $x_T$  converges the data distribution. As well as the authors proposed annealed Langevin dynamic, which starts with prior distribution, e.g., uniform distribution, and applies Langevin dynamic for a fixed number of iterations.

### 2.3.2. Denoising Diffusion Probabilistic Models (DDPMs)

**Forward process** The data distribution  $q(x_0)$  is gradually converted into a well-behaved distribution  $\pi(y)$  by repeated application of a Markov diffusion kernel  $T_\pi(y|y; \beta)$  for  $\pi(y)$ . Then,

$$q(x_t|x_{t-1}) = T_\pi(x_t|x_{t-1}; \beta_t) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I})$$

And the forward trajectory, starting at the data distribution and performing  $T$  steps of diffusion process, is following as:

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^T q(x_t|x_{t-1})$$

where  $x_1, x_2, \dots, x_T$  are latents of the same dimension as the data  $x_0$ . The forward process is that is that admits sampling  $x_t$  at an arbitrary timestep  $t$  in closed form: using the notation  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \sum_{s=1}^t \alpha_s$ , then, we obtain the analytical form of  $q(x_t|x_0)$ :

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

And we can easily obtain a sample in imedietate distribution of difussion process,

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

**Backward process** Diffusion models are latent vaiable models of the parameterized distribution  $p_\theta(x_0) = \int p_\theta(x_{0:T})dx_{1:T}$ . The reverse trajectory, starting at the prior distribution, is following as:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t)$$

where  $p(x_T) = \pi(x_T)$  and  $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ .  $\mu_\theta(x_t, t)$  and  $\Sigma_\theta(x_t, t)$  are training targets defining the mean and covariance of the reverse Markov transitions for a Gaussian distribution, respectively. To approximate between the parameterized distribution  $p_\theta(x_0)$  and data distribution  $q(x_0)$ , training is performed by optimizing the variational lower-bound on negative log likelihood:

$$\mathbb{E}_{x \sim q(x)}[-\log p_\theta(x)] \leq \mathbb{E}_{x \sim q(x)} \left[ -\log p(x_T) - \sum_{t \geq 1} \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] = \mathcal{L}_{vlb}$$

For efficient training, further improvement came by rewriting  $\mathcal{L}_{vlb}$ :

$$\begin{aligned} \mathcal{L}_{vlb} = \mathbb{E}_{x \sim q(x)} [ & D_{KL}(q(x_T|x_0) || p(x_T)) + D_{KL}(q(x_{t-1}|x_t, x_0) || p_\theta(x_{t-1}|x_t)) \\ & - \log p_\theta(x_0|x_1) ] \end{aligned}$$

And, above equation used KL divergence to directly compare  $p_\theta(x_{t-1}|x_t)$  against forward process posteriors. The posterior distributions are tractable when conditioned on  $x_0$ :

$$q(x_t|x_{t-1}) = q(x_{t-1}|x_t, x_0) \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)} = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

where  $\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_{t-1})}}{1-\bar{\alpha}_t}x_t$  and  $\tilde{\beta}_t = \frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}\beta_t$ .

Like NCSN, the loss function can write:

$$L_{simple} = \mathbb{E}_q [\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$$

After training, sample can be generated by starting from  $x_T \sim \mathcal{N}(0, \mathbf{I})$  and following the parameterized reverse Markov chain as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z$$

The generative process is still defined by  $p_\theta(x_{t-1}|x_t)$ , but the network predicts the noise from the perturbed image, not the mean and the covariance directly.

### 2.3.3. Score SDE

Score-based generative model and DDPMs can generate a high-fidelity image by using perturbing data with diffusion process of multiple noise scales. Therefore, the diffusion process can be generalized an infinite number of noises scales, and the perturbed data distributions of diffusion process construct according to a stochastic differential equation (SDEs). The diffusion process can be indexed by a continuous time variable, such that  $x_t$  and  $p_\theta(x_t)$  are data and prior distribution. The forward diffusion process evolves according to the Ito stochastic differential equation:

$$dx = f(x, t)dt + g(t)dw$$

where  $dw$  is the standard Wiener process (a.k.a, Brownian motion) and  $f(x, t)$ ,  $g(t)$  are a drift coefficient and diffusion coefficient of  $x_t$ . It is known that any diffusion process can define reverse-time diffusion process by B. The reverse-time SDE is given as:

$$dx = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\bar{w}$$

where  $dt$  and  $d\bar{w}$  are an infinitesimal negative time step and standard Wiener process when time flows backwards. NCSNs and DDPMs correspond to variance-exploding SDE (VESDE) and variance-preserving SDE (VPSDE), respectively. VESDE is given as:

$$dx = \sqrt{\frac{d[\sigma(t)^2]}{dt}} dw$$

And VPSDE is given as:

$$dx = -\frac{1}{2}\beta(t)xdt + \sqrt{\beta(t)}dw$$

Finally, training is performed so that the score network estimates the score function by generalizing the score matching objective, the following objective:

$$\mathbb{E}_{t \sim U[0, T], x_0 \sim p(x_0), x_t \sim p(x_t | x_0)} \left[ \lambda(t) \|s_\theta(x_t, t) - \nabla_{x_t} \log p_{0t}(x_t | x_0)\|^2 \right]$$

To solve the SDE, a numerical solver was used with a score-based Markov chain Monte Carlo (MCMC) approach, specifically a Langevin MCMC. The Predictor-Corrector (PC) sampler alternates between using a predictor, such as a reverse diffusion SDE solver [19], which computes the reverse-time SDE with a fixed discretization strategy, and a corrector, such as annealed Langevin dynamics, which can adjust the direction of gradient ascent. This allows us to accurately estimate the probability of the data being generated, which is essential for the MCMC approach.

#### 2.3.4. Consistency generation

The three models referred to as *NCSN*, *DDPM*, and *ScoreSDE* are stochastic models that lack consistency and don't have a paired between noise and data. While it offers significant advantages from a generative perspective, it is questionable whether this is advantageous in terms of reproducibility or latent representations. Therefore, consistency sampling is possible through the probability flow ordinary differential equation (PF ODE) used in the SDE-based framework [19] and denoising diffusion implicit model (DDIM) [24] sampling used in the diffusion-based model. The probability flow sampling method uses a score model, and the equation is as follows:

$$dx = \left[ f(x, t) - \frac{1}{2}g(t)^2 \nabla_x \log p_t(x) \right] dt$$

Unlike SDE sampling, there is no diffusion term, and it is also called a deterministic process. Sampling is possible using numerical methods such as Euler-Maruyama [58] or stochastic Runge-Kutta methods [59] that approximate SDE using the score function. Next, the DDIM sampling method also allows for consistency generation by setting the term corresponding to the diffusion term to zero in the DDPM sampling method. If we separate the "predicted  $x_0$ " term and the "direction pointing to  $x_t$ " term in the DDPM sampling method, the equation becomes:

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}} \left( x_t - \frac{\sqrt{1 - \bar{\alpha}_t} (1 - \alpha_t)}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sqrt{(1 - \alpha_{t-1} - \sigma_t^2)} \epsilon_\theta(x_t, t) + \sigma_t z$$

Setting  $\sigma_t = 0$  here enables consistency sampling, and this sampling method is called denoising diffusion implicit model (DDIM).

### 2.3.5. Controllable generation

---

**Algorithm 1** Controllable generation (inpainting) in VESDE

---

**Require:**  $s_{\theta^*}$  (score model),  $N$  (step of process),  $M$  (binary matrix),  $y$  (real data)

$x_N \sim \mathcal{N}(0, \sigma_{max}^2 I)$

**for**  $i = N - 1$  **to** 0 **do**

$x_i \leftarrow \text{Predictor}(s_{\theta^*}, x_{i+1}, i)$

$z \sim \mathcal{N}(0, I)$

$x_i \leftarrow x_i \odot (1 - \Omega) + (y + \sigma_i z) \odot \Omega$

$x_i \leftarrow \text{Corrector}(s_{\theta^*}, x_{i+1}, i, M)$

$z \sim \mathcal{N}(0, I)$

$x_i \leftarrow x_i \odot (1 - \Omega) + (y + \sigma_i z) \odot \Omega$

**end for**

**return**  $x_0$

---

Controllable generation, such as inpainting, colorization, and super-resolution, is also possible with the sampling method. First, the inpainting method sets a binary mask to indicate the area that needs to be inpainted in an image. During sampling, the condition is added by using this mask when generating the image. Specifically, the  $(t+1)$ -th state is

generated using a sampling method from the  $t$ -th state. Then, the original data is perturbed according to the  $t$ -th state, and inpainting is performed using the generated  $t$ -th state, the perturbed data, and the mask. **Algorithm 1** shows the inpainting algorithm.

Next, Image dimensions can be decoupled to map the grayscale image into a separate channel of a different space by an orthogonal linear transformation. In addition, imputation can be performed to complete the other channels before transforming everything back to the original image space.

First, an orthogonal matrix  $M$  was defined to decouple the color channel. The orthogonal matrix  $M$  is given by

$$M = \begin{pmatrix} 0.577 & -0.816 & 0 \\ 0.577 & 0.408 & 0.707 \\ 0.577 & 0.408 & -0.707 \end{pmatrix}$$

And colorization can be conducted by applying coupling and decoupling after the predictor and corrector.

$$\text{couple}(I, M) = \text{einsum}(bc_1hw, c_1c_2 \rightarrow bc_2hw, I, M^{-1})$$

$$\text{decouple}(I, M) = \text{einsum}(bc_1hw, c_1c_2 \rightarrow bc_2hw, I, M)$$

Finally, **Algorithm 2** shows the colorization algorithm.

---

**Algorithm 2** Controllable generation (colorization) in VESDE

---

**Require:**  $s_{\theta^*}$  (score model),  $N$  (step of process),  $M$  (orthogonal matrix),  $\Omega$  (colorization mask),  $g$  (grayscale image)

$x_N \sim \mathcal{N}(0, \sigma_{max}^2 I)$

**for**  $i = N - 1$  **to**  $0$  **do**

$x_i \leftarrow \text{Predictor}(s_{\theta^*}, x_{i+1}, i)$

$z \sim \mathcal{N}(0, I)$

$x_i \leftarrow \text{couple}(\text{decouple}(x_i, M) \odot (1 - \Omega) + (\text{decouple}(g, M) + \sigma_i z) \odot \Omega, M)$

$x_i \leftarrow \text{Corrector}(s_{\theta^*}, x_{i+1}, i, M)$

$z \sim \mathcal{N}(0, I)$

$x_i \leftarrow \text{couple}(\text{decouple}(x_i, M) \odot (1 - \Omega) + (\text{decouple}(g, M) + \sigma_i z) \odot \Omega, M)$

**end for**

**return**  $x_0$

---

In addition, controllable generation can be performed using independent diffusion models and classification models such as classifier-guidance [32], or label-based generation can be performed in classifier-free diffusion model [60]. Our objective is to sample from the conditional distribution  $p_{\theta,\phi}(x_t|x_{t+1},y)$  given the condition where label  $y$  is provided. Therefore, the conditional distribution can be expressed as:

$$p_{\theta,\phi}(x_t|x_{t+1},y) = Zp_{\theta}(x_t|x_{t+1})p_{\phi}(y|x_t),$$

where  $Z$  is a normalizing constant (proof in [32]). Next, the score function in condition distribution can be obtained.

$$\begin{aligned}\nabla_{x_t} \log p_{\theta,\phi}(x_t|x_{t+1},y) &= \nabla_{x_t} \log p_{\theta}(x_t|x_{t+1}) + \nabla_{x_t} \log p_{\phi}(y|x_t) \\ &= -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon(x_t) + \nabla_{x_t} \log p_{\phi}(y|x_t)\end{aligned}$$

Applying this, we can estimate a new epsilon that corresponds to the conditional distribution.

$$\hat{\epsilon}(x_t) = \epsilon(x_t) - w\sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log p_{\phi}(y|x_t),$$

where  $w$  is scale of the classifier guidance. By modifying from  $\epsilon$  to  $\hat{\epsilon}$  in DDPM or DDIM sampling methods, we can perform conditional sampling, which is called as *classifier-guidance diffusion model* or *ablated diffusion model (ADM)* [32].

However, ADM has the burden of training two models and the inconvenience of computing gradients. To address this, instead of training a classifier, we additionally train a conditional diffusion model  $\epsilon(x_t, y)$  and an unconditional diffusion model  $\epsilon(x_t) = \epsilon(x_t, \emptyset)$ , where  $\emptyset$  is a null token for the uncondition distribution. Then, the score function of  $p(y|x_t)$  is following:

$$\nabla_{x_t} \log p(y|x_t) = -\frac{1}{\sigma_t}[\epsilon(x_t, y) - \epsilon(x_t)]$$

Then, we defined a modified epsilon  $\tilde{\epsilon}(x_t, y) = \epsilon(x_t, y) + w[\epsilon(x_t, y) - \epsilon(x_t)]$ . The modified epsilon can also be used to replace epsilon in the DDPM or DDIM sampling method, and we call this method *classifier-free diffusion model* [60].



### 2.3.6. Latent Diffusion Models (LDMs)

The diffusion model is known for its high generative ability, as it can approximate data distribution very well [18, 61, 62]. However, it costs significant computational resources during the generation process. To address this issue, latent diffusion models (LDMs) consist of two training phases [26]. In the first stage, an autoencoder comprising an encoder  $\mathcal{E}$  and a decoder  $\mathcal{D}$  are trained. The encoder learns the encoding of data into a latent space, while the decoder maps from the latent space to the data distribution. In the second stage, a diffusion model is trained to generate the encoded latent space from the prior distribution. The latent space is expanded using KL- [2, 3] or VQ- [1] regularization and adversarial training methods [11] to enhance generative ability. Finally, the following objective is given as:

$$L = \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_{\theta}(z_t, t)\|^2]$$

Due to the advancements in LDM, image generation can be performed quickly, and conditional generation tasks can be carried out using various prompts, such as CLIP [63] or BERT [64].

### ***3. High-bit depth generation in computed tomography***

Computed tomography (CT) is the most used imaging modality for the initial diagnosis of acute stroke and is adopted worldwide [65]. In emergency situations such as trauma or acute severe headache, CT is more important than other clinical imaging modalities. In addition, as the brain is an elaborate and complex-functioning organ, the three-dimensional (3D) integrity of brain imaging is important for the diagnosis and treatment of brain injuries.

The 12-bit or 16-bit format is preferred for medical images such as X-ray images, CT scans and magnetic resonance imaging (MRI), as it provides more information than the 8-bit format. CT uses a specific quantitative measurement called the *Hounsfield unit* (HU), which ranges from -1024 HU to 3071 HU in 12-bit format. In deep-learning research, CT images are generally clipped within a dynamic range 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 [66, 67].

However, generating satisfactory high-resolution images [10] in the 12-bit format of real clinical settings is difficult. Therefore, we want to experiment with how well a generative model can create 12-bit format images, as opposed to natural imaging. We first attempted generation on 12-bit format using GANs (StyleGAN2 [14, 17] and StyleGAN3 [16]) and a diffusion model (ScoreSDE). Additionally, we will evaluate quantitatively whether 12-bit generation works well not only in the whole range but also in the windowing range.

#### ***3.1. Dataset and model architecture***

A total of 34,085 non-contrast brain CT scans and paired radiology reports were retrospectively collected from patients who visited an urban, tertiary, academic hospital between January 1, 2000, and August 31, 2018. Among the scans, we only selected CT scans, which consist of 32 slices, each of 5-mm thick. The study protocol was approved by

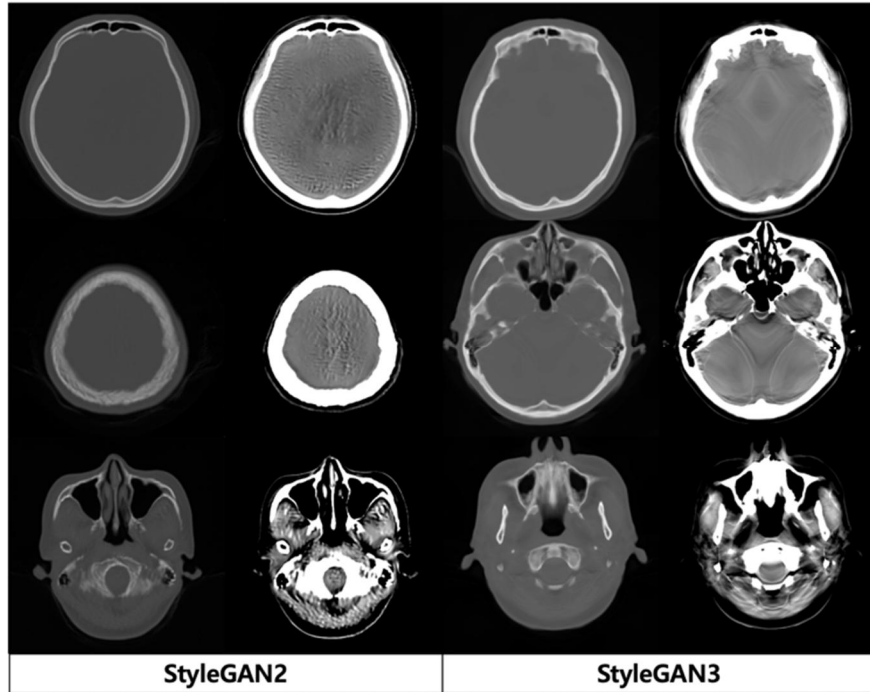
institutional review board of Asan Medical Center and Gangneung Asan Hospital. Finally, we split 1,500 brain CT scans at random to evaluate metrics.

### **3.2. Results of GAN**

It discusses 12-bit whole range generation with GANs in this Section. 12-bit format generation by two GAN models, StyleGAN2 and StyleGAN3, was evaluated in experiments. The models were good performance in the natural image, both models obtained good results over the whole range of HUs but also generated different artifacts in the windowing range like **Figure 2**. The anatomical structures, such as white matter and grey matter, collapsed in the 12-bit generation in two GAN models. On the other hand, bones and air were properly generated and keep their shape. The region that is not notably distinct from the surrounding area is not well generated, whereas other areas are well generated.

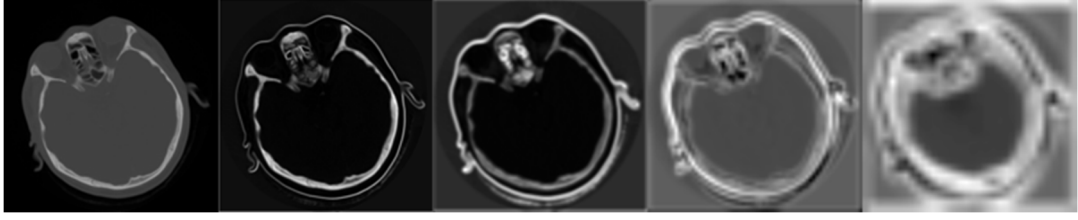
Hounsfield unit of bone (1000HU~) is larger and more deviated in comparison to other tissues (20HU ~ 30HU). The Hounsfield unit of air, on the other hand, is smaller and more deviated when compared to other tissues. As a result, these two tissues produce a high signal region in comparison to other tissues. Because the signal is strong due to a large deviation, most studies

use windowing as a preprocessing to attenuate signal except ROIs. [68, 69]



**Figure 2.** Result of StyleGAN2 and StyleGAN3. Both GAN models generate images well in the whole range, but different artifacts are observed when clipping in the windowing range.

The convolution operation was described as a high-pass filter [70], that amplifies the high-frequency components and is susceptible to high-frequency noises. Therefore, the GAN was not used for generating the 12-bit format because the convolution operation may not train the features from low-frequency signals. Also, GAN is trained by the feature of image through discriminator. As shown in **Figure 3**, the features from the low signal regions seem somewhat neglected. Therefore, generating with GAN is difficult because it uses adversarial loss from discriminator features.

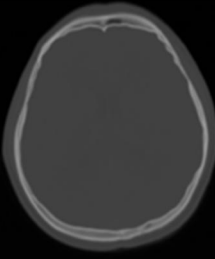
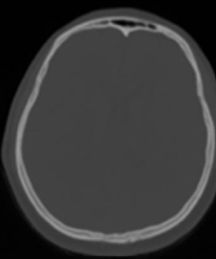
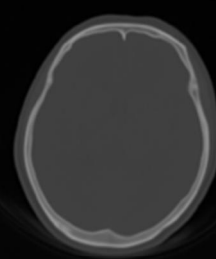
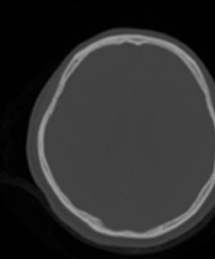
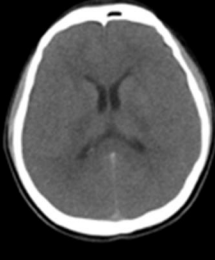
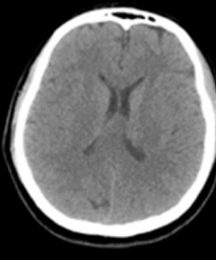
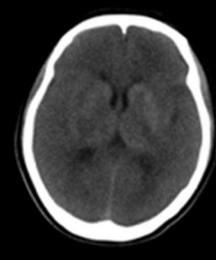
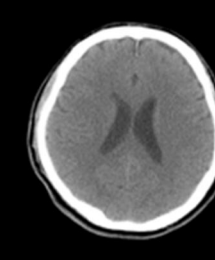


**Figure 3.** This is the feature map extracted from the discriminator using the CT images in the whole range. Low-signal regions, such as parenchyma, are not generated properly.

### 3.3. Results of score-based diffusion model

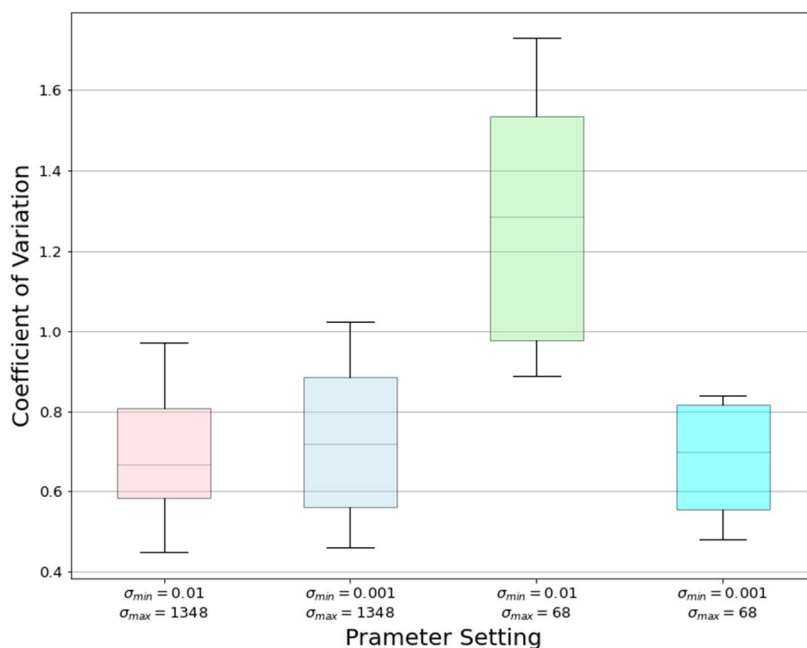
The score-based diffusion model trains the *score*, which defines the gradient of the log-density [23], rather than the image features. Therefore, the score-based diffusion model does not depend on image signals, despite being implemented as a convolution operation. In the noise schedule of [19, 56],  $\sigma_{min}$  was fixed at 0.01. This previous research performed noise scheduling on 8-bit natural images, not in 12-bit format. The models sufficiently generate the 8-bit format because the diffusion coefficient of the perturbation process cannot offset the low-frequency signals in an 8-bit format image. However, in 12-bit format, the diffusion coefficient can offset the low-frequency signals so the models cannot generate 12-bit formatted images.

To experimentally demonstrate this, we experimented two  $\sigma_{min}$  and  $\sigma_{max}$ ; 1) 0.01 and 1,348, 2) 0.001 and 1,348, 3) 0.01 and 68, and 4) 0.001 and 68, respectively. Setting  $\sigma_{min}$  to 0.01 generated a noisy image as shown in the first and third columns of **Figure 4**. Particularly, perturbed noise makes it impossible to distinguish between white matter and gray matter. Otherwise, when  $\sigma_{min}$  was reduced to 0.001, the generated image was well clarified in **Figure 4**. We also evaluated the coefficient of variation (CV) in the inner brain of the generated images (omitting the bones and air from the windowing) at both noise levels [71]. As shown in **Figure 4**, when the CV was low, anatomical structures are generated so that they may be distinguished. It was interpreted that reducing  $\sigma_{min}$  improved the quality of the generated image.

$\sigma_{min} = 0.01$ $\sigma_{max} = 1348$	$\sigma_{min} = 0.001$ $\sigma_{max} = 1348$	$\sigma_{min} = 0.01$ $\sigma_{max} = 68$	$\sigma_{min} = 0.001$ $\sigma_{max} = 68$
			
			
CV : 0.8485	CV : 0.6648	CV : 1.4128	CV : 0.5482

**Figure 4.** Results according to  $\sigma_{min}$  and  $\sigma_{max}$ . CV: coefficient of variation.

To quantitatively demonstrate the settings, we randomly generated 1,000 slices by each parameter and measured the CV. As shown **Figure 5**, the variance of CV was lowest when  $\sigma_{min}$  and  $\sigma_{max}$  were 0.001 and 68, respectively. Compared to other parameters, the cleaner the images were generated the lower the noise variations were measured. Furthermore, a board-certified radiologist also qualitatively assessed that the images generated with a lower  $\sigma_{min}$  showed a cleaner image. Also, setting  $\sigma_{max}$  to 68 is theoretically plausible according to previous study [21] because we preprocessed CT slices in the range of -1 to 1.



**Figure 5.** Results of coefficient of variation according to  $\sigma_{min}$  and  $\sigma_{max}$ .

Unlike GAN, the score-based diffusion model can generate 12-bits format images. The clarity of the image is determined by the diffusion coefficient in the final stage, which generates the fine-grained regions. If the diffusion coefficient is approximately 0.01, there is noise smaller than 2 values if this value is in the 8-bit range (0 to 255). However, there will be noise within 10 values if the value is in the 12-bit range (-1024, 3071HU). Therefore, since  $\sigma_{min}$  has the greatest impact on the diffusion coefficient, it was crucial to lower its value.

#### ***4. Data augmentation with isocitrate dehydrogenase type in glioma***

The World Health Organization (WHO) recently advanced the role of molecular diagnosis in central nervous system tumor classification, and molecular results have now been added as biomarkers for tumor grading [72]. For example, isocitrate dehydrogenase (IDH) mutant astrocytomas are graded as WHO grade 2, 3, or 4, whereas all IDH-wild types are grade 4 [73]. The grade is therefore no longer restricted to being a histological grade, and the importance of molecular biomarkers such as IDH mutation is emphasized. Therefore, imaging phenotypes representative of IDH mutation need to be separately learned from grading phenotypes such as the degree of contrast-enhancement or tumor size; high-grade glioma tends to show larger tumor size and more contrast enhancement than lower-grade glioma [74, 75].

The natural prevalence of IDH mutation differs between grades, with all IDH-wild types being glioblastomas (WHO grade 4), whereas most IDH-mutant types are lower-grade gliomas (WHO grade 2 or 3). Several previous studies developed deep learning-based classification algorithms to predict IDH mutation status using the natural prevalence reflected in the study population [76, 77]. However, only a limited number of patients with IDH-mutant high-grade glioma were included in these studies, which raises a concern that deep learning algorithms may learn imaging phenotypes to distinguish lower- and higher-grade gliomas, rather than molecular subtypes. Since most neural networks use black-box type analytic engines that generate unexplainable feature vectors with limited insight into the underlying mechanism for image classification, these potential risks are worthy of further consideration [78].

Image synthesis and augmentation with a generative model is a potential solution for medical deep learning models dealing with small or imbalanced clinical samples [79]. However, whether generated images can simulate various imaging phenotypes to improve classification performance remains unclear. Moreover, the optimal proportion of generated images to add to a classification model and the effects of adding specific imaging phenotypes



have not been studied. We therefore developed an imaging phenotype-based generative model augmentation (GMA) method and tested whether it could improve the classification of IDH type in glioma.

#### **4.1. Materials and Methods**

##### **4.1.1. Dataset**

This study is reported in accordance with the Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015 guidelines [80]. The study protocol was approved by the institutional review board of our institution, a tertiary referral hospital, which waived the requirement for informed consent as patient images were collected retrospectively and identification information was removed to achieve compliance with Health Insurance Portability and Accountability Act (HIPAA). The inclusion process for the study patients is shown in **Figure 6**. A total of 839 patients (703 from dataset 1 and 136 from the Cancer Genome Atlas [TCGA] and Cancer Imaging Archive [TCIA]) who underwent preoperative MRI for newly diagnosed glioma (grades 2, 3, and 4) between August 2008 and September 2020 were considered for inclusion. The inclusion criteria were as follows: (i) pathologically confirmed glioma, (ii) known IDH mutation status according to WHO 2016 and WHO 2021 criteria [73, 81], (iii) preoperative MRI including contrast-enhanced T1-weighted imaging (CE-T1WI) and T2-weighted fluid-attenuated inversion recovery (FLAIR) imaging, and (iv) age  $\geq 18$  years. The exclusion criteria were as follows: (i) previous history of biopsy or surgery for brain tumor ( $n = 14$ ), (ii) absence of CE-T1WI or T2 FLAIR images ( $n = 15$ ), (iii) inadequate image quality ( $n = 12$ ), and/or (vi) unknown IDH status ( $n = 28$ ). Finally, 664 patients from dataset 1 and 106 patients from TCGA were enrolled. These patients were allocated into development with stratified random sampling ( $n = 651$ , both dataset 1 and TCGA patients) and internal test sets ( $n = 119$ , dataset 1 patients). The development set was subsequently divided into training ( $n = 565$ ) and tuning ( $n = 86$ ) sets. The patients in the development set were taken from two

different datasets to account for diverse imaging sequences from multiple centers. For an external test set, 108 patients imaged between July 2017 and October 2018 at the dataset 2 were enrolled, in accordance with the previously stated inclusion and exclusion criteria.

**Table 1.** Patient characteristics and molecular subtypes

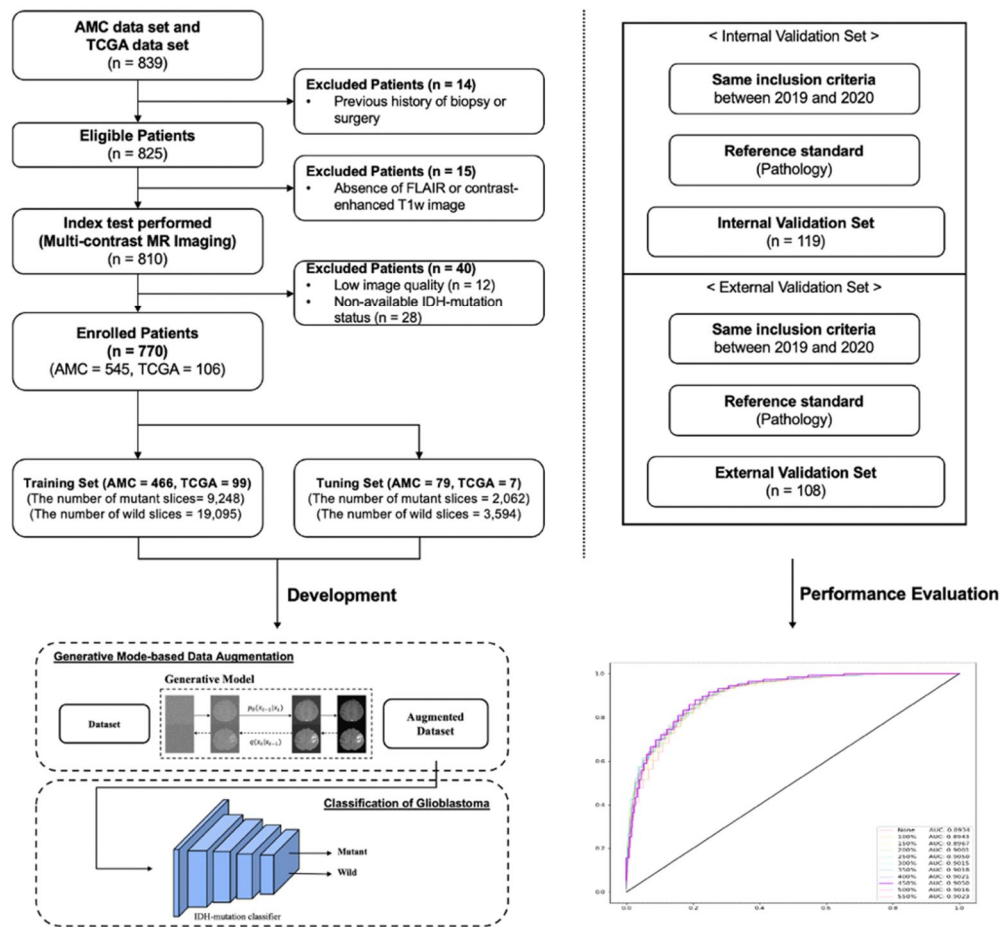
	Training set	Internal	<i>P</i>	External	<i>P*</i>
Number of patients	565	119		108	
Dataset	AMC + TCGA	AMC		Severance	
Age (years)	53.5 ± 14.6	56.4 ± 13.9	.06	56.9 ± 15.6	.05
Sex			.95		.70
Male	320 (56.6%)	67 (56.3%)		59 (54.6%)	
Female	245 (43.4%)	52 (43.7%)		49 (45.4%)	
IDH status			.38		.29
Wild type	346 (61.2%)	78 (65.5%)		72 (66.7%)	
Mutant type	219 (38.8%)	41 (34.5%)		36 (33.3%)	
WHO grade			<b>.02</b>		.18
2	98 (17.3%)	31 (26%)		11 (10.2%)	
IDH wild type	26	9		2	
IDH mutant + 1p/19q non-codeletion	72	20		9	
IDH mutant + 1p/19q codeletion	0	2		0	
3	97 (17.2%)	12 (10.1%)		20 (18.5%)	
IDH wild	48	6		10	
IDH mutant + 1p/19q non-codeletion	30	4		2	
IDH mutant + 1p/19q codeletion	19	2		8	
IV	370 (65.5%)	76 (63.9%)		77 (71.3%)	
IDH wild	307	63		60	
IDH mutant + 1p/19q non-codeletion	8	1		0	
IDH mutant + 1p/19q codeletion	2	1		0	
IDH mutant + 1p/19q status not specified	53	11		17	

Note: *P* indicates statistical significance between training and internal validation sets. *P\** indicates statistical significance between training and external validation sets. Data are expressed as mean ± standard deviation. Abbreviation: IDH = isocitrate dehydrogenase.

#### ***4.1.2. IDH mutation status and image preprocessing***

IDH mutation status was analyzed by members of the pathology division of our hospital, who were blinded to the radiologic results. Before 2017, the reference standard consisted of immunohistochemical determination of IDH1 (R132H) protein expression [82]. Mutations in *IDH1* and *IDH2* genes were determined by DNA pyrosequencing at diagnosis. From 2017, next generation sequencing was performed as routine practice, and IDH gene mutation status was diagnosed.

All enrolled patients from our institution underwent MRI on a 3.0 Tesla scanner (Achieva or Ingenia, Philips Medical Systems) using a 16 channel or 32 channel head coils. The MRI protocols included T2-weighted imaging (T2WI), FLAIR imaging, T1-weighted imaging (T1WI) , and CE-T1WI. The CE-T1WI images were obtained as a high-resolution three-dimensional (3D) volume, using a gradient-echo T1WI with the following parameters: repetition time (TR)/echo time (TE), 9.8/4.6 ms; flip angle, 10°; field of view (FOV), 256 x 256 mm; matrix, 512 × 512; and slice thickness, 1 mm with no gap. The parameters for FLAIR imaging included TR/TE, 9000/135 ms; flip angle, 90°; FOV, 240 x 240mm; matrix, 512 × 512; and slice thickness, 4 mm with no gap. To prepare the training data, the preprocessing methods were used co-registration between the 3D contrast-enhanced T1-weighted and FLAIR images using rigid transformations with six degrees of freedom in SPM12 [83], skull stripping using HD-BET algorithms [84] to remove non-brain tissues, intensity clipping to convert the original 16-bit MR images to 8-bit by clipping the lower 1% of intensities, slice selection by including only slices with tumors larger than 100 pixels, data cleansing by identifying and excluding incorrect and inconsistent data types, and anonymization by removing all identifying information from the images and using de-identified data for analysis.



**Figure 6.** Flow diagram of the training, model development, and internal and external testing. AMC = Asan Medical Center; TCGA = The Cancer Genome Atlas.

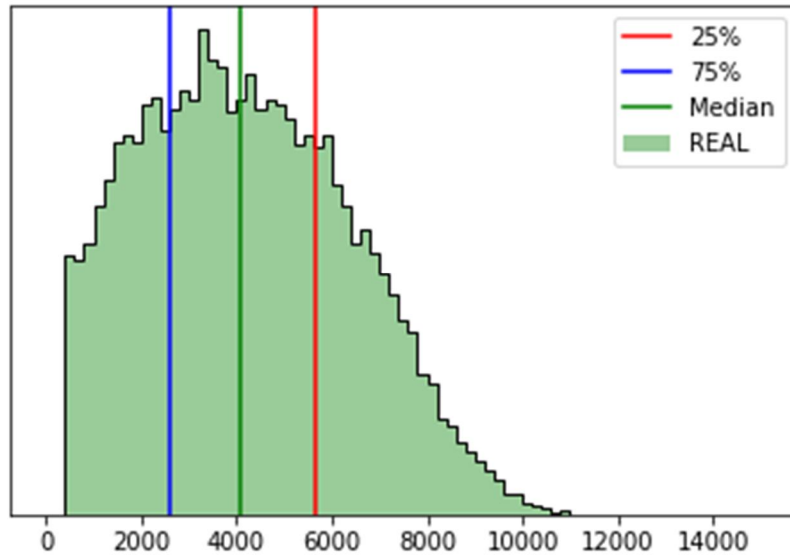
For the quantitative evaluation of generative model, Fréchet inception distance (FID) scores of IDH-mutant and IDH-wild type were 10.54 and 8.33, respectively.

#### 4.1.3. Imaging phenotype

The primary endpoints of our study were as follows: (1) optimization of the number of training images using a deep generative model for IDH mutation prediction with GMA, and (2) exploration of imaging phenotype-based GMA and determination of the optimum imaging phenotype-based GMA to improve IDH mutation prediction via a deep generative model.

The imaging phenotype-based GMA was developed using the Visually AcceSAbly Rembrandt Images (VASARI) lexicon, which is a rule-based lexicon to improve the reproducibility of interpretation of gliomas and attempts to standardize visual interpretation of malignant gliomas [85]. Among the qualitative features, we chose the semiquantitative contrast-enhancement pattern for imaging phenotype-based GMA: the proportion of contrast-enhancement was measured, and more than 30% was defined as predominant enhancement, whereas less than 5% was defined as no enhancement. The segmentation output included enhancing tumor, nonenhancing tumor, and internal necrosis (if present). The proportion of contrast enhancement (CE) was calculated as follows:  $(\text{enhancing tumor} / [\text{enhancing tumor} + \text{any necrosis} + \text{nonenhancing tumor}] \times 100)$ . This parameter was used as the imaging phenotype of CE.

Next, the tumor size was defined as large or small according to whether it was above the 75<sup>th</sup> percentile or below the 25<sup>th</sup> percentile of the real data. The median and range of the whole tumor area (enhancing tumor + necrosis + nonenhancing tumor) were 4059 voxels (40.59 mm<sup>2</sup>). On the basis of the real data, we defined regions containing more than 5655 voxels (56.55 mm<sup>2</sup>) as large (>75<sup>th</sup> percentile) and those with less than 2589 (25.89 mm<sup>2</sup>) as small (<25<sup>th</sup> percentile) in **Figure 7**.



**Figure 7.** Histogram of tumor size in the real data. A large tumor size was defined as that above the 75<sup>th</sup> percentile of tumor size and a small tumor as below the 25<sup>th</sup> percentile.

## 4.2. Results

### 4.2.1. Evaluation by human readers

To assess the authenticity of generated images, 200 pairs of CE-T1WI and FLAIR images were randomly selected from both the generated and real datasets for evaluation. Two neuroradiologists with different levels of experience in neuro-oncologic imaging were asked to independently determine whether each pair of images was real or fake through a visual Turing test. The accuracy of their classifications was then calculated as the percentage of correct identifications between real and fake images.

To validate the accuracy of the imaging-based phenotypes, 200 pairs of contrast-enhanced T1WI and FLAIR images were randomly selected from both the IDH wild-type and IDH-mutant generation networks. However, 13 pairs from the IDH wild-type network and 36 pairs from the IDH-mutant network were excluded because they did not show the maximum tumor

portion. As a result, a total of 151 paired images were evaluated based on their imaging-based phenotype.

Two neuroradiologists, H.H.M and J.E.P, who respectively had 2 and 8 years of experience in neuro-oncologic imaging, evaluated the generated images without knowledge of their corresponding generative network. They assessed the images using VASARI features such as tumor location, proportion of enhancement, multifocal or multicentric features, and cortical involvement, as well as other reproducible qualitative features like the presence of necrosis and the margin of non-enhancing lesions [86]. The tumor location was specified according to the tumor epicenter.

In generated images, the tumor location, proportion of enhancement, presence of cortical involvement, and presence of necrosis were significantly different between IDH-wild and IDH-mutant types according to chi-square tests. The IDH wild-type group had higher rates of nonlobar location (4.5% [4/87] in IDH-wild type and 0% [0/64] in mutant-type,  $P=.027$ ) and necrosis (70.1% [60/87] in IDH-wild type and 25% [16/64] in mutant-type,  $P < .001$ ), and had a higher proportion of enhancement than the IDH-mutant type group ( $P = .025$ ). Of 87 cases in the IDH wild-type group, 18 (20.7%) were classified as 68–100% enhancement, 36 (41.4%) as 34–67%, 12 (13.8%) as 6–33%, and 21 (24.1%) as <5%. Of the 64 cases in the IDH-mutant type group, 3 (4.7%) were classified as 68–100% enhancement, 17 (26.6%) as 34–67%, 5 (7.8%) as 6–33%, and 39 (60.9%) as <5%. The IDH-mutant type group had more cases with a frontal location (40.2% [35/87] in IDH-wild type and 59.4% [38/64] in mutant-type,  $P = .027$ ) and cortical involvement (60.9% [53/87] in IDH-wild type and 78.1% [50/64] in mutant-type,  $P = .025$ ). However, a multifocal or multicentric distribution and the margin of nonenhancing lesion showed no significant difference between the two subtypes ( $P = .871$  and  $.093$ , respectively).

In real images, the tumor location, proportion of enhancement, presence of cortical

involvement, and presence of necrosis, and margin of nonenhancing lesion were significantly different between IDH-wild and IDH-mutant types. The IDH wild-type group had higher rates of nonlobar location (% [8/82] in IDH-wild type and % [2/89] in mutant-type,  $P < .001$ ) and necrosis (56.1% [46/82] in IDH-wild type and 23.6% [21/89] in mutant-type,  $P < .001$ ), and had a higher proportion of enhancement than the IDH-mutant type group ( $P < .001$ ). The IDH-mutant type group had more cases with a frontal location (29.3% [24/82] in IDH-wild type and 66.3% [59/89] in mutant-type,  $P < .001$ ) and cortical involvement (59.8% [49/82] in IDH-wild type and 88.8% [79/89] in mutant-type,  $P < .001$ ). However, a multifocal or multicentric showed no significant difference between the two subtypes ( $P = .425$ ).

#### 4.2.2. Deep learning-based prediction of IDH type using the real data and nonselective GMA

The study evaluated the diagnostic performance of the model for classifying IDH mutation status in an internal test set. In **Table 2**, the performance was assessed based on different levels of data augmentation. The null model without augmentation had an accuracy of 81.5% and an AUC of 0.900. The best performance was achieved with 110,000 generated images added to the original images, resulting in an AUC of 0.938 and an accuracy of 85.7%. The model was also evaluated on an external test set, where it achieved an AUC of 0.833 and an accuracy of 75.0% with the optimal level of data augmentation. Sensitivity, specificity, PPV, and F1 score were also reported for each case.

**Table 2.** Diagnostic performance for the classification of IDH-mutation status with addition of generated images without imaging feature selection in internal test set (per patient)

	AUC (95% CI)	Accuracy	Sensitivity	Specificity	PPV	F1 score
0	0.900 (0.833-0.966)	81.5% (97/119)	56.1% (23/41)	94.9% (74/78)	0.852 (23/27)	0.677
1U	0.922 (0.862-0.981)	78.2% (93/119)	90.2% (37/41)	71.8% (56/78)	0.627 (37/59)	0.740



2U	0.895 (0.826-0.963)	82.4% (98/119)	53.7% (22/41)	97.4% (76/78)	0.917 (22/24)	0.677
3U	0.913 (0.850-0.975)	82.4% (098/119)	56.1% (23/41)	96.2% (75/78)	0.885 (23/26)	0.687
4U	0.906 (0.842-0.971)	84.9% (101/119)	80.5% (33/41)	87.2% (68/78)	0.767 (33/43)	0.786
5U	0.907 (0.842-0.971)	80.7% (96/119)	87.8% (36/41)	76.9% (60/78)	0.667 (36/54)	0.758
6U	0.929 (0.873-0.986)	86.6% (103/119)	73.2% (30/41)	93.6% (73/78)	0.857 (30/35)	0.790
7U	0.928 (0.871-0.985)	84.9% (101/119)	73.2% (30/41)	91.0% (71/78)	0.811 (30/37)	0.769
8U	0.881 (0.809-0.953)	66.4% (79/119)	92.7% (38/41)	52.6% (41/78)	0.507 (38/75)	0.655
9U	0.900 (0.834-0.967)	82.4% (98/119)	58.5% (24/41)	94.9% (74/78)	0.857 (24/28)	0.696
10U	0.914 (0.852-0.976)	83.2% (99/119)	63.4% (26/41)	93.6% (73/78)	0.839 (26/31)	0.722
11U (*)	0.938 (0.885-0.991)	85.7% (102/119)	75.6% (31/41)	91.0% (71/78)	0.816 (31/38)	0.785
12U	0.882 (0.811-0.954)	83.2% (99/119)	68.3% (28/41)	85.9% (67/78)	0.800 (28/35)	0.737
13U	0.942 (0.890-0.993)	84.0% (100/119)	75.6% (31/41)	94.9% (74/78)	0.775 (31/40)	0.765
14U	0.918 (0.857-0.979)	82.4% (98/119)	87.8% (36/41)	84.6% (66/78)	0.692 (36/52)	0.774
15U	0.905 (0.839-0.970)	84.0% (100/119)	75.6% (31/41)	100% (78/78)	0.775 (31/40)	0.765

Note: 1U (unit) refers to a set of 10,000 synthetic images, consisting of 5,000 images of IDH mutant and 5,000 images of IDH wild types. AUC: area under the receiver operating characteristics curve; PPV: positive predictive value. (\*): optimal augmentation.

#### ***4.2.3. Deep learning-based prediction of IDH type using imaging phenotype-based GMA according to tumor size***

The effects of tumor size-based GMA on the internal and external test sets are summarized according to tumor size in **Table 3**. Compared with GMA, imaging phenotype-based GMA with a large tumor size showed similar results in both internal and external test sets (internal test set: AUC 0.956, 95% CI: 0.911–1, accuracy 86.6% [103/119]; external test set: AUC 0.810,

95% CI: 0.716–0.903, accuracy: 77.8% [84/108]). Adding a large tumor size improved classification model accuracy and specificity in both internal and external test sets (internal test set: accuracy 86.6% [103/119], specificity 100% [78/78]; external test set: accuracy 77.8% [84/108], specificity 87.5% [63/72]), but it did not reach statistical significance. The specificity of the classification model represented the ability to predict the IDH-wild subtype in true IDH wild-type patients.

Compared with optimal GMA, imaging phenotype-based GMA with a small tumor size reduced accuracy and sensitivity of the classification model in both internal and external test sets (internal test set: accuracy 70.6% [84/119], sensitivity 14.6% [6/41],  $P = <.00$ , both; external test set: accuracy 69.4% [75/108], sensitivity 13.9% [5/36],  $P = <.00$ , both) while increasing specificity (internal test set: 94.9% [74/78]; external test set: 97.2% [70/72],  $P = .02$  and  $<.00$ , respectively) with statistical significance. The specificity of the classification model represented the ability to predict the IDH-wild subtype in true IDH wild-type patients.

**Table 3.** Effect of image features in the generated images according to size on the classification of IDH-mutation status

Parameter	Internal				
Size	Optimal	Small	Large	<i>P</i>	<i>P*</i>
F1 score	0.785	0.255	0.784		
AUC	0.938 (0.885-0.991)	0.896 (0.828-0.967)	0.956 (0.911-1.000)	.09	.22
Accuracy	85.7% (102/119)	70.6% (84/119)	86.6% (103/119)	<b>&lt;.001</b>	.23
Sensitivity	75.6% (31/41)	14.6% (6/41)	70.7% (29/41)	<b>&lt;.001</b>	.69
Specificity	91.0% (71/78)	94.9% (74/78)	100% (78/78)	<b>.02</b>	.38
Parameter	External				
Size	Optimal	Small	Large	<i>P</i>	<i>P*</i>

F1 score	0.64	0.233	0.636		
AUC	0.833 (0.744-0.922)	0.847 (0.761-0.933)	0.810 (0.716-0.903)	.68	.53
Accuracy	75.0 % (81/108)	69.4% (75/108)	77.8% (84/108)	<.001	.08
Sensitivity	66.7% (24/36)	13.9% (5/36)	58.3% (21/36)	<.001	.51
Specificity	79.2% (57/72)	97.2% (70/72)	87.5% (63/72)	<.001	.15

Note: P indicates statistical significance between optimal augmentation and small augmentation using DeLong's test and McNemar's test. P\* indicates statistical significance between optimal augmentation and large size using DeLong's test and McNemar's test.

#### ***4.2.4. Deep learning-based prediction of IDH type using imaging phenotype-based GMA according to CE***

The effect of imaging phenotype-based GMA performed according to CE in the internal and external test sets is summarized in **Table 4**. Compared with optimal GMA, CE-based GMA showed similar results in internal test set with the addition of either predominant CE (AUC 0.922, 95% CI: 0.863–0.982, accuracy: 83.2% [99/119]) or no CE (AUC 0.903, 95% CI: 0.838–0.969, accuracy: 81.5% [97/119]). Imaging phenotype-based GMA with no CE reduced AUC, accuracy, and specificity of the classification model in external test sets (AUC 0.765, 95% CI: 0.664–0.867, accuracy: 68.5% [74/108], specificity: 65.3% [47/72],  $P = .04$ ,  $<.00$ , and  $.02$ , respectively) while increasing sensitivity (75.0% [27/36],  $P = .02$ ) with statistical significance. The sensitivity of the classification represented the prediction of IDH mutation type in the true IDH mutation patients.

Representative cases of fully automated IDH mutation classification with CAM are shown in **Figure 8**. The main activation area was the enhancing tumor areas if the tumor had predominant contrast enhancement, whereas the main activation area involved the entire tumor and peritumoral edema if the tumor had no enhancement.

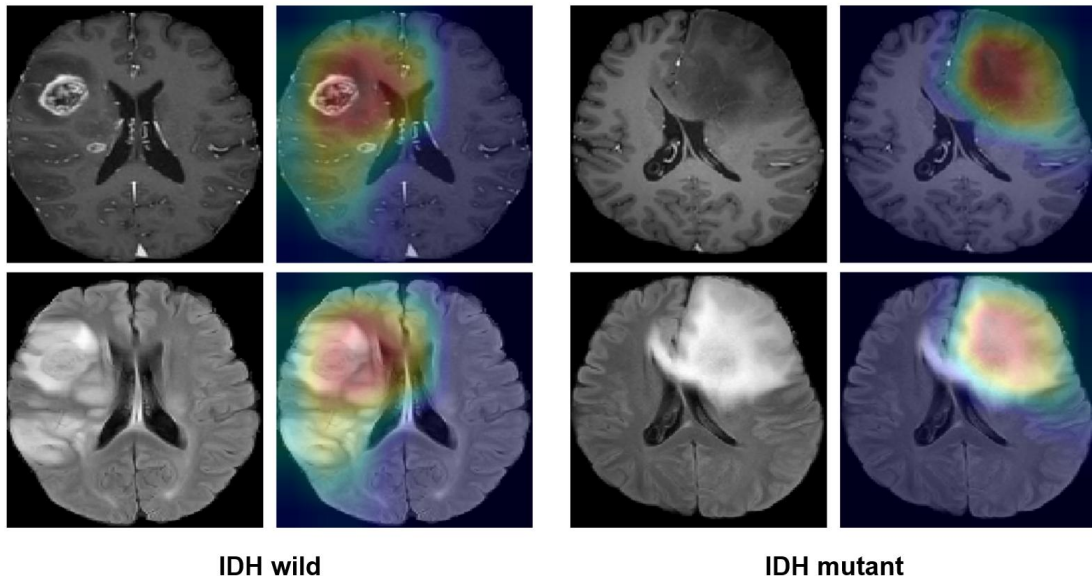
**Table 4.** Effect of image features in the generated images according to CE on the classification

of IDH-mutation status

Parameter	Internal				
Contrast enhancement	Optimal	No CE	Predominant CE	<i>P</i>	<i>P*</i>
F1 score	0.785	0.732	0.767		
AUC	0.938 (0.885-0.991)	0.903 (0.838-0.969)	0.922 (0.863-0.982)	.36	.19
Accuracy	85.7% (102/119)	81.5% (97/119)	83.2% (99/119)	.12	.66
Sensitivity	75.6% (31/41)	73.2% (30/41)	80.5% (33/41)	.73	1
Specificity	91.0% (71/78)	85.9% (67/78)	84.6% (66/78)	.13	.39
Parameter	External				
Contrast enhancement	Optimal	No CE	Predominant CE	<i>P</i>	<i>P*</i>
F1 score	0.64	0.614	0.587		
AUC	0.833 (0.744-0.922)	0.765 (0.664-0.867)	0.749 (0.646-0.853)	<b>.04</b>	.09
Accuracy	75.0% (81/108)	68.5% (74/108)	58.3% (63/108)	<b>&lt;.001</b>	<b>.01</b>
Sensitivity	66.7% (24/36)	75.0% (27/36)	88.9% (32/36)	<b>.02</b>	.45
Specificity	79.2% (57/72)	65.3% (47/72)	43.1% (31/72)	<b>&lt;.001</b>	<b>.02</b>

Note: *P* indicates statistical significance between optimal augmentation and no contrast enhancement (CE) augmentation using DeLong's test and McNemar's test. *P\** indicates statistical significance between optimal augmentation and predominant CE augmentation using DeLong's test and McNemar's test.

Representative cases of fully automated IDH mutation classification with CAM are shown in **Figure 8**. The main activation area was the enhancing tumor areas if the tumor had predominant contrast enhancement, whereas the main activation area involved the entire tumor and peritumoral edema if the tumor had no enhancement.



**Figure 8.** Representative cases of fully automated IDH-mutation classification with class activation maps (CAM) for paired contrast-enhanced T1-weighted images and FLAIR images. IDH-wild type glioma is shown on the left, and IDH-mutant type glioma is shown on the right. IDH-wild type glioma presented as a predominantly enhancing mass with central necrosis [87]. The main activation area of the CAM involved enhancing tumor. IDH-mutant type glioma presented as a nonenhancing tumor with FLAIR hyperintensity. The main activation area of the CAM involved the entire tumor and peritumoral edema.

### ***5. Image-to-image translation with H&E staining normalization of whole slide imaging***

Digital pathology has made whole slide imaging (WSI) useful in different areas such as surgical pathology, diagnostic pathology, and education [88-91]. This technology has also enabled the development of deep learning (DL) models for pathology, leading to the creation of various DL models for tasks like pathology image segmentation and brain tumor type classification [92, 93].

Hematoxylin and eosin (H&E) staining is the standard in medical pathology, but the dosage ratio is not standardized, and the stain fades at different rates, causing variations in images. Stain normalization is crucial in DL of digital pathology to address these issues. Various methods, such as conventional techniques, autoencoders [94], and GANs [95], have been developed to normalize histopathologic images. However, conventional methods like Macenko [96] and Vahadane [97] can lead to tissue structure and texture loss during stain normalization based on reference images.

To address the challenges in stain normalization of WSIs, patch-wise DL-based normalization approaches were developed due to the large image size. However, these methods can result in grid artifacts at the patch boundary caused by the 3x3 kernel size of the convolution operation, which performs a weight summation among local pixels [98]. While using a 1x1 kernel or weighted summation of neighborhood output's pixel can mitigate the issue, there are still potential risks like overfitting and contrast differences between adjacent patches.

As mentioned in **Section 2.3.5**, the application of diffusion models in inpainting and colorization has also shown significant results. Therefore, we attempted to apply this method to stain normalization. Paradoxically, the high generative ability of diffusion models can lead to color mistransfer in stain normalization. To prevent this issue, researchers have proposed a stain separation technique that decomposes the H&E stain into hematoxylin and eosin components. This approach aims to limit the capabilities of diffusion models and ensure accurate stain normalization results.

To prevent grid artifacts in the normalized WSIs, researchers used an overlapped moving window patches approach for inpainting. Additionally, they developed a patch-wise stain normalization technique using a diffusion models with stain separation and overlapped moving window patch strategies. This approach aimed to improve the accuracy of stain normalization in digital pathology images.

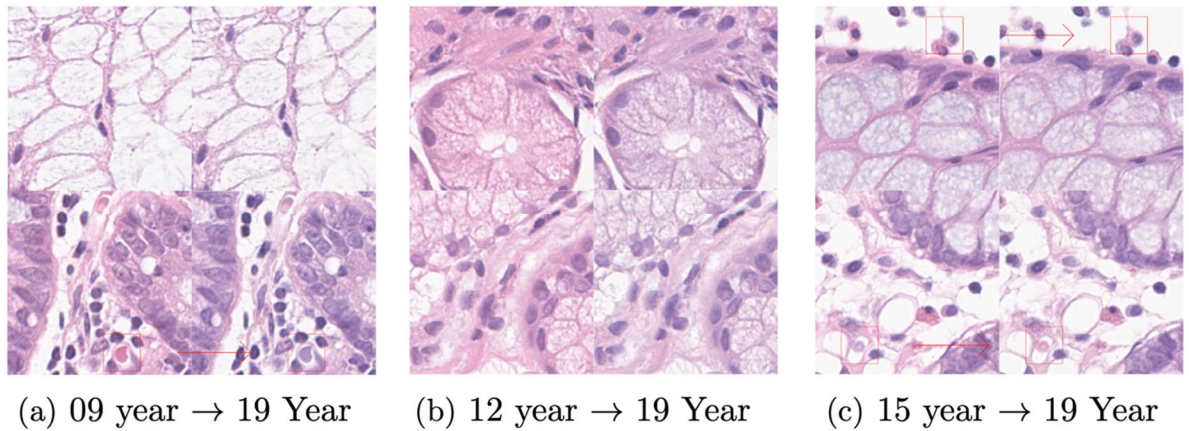
### ***5.1. Dataset***

This study used histopathologic WSIs of the colon stained with H&E from Asan Medical Center in 2009, 2012, 2015, and 2019. The WSIs from 2019 were used to train the SDM model, and 100,000 patches of 256×256 size were extracted from each WSI [99]. The CAMELYON 16 [100] and PAIP2019 [101] datasets, which consist of WSIs of hepatic lymph nodes and hepatocellular carcinoma stained with H&E, were used for external validations.

### ***5.2. Stain normalization without stain separation***

The researchers began by converting patch images from different years into grayscale images by taking the average of their RGB channels. Then, they applied **Algorithm 2 of Section 2.4** to normalize the grayscale images. As shown as **Figure 9**, although the resulting images showed consistent normalization, there was a critical issue in which red blood cells and eosinophils appeared purple in some regions of the images, instead of the expected red color. This could lead to misdiagnosis of important pathologies, such as inflammation, as normal. The researchers suspected that this problem arose from the loss of information that

occurred during the averaging process of RGB channels into grayscale images.



**Figure 9.** The results of stain normalization w/o stain separation in 09, 12 and 15 years. Images of left column is original, and right is normalized image. Red boxes show the regions where it should be red, such as red blood cells or eosinophils, turned into purple.

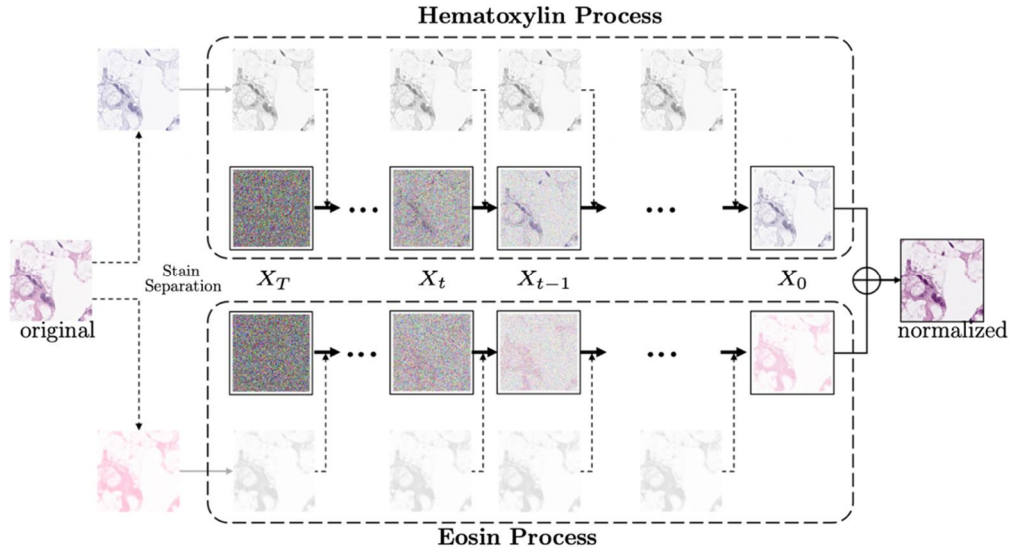
### ***5.3. Stain normalization with stain separation***

To address the challenge of high diversity in histological images, the study utilized stain separation to separate hematoxylin and eosin stains from the H&E-stained images. For this purpose, sparse non-negative matrix factorization (SNMF) [102] was employed, which uses color deconvolution in the Vahadane [97] method to provide stain information for the model. This approach was developed to overcome the limitations of non-negative matrix factorization [103], which is not suitable for large histological image datasets with considerable color variation [97].

To perform stain normalization, two separate SDM models were trained on the hematoxylin and eosin stain spaces. Prior to training, SNMF method was used to perform color deconvolution and separate the stains. After training, the SDM models were applied to the separated stain images from the source image. The normalization was performed using PC



sampling **Algorithm 2** and the resulting normalized images were obtained. The overall process is illustrated in **Figure 10**.



**Figure 10.** Flow chart of the method using hematoxylin and eosin processes. Using sparse non-negative matrix factorization (SNMF), histological images separate to hematoxylin and eosin representation.

To normalize whole slide images (WSI), a moving window approach is being used. Initially, the first patch of the hematoxylin and eosin WSI is normalized using Algorithm 1. Then, the window is shifted by the overlapped ratio of the previously normalized patch to prepare the patch for subsequent normalization. By alternating colorization and inpainting, the non-overlapping regions are also normalized. This method has been shown to produce realistic and consistent results in previous studies [104]. In contrast to other patch-wise stain normalization methods, grid artifacts have not been observed in the results of this approach.

#### 5.4. Performance evaluation criterion

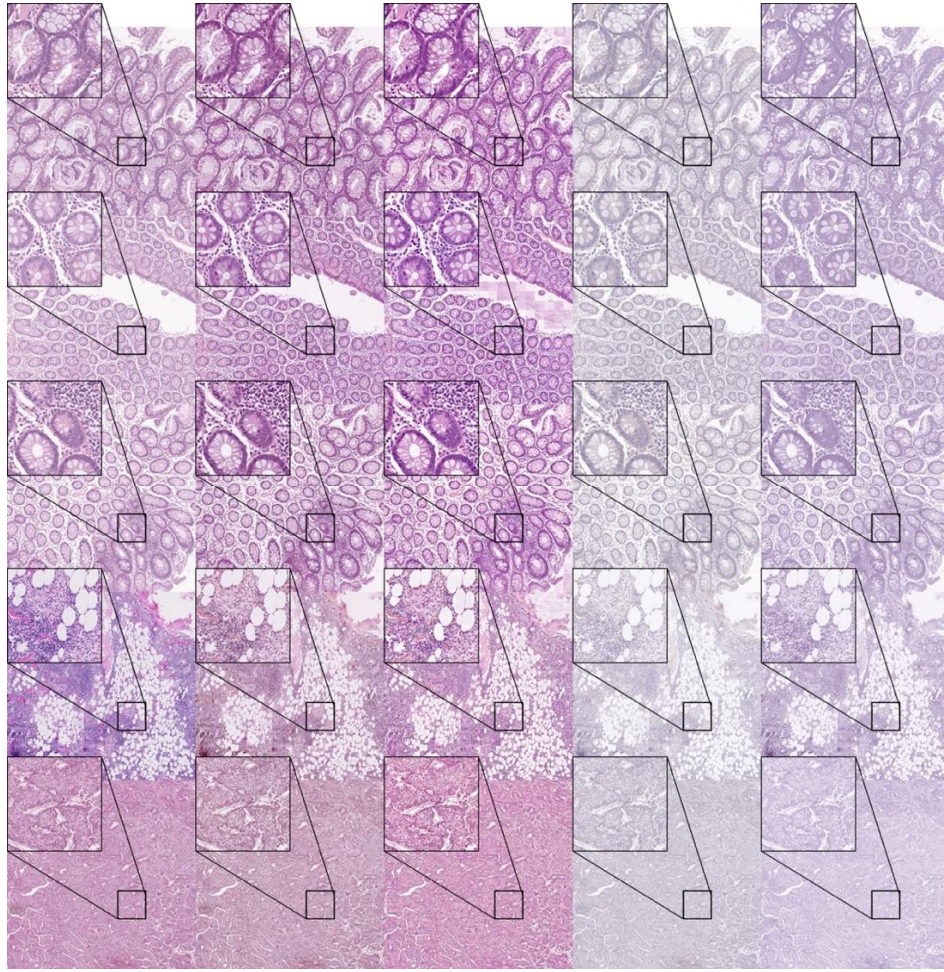
The Pearson correlation coefficient (PCC) is used in this experiment to evaluate color consistency and compare intensity statistics over the whole slide image. where S and P are the source whole slide image and processed whole slide image, respectively. And  $\mu_S$  and  $\mu_P$  are the mean of original and processed images, respectively. Also, the image quality is evaluated as multiple quantitative measurements, including universal quality image index (UQI) [105], erreur relative globale adimensionnelle de synth`ese (ERGAS) [106], multiscale structural similarity index (MS-SSIM) [107], peak signal-to-noise ratio (PSNR), root mean square error (RMSE).

$$PCC(S, P) = \frac{\sum_i (S_i - \mu_S)(P_i - \mu_P)}{\sqrt{\sum_i (S_i - \mu_S)^2} \sqrt{\sum_i (P_i - \mu_P)^2}}$$

## 5.5. Result

### 5.5.1. Quantitative and Qualitative Results

Stain normalization was performed on WSI with a patch-wise SDM model trained using stain separation. We tested the SDM model with an overlapping ratio of  $\gamma_{ratio} = 0.05$ . The Pix2PixHD model was also trained in a self-supervised manner with the grey2color method [108]. Two Pix2PixHD models, like ours, were trained to translate from grey to hematoxylin and eosin stains, respectively. In addition, we tested the Macenko [96] and Vahadane [97] method using a conventional method. The result of Pix2PixHD has observed the grid artifact, which is performed by patch-wise stain normalization. However, the SDM model has not observed the grid artifact even though performed by patch-wise stain normalization. The result is shown in **Figure 11**.



**Figure 11.** Results of stain normalization with whole slide image of 09, 12 and 15 years, and two external datasets, Camelyon and PAIP.

**Table 5** shows a quantitative metrics with two-conventional stain normalization methods [96, 97], GAN-based method [108] and ours. Our method outperformed Vahadane and Macenko in most metrics. We interpreted that the conventional methods were successful since we normalized to a histopathological image stained by the same institute. However, as shown in **Figure 11**, when normalizing histopathological images from different protocols, the color of the cell does not normalize compared to tissue in the conventional methods. On the other hand, our method's obvious normalization between cell and tissue was performed. Also, all quantitative results from external validation are inferior when using conventional

methods. Therefore, because the conventional method overfits the histopathological image of the same institution, the metric was showed to be high. Furthermore, forcing the reference image in conventional methods has some limitations.

**Table 5.** The quantitative comparisons with whole slide image from 09, 12, and 15 years. As well, the PAIP19 and CAMELYON16 datasets were evaluated the metrics: UQI (uni- versal quality index), ERGAS (erreur relative globale adimensionnelle de synth`ese), MS- SSIM (multi-scale structural similarity index), PSNR (Peak signal-to-noise ratio), RMSE (Root mean square error) and PCC (Pearson correlation-coefficient).

	UQI	ERGAS	MS-SSIM	PSNR	RMSE	PCC
09 Years						
Ours	0.9905	3356	0.9756	24.31	15.52	0.9901
Pix2PixHD	0.9775	4759	0.9263	21.90	20.76	0.9639
Macenko	0.9928	2744	0.9730	25.96	13.80	0.9706
Vahadane	0.9863	4101	0.7740	21.75	21.56	0.8419
12 Years						
Ours	0.9809	4351	0.9516	21.61	21.18	0.9900
Pix2PixHD	0.9739	5599	0.8967	20.38	24.68	0.9555
Macenko	0.9970	1836	0.9861	29.02	9.33	0.9758
Vahadane	0.9927	3120	0.8905	24.05	16.09	0.9380
15 Years						
Ours	0.9921	2399	0.9795	26.42	12.18	0.9928
Pix2PixHD	0.9833	4966	0.7070	19.97	25.57	0.9786
Macenko	0.9937	2481	0.9792	26.11	13.02	0.9779
Vahadane	0.9907	3178	0.8424	23.30	17.66	0.8964
PAIP19						
Ours	0.9794	4791	0.9697	23.85	16.20	0.9485
Pix2PixHD	0.9661	6363	0.9567	20.22	28.46	0.9324
Macenko	0.9327	8661	0.9099	17.06	40.47	0.9476
Vahadane	0.9280	9237	0.7930	16.00	44.13	0.8733
CAMELYON16						
Ours	0.9297	8513	0.9578	19.79	28.66	0.9568
Pix2PixHD	0.9063	9759	0.9592	17.52	37.62	0.9344
Macenko	0.8182	13729	0.8372	12.45	63.54	0.9016
Vahadane	0.8191	13395	0.6862	12.77	63.41	0.8130

### 5.5.2. Result of overlapping

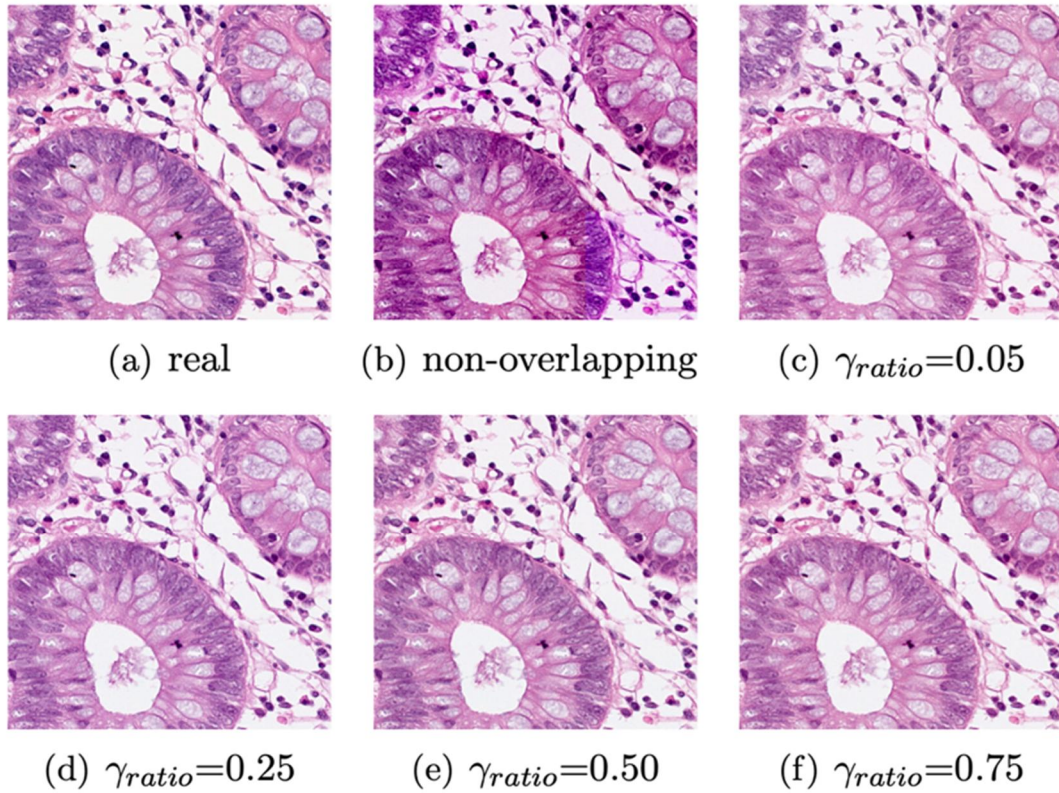
To confirm grid artifacts can be overcome by overlapping and to investigate the optimal coefficient of overlapping ratio, we normalized ratio  $\gamma_{\text{ratio}} = 0$  (non-overlapping), 0.05, 0.25, 0.50, and 0.75. Without overlapping, each patch can be normalized in various colors, as shown in **Figure 12(b)**, due to the powerful generative performance of the SDM model. However, the grid artifact was not observed when overlapping was used regardless of overlapping ratio  $\gamma_{\text{ratio}}$ .

**Table 6** shows a comparison of relative time cost and MS-SSIM, for  $\gamma_{\text{ratio}} = 0.00$  (non-overlapping), 0.05, 0.25, 0.50, and 0.75. The relative time cost can be saved quadratically as the overlapping ratio decreases while the performance improvement can be preserved.

**Table 6.** The relative time cost, MS-SSIM (multi-scale structural similarity index measure), PCC(Pearson correlation coefficient) and PSNR(peak signal-to-noise ratio) of patch, which is a  $512 \times 512$  patch of 2015 year WSI.  $\gamma_{\text{ratio}}$  refers to the overlapping ratio.

$\gamma_{\text{ratio}}$	.75	.50	.25	.05	Non-overlapping
Relative time cost	100%	30.9%	19.8%	11.1%	10.1%
PCC	0.9917	0.9920	0.9893	0.9883	0.9878
MS-SSIM	0.9237	0.9167	0.9097	0.9113	0.9029
PSNR	27.69	28.90	28.15	27.80	22.96





**Figure 12.** The results of various overlapping ratio.  $\gamma_{ratio}$  refers to the overlapping ratio.

### 6. 3D generation in brain CT

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 anatomy and pathology. However, access to 3D medical imaging datasets is severely limited by patient privacy laws. This inaccessibility can be largely circumvented by generating very realistic fake data. As is well known, data insufficiency or data imbalance can be overcome with a well-trained generative model [12, 13]. However, generating images with intact integrity and distribution in the 3D volume is very difficult because resources are limited. Generating satisfactory high-resolution images [10] in the 12-bit format of real clinical settings is also difficult. The present study proposes a 2D-based 3D-volume generation method that uses the previous slice to

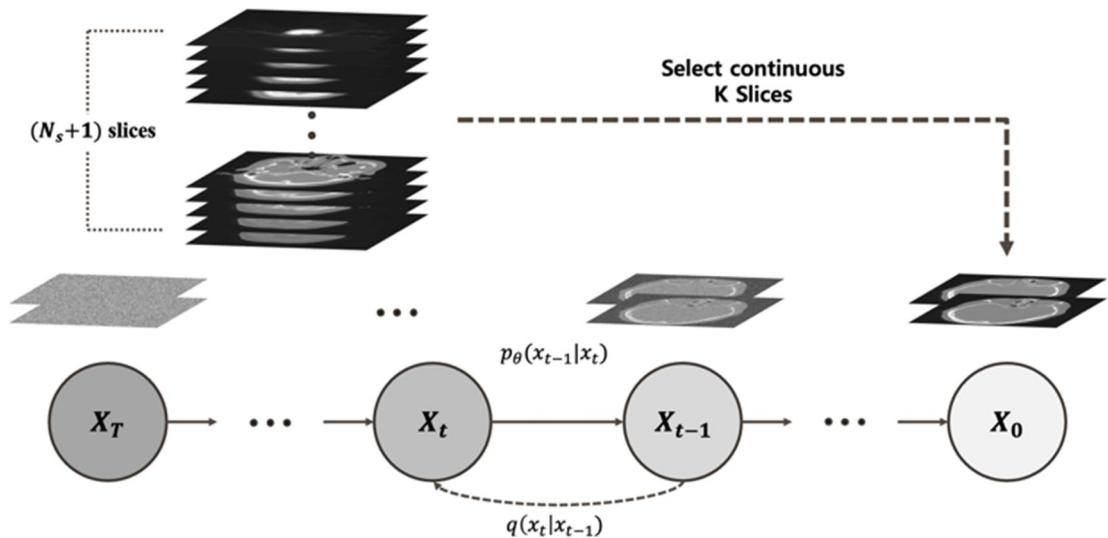
generate an adjacent axial slice. We call this method adjacent slice-based conditional iterative inpainting, *ASCII*.

*ASCII* is combined with a score-based diffusion model to generate images in 12-bit format. Experiments demonstrated that *ASCII* could generate 3D volumes with intact 3D integrity and distribution. However, in 12-bit format, the intensity windowing of cerebral parenchymal tissues was improperly calibrated in some slices. To resolve this problem, we proposed an additional novel intensity-calibration network (IC-Net), which is trained in a self-supervised manner to match the intensities of the previous and next-generated slices.

Finally, we will discuss 12-bit whole range generation. The anatomical structures, which are low signal regions like the parenchyma, collapse at the windowing view when GANs or VAEs generates in 12-bit Hounsfield unit whole range. We address the limitations of GAN and VAE for 12-bit whole range generation as well as the possibilities of the diffusion model.

## 6.1. Material and Methods

### 6.1.1. Adjacent Slice-based Conditional Iterative Inpainting, *ASCII*



**Figure 13.** The progress of score-based diffusion model for continuous K slices. We selected

continuous  $K$  slices among total  $N_s$  slices. To generate a continuous  $K$  slices, the score-based diffusion model is trained.

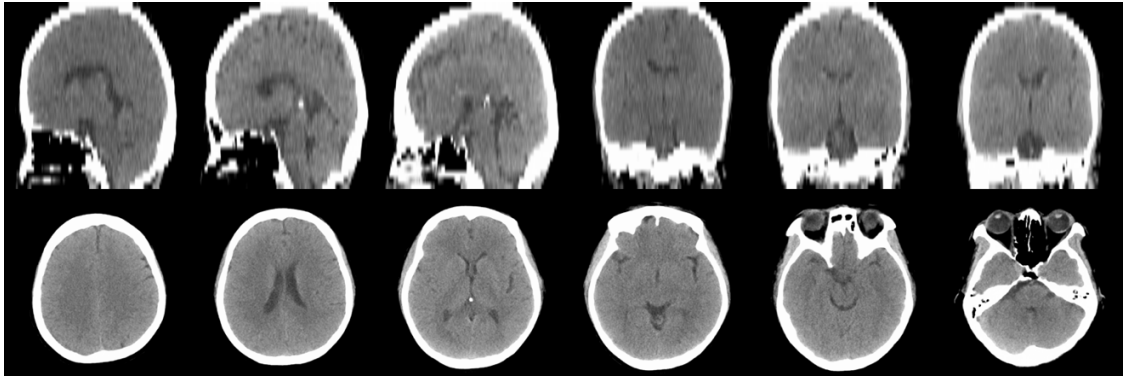
Our goal is to generate a 3D volumetric image in a slice-wise manner using a binary mask, which is moved along the channel axis and  $K = 2$  or  $3$  in the VESDE. A slice  $\mathbf{x}^0 = \{-1024HU\}^D$  filled with intensity of air was padded before the first slice  $\mathbf{x}^1$  of CT. Then, the input of the model was given by  $[\mathbf{x}^t; \mathbf{x}^{t+K-1}]$ , where  $t \in [0, N_s - K + 1]$  and  $N_s$  is the total slice number of CT. In addition, we omit augmentation because the model itself might generate augmented images. As shown in Figure 13, it presents a training schema.

After training the score network, we can solve unconditional stochastic process  $x_t^n$ , where  $x_t^n \sim \mathcal{N}(x_t^n | x^n, \sigma(t)^2)$  for generation. With the previous slice given as seed, to generate the next slice by inpainting is the key to our method. First, we defined a channel mask, which is a diagonal matrix  $\Lambda \in 0,1^{K \times K}$  with  $tr(\Lambda) = K - 1$  and  $\Lambda_{K,K} = 0$ . Next, let  $x^0$  be the initial seed slice of inpainting step, then  $x^1$  was generated by conditional stochastic process  $\{\hat{x}_t^1 | x^0\}_{t \in [0,1]}$ , where  $\hat{x}_t^1 = (I - \Lambda) \otimes x_t^1 + \Lambda \otimes x_t^0$ , where  $\otimes$  is channel-wise product as  $bchw, cc' \rightarrow bc'hw$  in Einstein notation and it was set to be the seed of generating next slice  $x^2$ . After setting the slice generated from the previous step as the seed of next step, the model iteratively generates a next adjacent slice. Finally, when the whole process is performed iteratively, the volumetric CT was sequentially generated by solving conditional stochastic processes  $\{\hat{x}_t^n | x^{n-1}\}_{t \in [0,1], n \in [1, N_s]}$ , where  $\hat{x}_t^n = (I - \Lambda) \otimes x_t^n + \Lambda \otimes x_t^{n-1}$ . We call this method *adjacent slice-based conditional iterative inpainting, ASCII*. We performed two experiments with CT volumetric image generation using ASCII. The first experiment performed in the windowing range generation, and the second performed in the whole range generation.



### 6.1.2. Result of ASCII

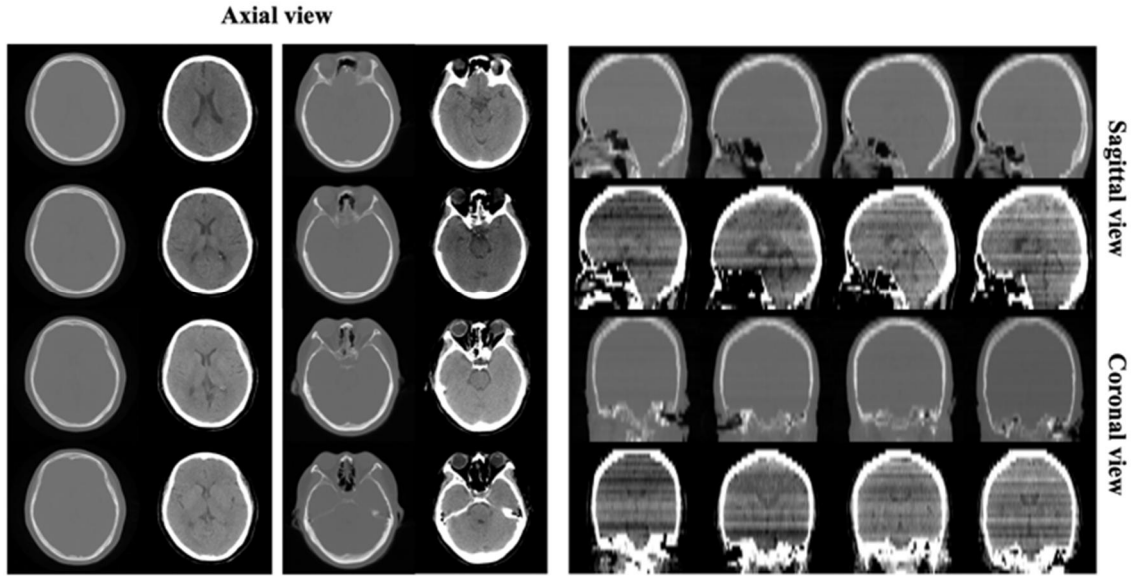
We experimented  $\sigma_{min}$ ,  $\sigma_{max}$  and K to 0.01, 1,348 and 2 with VESDE, respectively. Also, total timesteps of stochastic process  $T$  set 1,000. And, the CT slices were clipped by [-10HU, 70HU] and normalized from [-10HU, 70HU] to [-1, 1]. **Figure 14** shows the generated results, there is no discernible difference between reals and fakes. Both white matter and grey matter can be clearly distinguished, as well as the continuity between adjacent slices is kept properly. The result of sequential synthesis with previous slices is presented in Figure 2.



**Figure 14.** (Up) Results of ASCII trained in windowing range with sagittal and coronal views. (Down) Results of ASCII trained in windowing range with axial view.

### 6.1.3. Result of ASCII in 12-bit whole range

We experimented  $\sigma_{min}$ ,  $\sigma_{max}$  and K to 0.001, 68 and 2 with VESDE, respectively. And, the CT slices were normalized from [-1024HU, 3071HU] to [-1, 1]. The result of sequential synthesis with previous slices is presented in Figure 5, the model is well generated in each axial view image. The anatomical structure was consistent and well-generated even in the windowing view. However, the contrast of each image is not calibrated and therefore, the stripe artifacts are created in sagittal and coronal image as shown in **Figure 15**. Theoretically,  $\sigma_{min}$  and schedule are expected to be solved by creating smaller and longer, but this cannot be performed because the computational cost grows.



**Figure 15.** (Left) Results of ASCII with axial view in whole and windowing range. (Right) Results of ASCII with sagittal and coronal view in whole and windowing range.

#### 6.1.4. Intensity Calibration Network

It was noted that this problem only occurs in 12-bit generation. To normalize the intensity of the parenchyma area, we used a conventional non-trainable post-processing, such as histogram matching. The intensities of each slice of 3D CT can be calibrated by histogram matching, however, anatomical region is collapsed accordingly. Because each slice has different anatomical structure, the histogram of each slice image was fitted to their subtle anatomical variation. Finally, we proposed a solution for this intensity mismatching through trainable intensity calibration network: IC-Net.

In a recent study [109], additional network was trained to generate elaborate images. Our objective is to calibrate intensity, we proposed a training method of 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  $\hat{x}_c^{t+1}$ ,

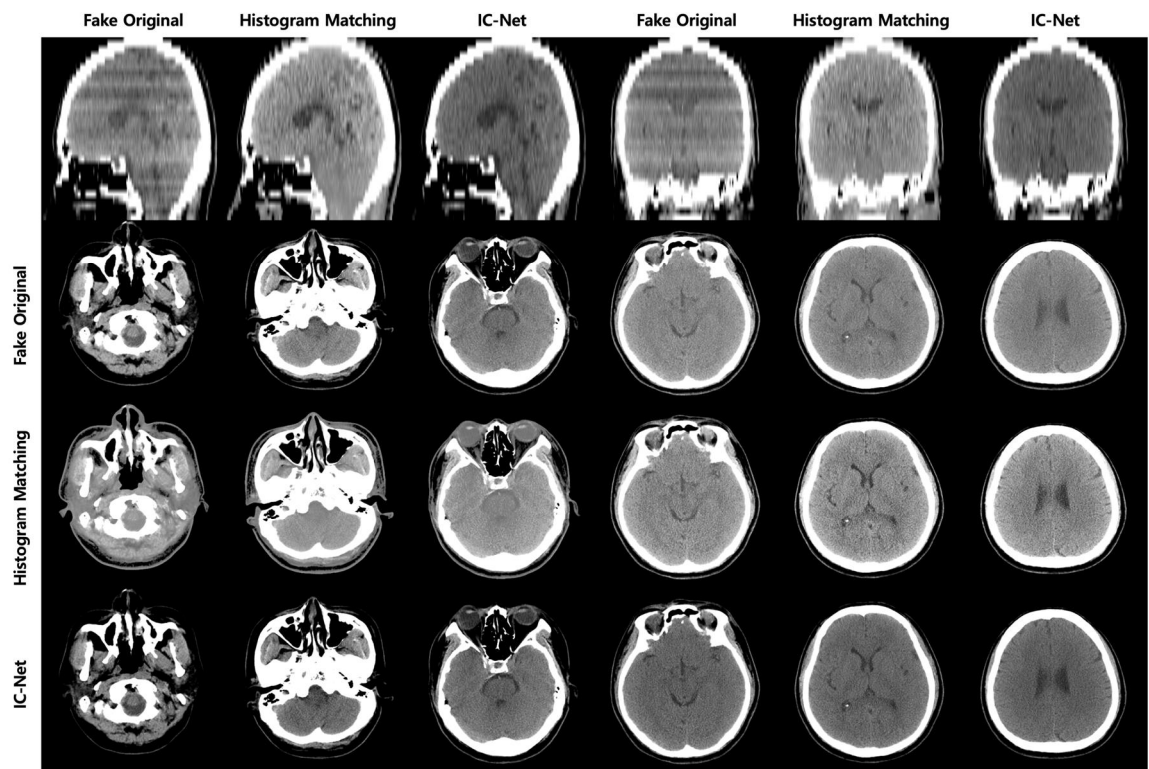
$$\hat{x}_\mu^{t+1} = (x^{t+1} - \overline{x^{t+1}}) * \mu + \overline{x^{t+1}}$$

where  $\overline{x^{t+1}}$  and  $\mu$  are the mean of  $x^{t+1}$  and shifting coefficient, respectively.

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 to preserve the subtle texture and the shape of a generated slice and only calibrate the intensity of  $x^t$ . The current slice is normalized by the IC-Net using the prior slice. The loss function of IC-Net is given by,

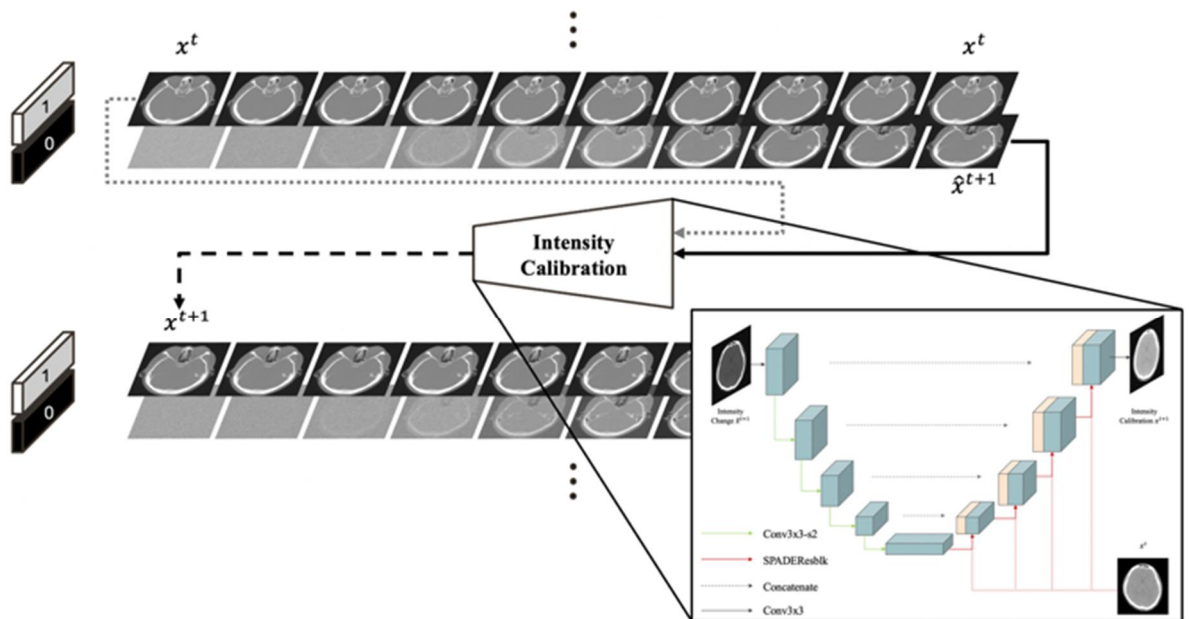
$$\mathcal{L}_{IC} = \mathbb{E}_t \mathbb{E}_{\mu \sim U[-0.7, 1.3]} [ |IC-Net(\hat{x}_\mu^{t+1}, x^t) - x^{t+1}| ]$$

As shown in **Figure 16**, some important anatomical structures, such as midbrain, pons, medulla oblongata, and cerebellar areas, are blurred and collapsed when histogram matching was used. This is a risky method as the outcomes vary depending on the matching seed. On the other hand, In the anatomical structure, the IC-Net did not collapse. Also, there is no requirement to specify the seed because normalization is carried out using the produced adjacent slice.



**Figure 16.** Result of post-processing. First row is generated by ASCII and second and third row are post-processing of first row using IC-Net and histogram matching, respectively.

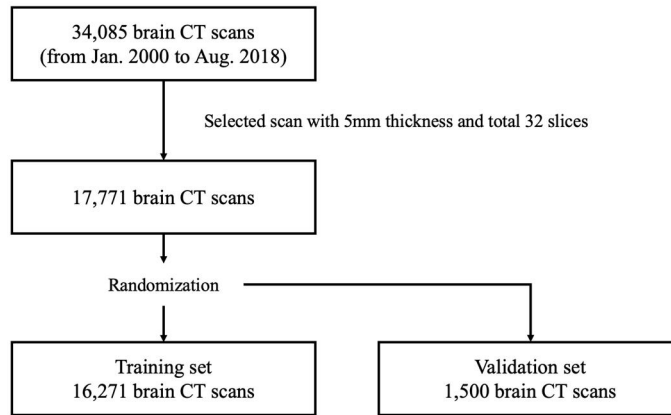
Finally, the overview of *ASCII with IC-Net* is shown in **Figure 17**.



**Figure 17.** The overview of ASCII(2) with IC-Net. We set seed  $x^0$  to fill with intensity of air (-1024 HU) and channel mask  $\Lambda$ .  $x^1$  was generated by using seed  $x^0$ . The brain CT volume data was generated by iterative processing this procedure with changed seeds. The model architecture of IC-Net is presented.

### 6.1.5. Dataset and model architecture

A total of 34,085 non-contrast brain CT scans and paired radiology reports were retrospectively collected from patients who visited an urban, tertiary, academic hospital between January 1, 2000, and August 31, 2018. Among the scans, we only selected CT scans, which consist of 32 slices, each of 5-mm thick. The study protocol was approved by institutional review board of Asan Medical Center and Gangneung Asan Hospital. Finally, we split 1,500 brain CT scans at random to evaluate metrics. **Figure 18** summarizes the data collection and curation process.



**Figure 18.** Data flow of diagram and process of curation.

We defined a score network as  $ncsnpp^1$  suggested. The IC-Net has a U-Net [93] based architecture and instance normalization [110-112], and the decoder is defined by SPADE [52] residual block for embedding the previous slice.

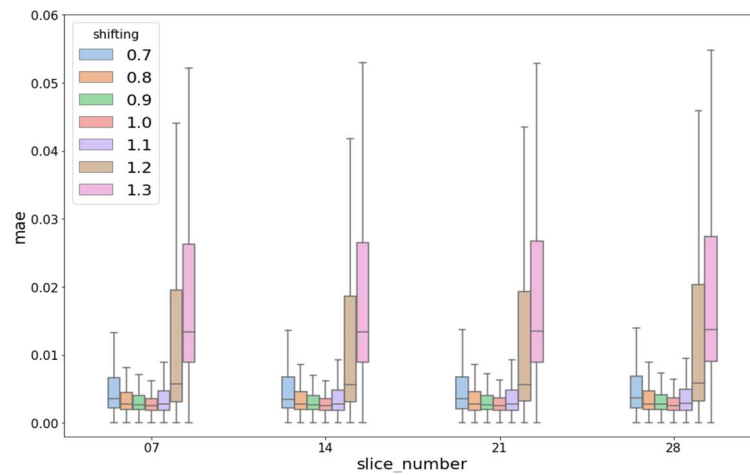
## 6.2. Experiments

### 6.2.1. Results of intensity calibration network

To demonstrate the performance of IC-Net, we conducted experiments with the 7, 14, 21 and 28th slices, which are complex enough 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.

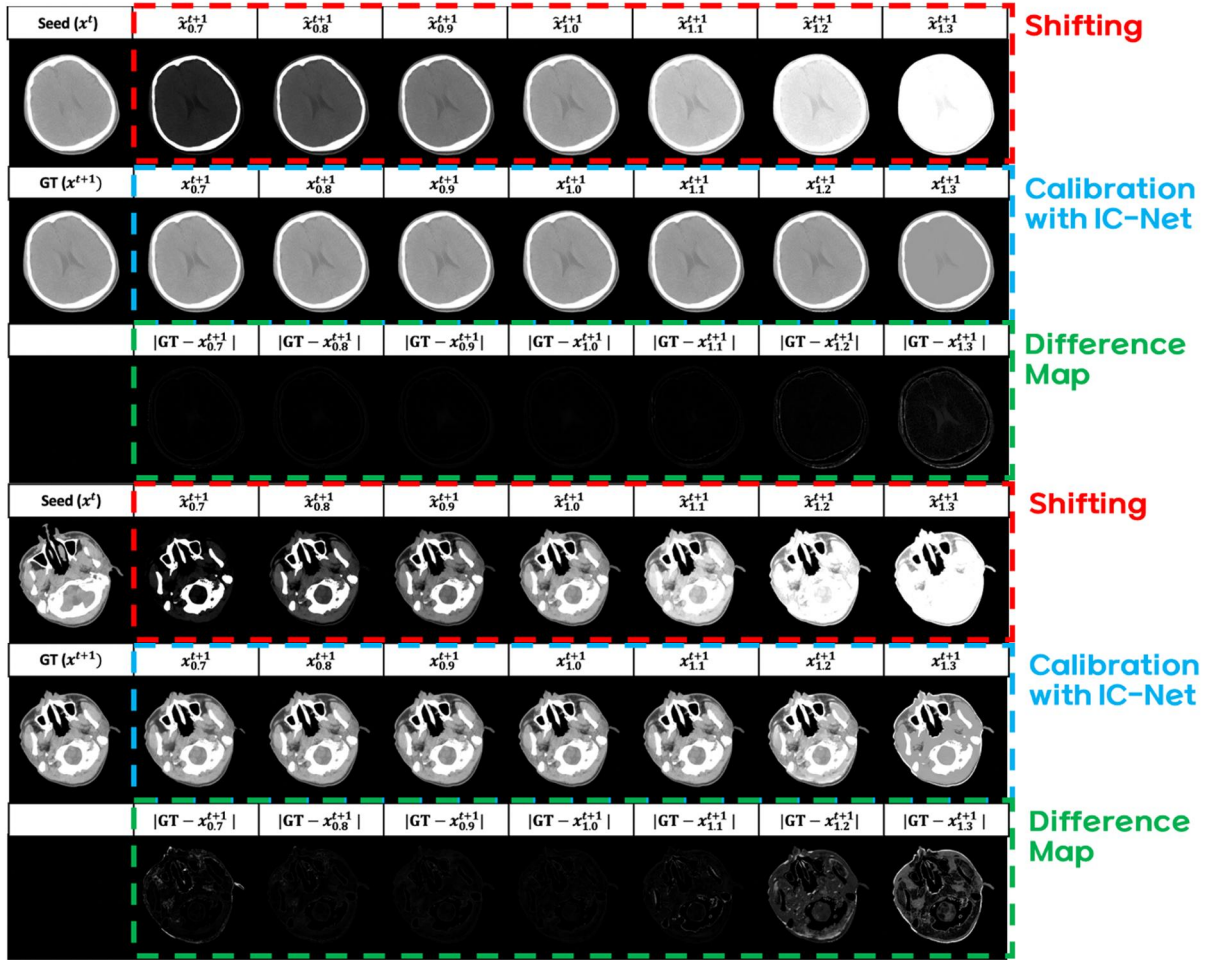
---

<sup>1</sup> [https://github.com/yang-song/score\\_sde\\_pytorch](https://github.com/yang-song/score_sde_pytorch)



**Figure 19.** The difference of between GT and prediction by each slice number. All slices were normalized from [-150HU, 150HU] to [-1, 1] and the mean average error was shown as normalized range on vertical axis.

As shown in **Figure 19**, 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, qualitatively, IC-Net can calibrate intensity to some extent even in the collapsed images as shown **Figure 20**.



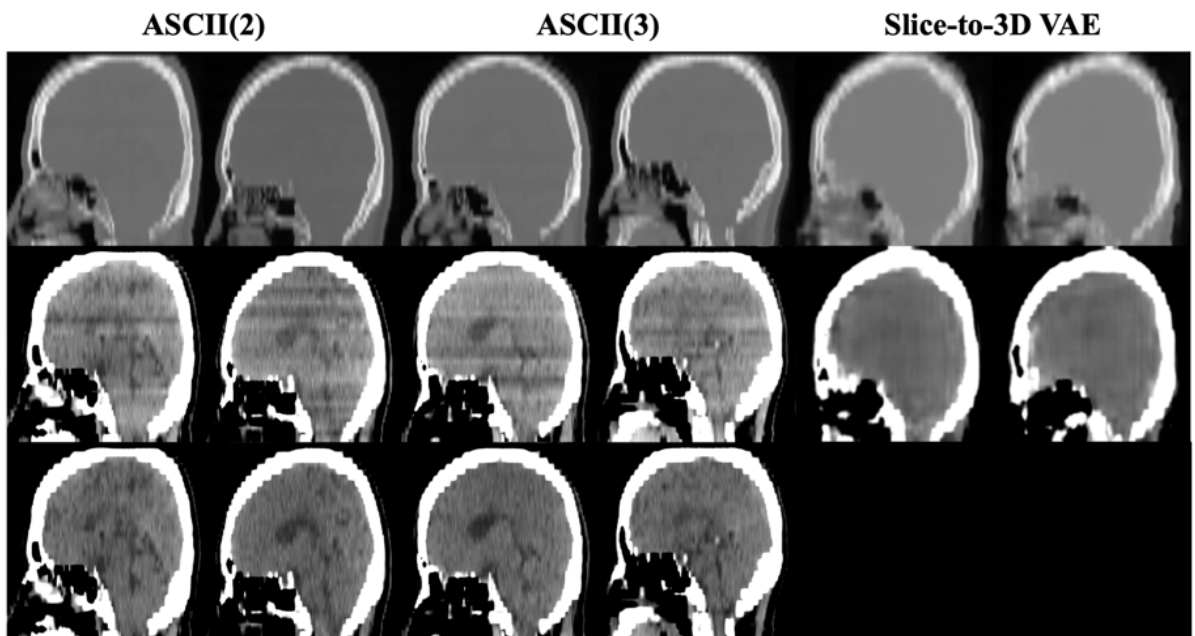
**Figure 20.** Result of IC-Net with slices shifted by fixed values. We used a fixed value from 0.7 to 1.3. Using a seed and shifted slice for intensity calibration, the difference map with GT was illustrated.

### 6.2.2. Results of ASCII with IC-Net in whole range

We experimented on continuous K slices of K=2 and 3 called ASCII(2) and ASCII(3), respectively. 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 [113]. **Figure 21** 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 **Figure 2 of 3.2**, was also shown in Slice-to-3D VAE.



**Figure 21.** Result of ASCII(2) and ASCII(3) with and without IC-Net, and. And last column is result of bone rendering using 3D Slicer [114].

### 6.2.3. Quantitative Evaluation

We generated a head & neck CT images via ASCII(2) and ASCII(3) with and without IC-Net, and slice-to-3D VAE [113]. The volumes were generated in a whole dynamic range of Hounsfield unit. Therefore, the performance of 3D generation can be evaluated in both the whole range and the windowing range. The axial middle slice, sagittal middle slice, and

coronal middle slice of the generated volumes were evaluated using 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<sup>2</sup>) were evaluated to quantitative metrics. The MS-SSIM and bMMD<sup>2</sup> are known to be able to properly measure the diversity of generated images and the distance between two distributions of finite samples of mini-batch, estimated with kernel functions in the reproducing Hilbert space.

The quantitative results in whole range are shown in **Table 7**. In general, ASCII(2) performs better than ASCII(3). 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 generate it diverse enough.

**Table 7.** Quantitative result of generated image by ASCII(2), ASCII(3) and Slice-to-3D VAE in whole range and windowing range.

	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
<b>Whole Range</b>					
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
bMMD <sup>2</sup>	425704	429120	428045	432665	311080
<b>Windowing Range</b>					
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 <sup>2</sup>	1975336	1854921	2044218	1858850	1894911

*Note: IC-Net, Intensity Calibration Network; FID, Fréchet Inception Distance; Ax, Axial; Cor, Coronal; Sag, Sagittal; MS-SSIM, Multi-Scale Structural Similarity Index Measure; bMMD<sup>2</sup>, batch-wise squared Maximum Mean Discrepancy. And whole range and windowing range are [-1024HU, 3071HU] and [-10HU, 70HU], respectively.*

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.

#### **6.2.4. Qualitative Evaluation**

Slice-based methods typically have issues with weak connectivity and three-dimensional (3D) integrity among generated slices. To evaluate the 3D integrity of the generated 3D images, the images were evaluated by an expert radiologist with more than 15 years of experience. 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. Fifty real and fifty fake CT scans were blindly evaluated by visual scoring on three scales focusing on the continuity of eight anatomical structures: skull (bone morphology and suture line), skull base (foramina and fissure), facial bone, ventricles, brain sulci and fissures, the basilar artery, the cerebral venous sinus, and the ascending and descending nerve fiber tracts through the internal capsule. Scales of 1, 2, and 3 represent 'discontinuity', 'strained continuity', and 'well preserved continuity', respectively. The visual scoring results are shown in Table 2. Most of the fake images were scored similarly to the real CT scans, but the continuity of the basilar arteries was evaluated as broken in the fake images (**Table 8**). The basilar artery is a small region, especially in the axial view, and was frequently not generated. As the model was trained on 5-mm thickness non-contrasted enhanced head &

neck CT scans, preserving the continuity of the basilar artery is excessively demanding.

**Table 8.** Result of the integrity evaluation.

	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 and descending nerve fiber tract through internal capsule	3.00	3.00

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

### ***7. Post-surgery imaging generation in cephalogram***

Orthognathic surgery (OGS) has been widely performed to correct severe dentofacial deformities. The objective of surgery is to obtain balance among esthetics, function, and stability and ensuring patient satisfaction. Orthognathic surgery is a surgical procedure aimed at correcting severe disharmony in the facial skeletal structure to restore jaw and facial function and enhance facial aesthetics. Among them, patients place greater emphasis on the improvement of facial aesthetics [115, 116]. Therefore, predicting the changes in facial appearance after surgery becomes an important factor in the patient's decision-making process.

Our method consists of three main steps. Firstly, we measure the landmark points for a total of 45 points on the pre-operation images. Then, using the images and landmark points, we will employ a model called Surgical Movement Prediction (SMP) to predict the displacement of the landmark points after surgery. Subsequently, we will perform post-surgical image generation using the predicted movement, landmark points, and the patient's profile line as

prompts through the Pre2Post model. Our predicted results can assist dentists in surgical planning.

## **7.1. Material and Methods**

### **7.1.1. Datasets**

As the nine university hospitals had different types of cephalography machines and radiation exposure protocols, these conditions could produce different lateral cephalogram qualities (Table 2). A total of 800 pairs of lateral cephalograms, taken at pre- (T0) and post- (T1) surgery, were deidentified and stored in a Digital Imaging and Communications in Medicine file format as 12-bit deep grayscale images.

Among them, 599 pairs of T0 and T1 lateral cephalogram image data from university hospitals Seoul National University Dental Hospital (SNUDH), Kyung Hee University Dental Hospital (KHUDH), and Kyungpook National University Dental Hospital (KNUDH) were used as internal datasets. As most patients at hospitals SNUDH and KHUDH had undergone two-jaw surgery ( $n=345/369$  and  $n=153/193$ , respectively), data imbalance may occur. To solve this problem, we added data from hospital KNUDH, where most patients had undergone one-jaw surgery ( $n=30/37$ ). Among 599 pairs of lateral cephalogram images, 399 pairs were used as the training set, 100 pairs as the validation set, and 100 pairs as the internal test set. Consequently, 201 pairs of T0 and T1 lateral cephalogram image data from university hospitals Wonkwang University Dental Hospital, Korea University Dental Hospital, Ehwa University Medical Center, Chonnam National University Dental Hospital, Ajou University Dental Hospital, and Asan Medical Center were used as external test sets.

Before training, all lateral cephalogram images were resized to unify the pixel interval to 0.1 and trimmed with the landmark coordinates to learn the area required for surgical planning from original lateral cephalogram images. Because of the original image's ratio, the cropped images were adjusted to be smaller than  $1024 \times 1024$  pixels.

When cropping the image from the landmark coordinates, the modified landmark coordinates were obtained by subtracting the coordinates of the upper left corner. Subsequently, the cropped image was divided by the maximum pixel value of the image. Pixel normalization was performed such that the pixel value was within 0~1. Before input to the AI model, the x- and y-axis distances were divided by the width and height of the cropped picture, and normalization was performed such that the feature value was within the range of 0 ~ 1. T0 and T1 lateral cephalogram images were superimposed by the sella-nasion (SN) line.

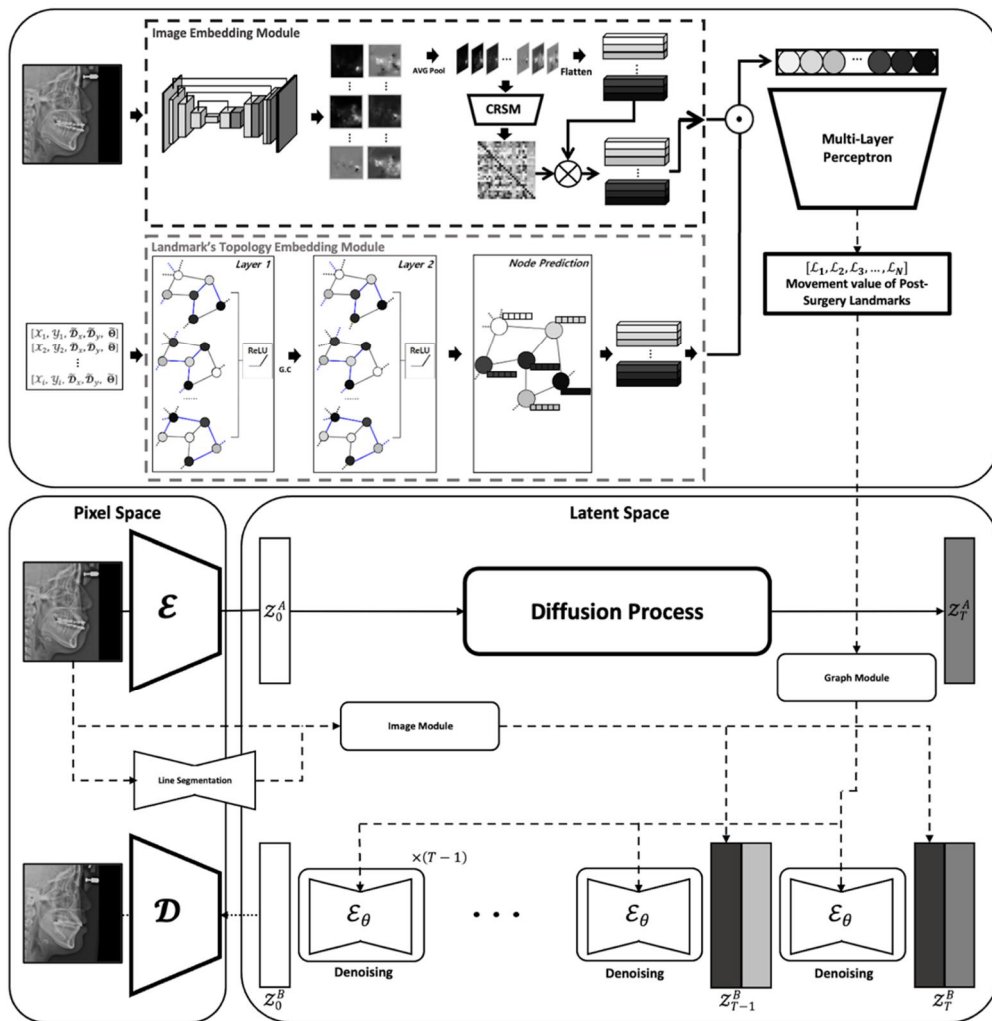
### ***7.1.2. Model Architecture***

We proposed using DentalNet to generate T1 lateral cephalogram images. DentalNet consists of two models: the surgical movement prediction (SMP) model, which predicts the movement of landmarks by the surgery, and the Pre2Post model, which generates T1 lateral cephalogram images.

SMP model predicts the movement of landmarks from orthognathic surgery. To predict this, we used an image embedding module (IEM) based on the high-resolution network (HRNet) [117], which is suggested as a method for maintaining high-resolution representations by connecting other resolution convolution streams in parallel and exchanging information across resolutions, architecture to embed high-resolution cephalometric representations. And we used a graph convolution network (GCN)-based [118] landmark topology structure embedding module (LTEM) that training the topological structures of 27 hard tissue landmarks, including the spatial relationships between landmarks. Furthermore, the T1 coordinates of the 24 landmarks were predicted using a multi-layer perceptron (MLP) module, by concatenating the outputs of the IEM and LTEM modules and passing them through the MLP module.

To generate post-cephalogram images, we used various prompts, including the movement of landmarks obtained through SMP. The purpose is to ensure more realistic and detailed generation quality using various prompts to provide information for surgery planning. To

guarantee a minimum generation ability, we trained an autoencoder using not only a labeled dataset of T0 and T1 but also an unlabeled cephalogram dataset. We trained a latent diffusion model for image generation in the encoding space with various prompts as conditioning factors through the Pre2Post module. Specifically, T0, T0's landmark, line segmentation, and movement prediction value were used as prompts. Each time we added more prompts, we observed an improvement in the quality and fidelity of generated images. The combined two modules are referred to as *DentalNet*, and an overview of *DentalNet* is shown in **Figure 22**.



**Figure 22.** The overview of *DentalNet*. (*Top*) We utilize a surgical movement prediction (SMP) model composed of an image embedding module and a graph-based module to predict the displacement of landmark points after surgery. (*Bottom*) Afterwards, using the measured

displacement and various prompts, we generate post operation images using the diffusion model (Pre2Post).

### ***7.2. Surgical Movement Prediction***

We evaluated the performance of the smp model in predicting the movement amount of hard and soft tissues using internal and external data. Only landmarks that were displaced by the surgery, including 19 hard tissue and 14 soft tissue landmarks, were used in the experiments. The results from the experiments using internal data showed that the hard and soft tissues had errors of 1.34 and 1.18, respectively, while the errors were 1.58 and 1.28, respectively, in the experiments using external data. Statistical analysis of the experimental results revealed no significant difference between the internal and external data. These findings suggest that the proposed smp model has robust performance not only with internal data but also with external data, making it a valuable tool for predicting hard and soft tissue displacement in post-surgery x-ray generation using dental imaging.

### ***7.3. Post-surgery imaging generation***

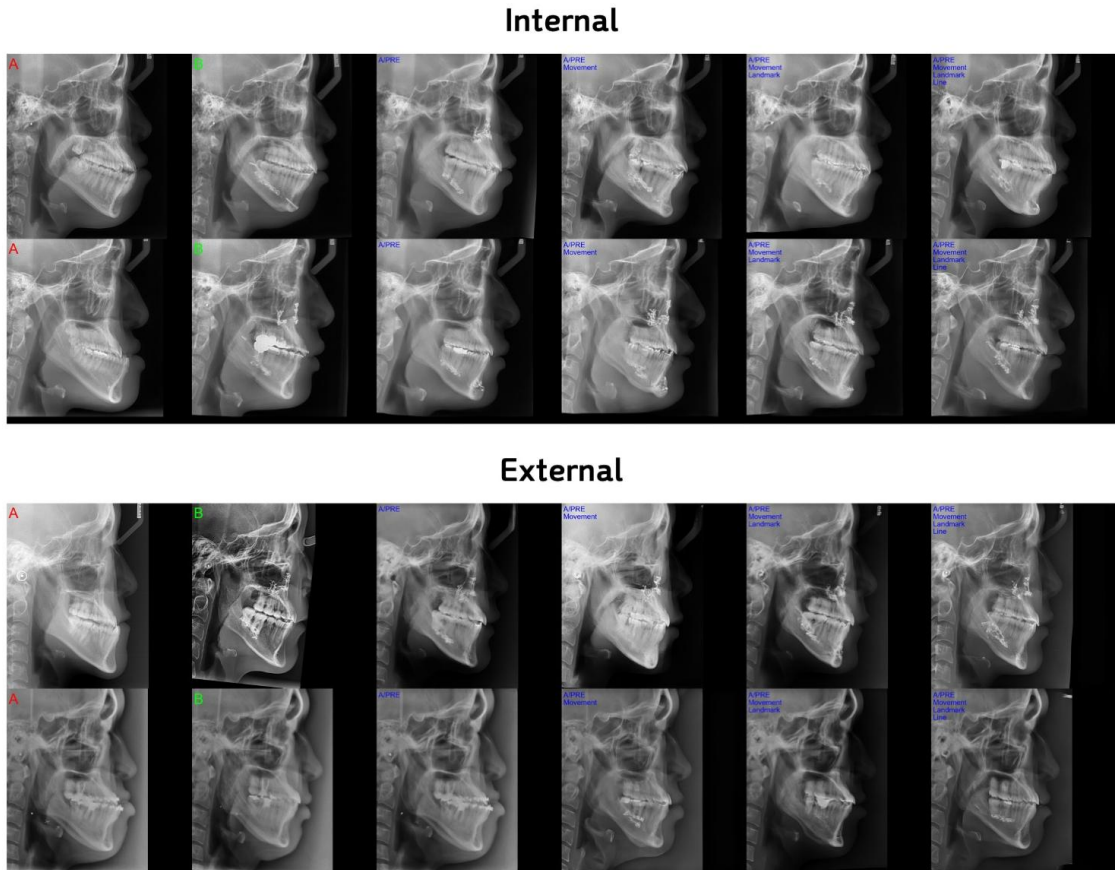
To ensure minimal generation capability, we trained an autoencoder using both pre-surgery images and post-surgery images, as well as unlabeled data. Additionally, we incorporated vectorize quantization (*vq-f16*) and adversarial learning to enhance training stability and achieve high fidelity. Subsequently, using the SMP module, we generated post-surgery imaging utilizing not only the predicted movement values but also the pre-surgery images, pre-surgery landmarks, and profile line as prompts. We trained a latent diffusion model by alternating between pre and post with appropriate probabilities. During the training of pre-surgery, we used null prompts for all prompts except the pre-surgery image, whereas during the training of post, we utilized all prompts. The objective function is as follows.



$$\mathcal{L}_{ldm} = \begin{cases} \|\epsilon_{\theta}(z_{pre}, t, \emptyset) - \epsilon\|_2 \\ \|\epsilon_{\theta}(z_{post}, t, c) - \epsilon\|_2 \end{cases}$$

where  $z_{pre}$ ,  $z_{post}$ ,  $\emptyset$  and  $c$  are the encoded pre-operation image, encoded post-operation image, null prompt, and prompts, respectively.

During inference, we employed DDIM sampling for consistency. Additionally, we conducted experiments using four different prompt combinations: (1) pre-surgery image, (2) pre-surgery image and movement value, (3) pre-surgery image, movement value, and landmark, and (4) pre-surgery image, landmark, profile line and movement value. The result for each prompt is shown in **Figure 23**.



**Figure 23.** Results of the predicted post-surgery images. **A** and **B** represent the pre-surgery image and post-surgery image, respectively. Additionally, the prompts written at the top of

each image indicate the results of the trained model using the respective prompts.

In orthognathic surgery, there should be minimal changes in the upper facial bone. Therefore, the fact that our predicted post-surgery images show minimal differences compared to the pre-surgery photographs in terms of the upper facial bone is medically plausible.

Additionally, unless undergoing surgical intervention, the menton should remain unchanged.

We can also observe that it undergoes minimal changes, especially when the line profile is added, resulting in a highly similar appearance of the menton in the generated post surgery images.

**Table 9.** The generated images in terms of the 30 changing landmarks due to surgery are compared with the ground truth (GT) using an error metric. These landmarks were measured by DDS with more than 10 years of experience. After measurement, like the preprocessing steps, alignment was performed based on the SN line for accurate measurements.

	Internal	External		Internal	External
A-Point	1.29 ± 0.97	1.47 ± 0.71	Mandible 6 distal	1.58 ± 1.05	1.78 ± 1.31
PM	1.51 ± 0.90	1.62 ± 1.10	Mandible 6 root	1.7 ± 1.12	1.93 ± 1.38
Pogonion	1.46 ± 0.85	1.62 ± 0.95	Point on Upper profile	1.50 ± 0.96	2.01 ± 2.40
B-point	1.51 ± 0.98	1.71 ± 1.16	Pronasale	0.96 ± 0.78	1.11 ± 0.93
Posterior Nasal Spine	1.70 ± 1.35	1.82 ± 0.92	Columella	0.85 ± 0.64	1.06 ± 0.92
Anterior Nasal Spine	1.38 ± 0.87	1.83 ± 1.15	Subnasale	0.99 ± 0.61	1.04 ± 0.67
R1	1.52 ± 0.82	1.90 ± 0.99	Soft tissue A	1.05 ± 0.73	1.07 ± 0.84
Ramus down	2.55 ± 1.67	3.10 ± 1.64	Labrale superius	1.07 ± 0.67	1.12 ± 0.85
Corpus left	2.72 ± 1.86	2.76 ± 1.53	Upper Lip	0.98 ± 0.53	1.21 ± 0.98
Menton	1.33 ± 0.86	1.68 ± 1.03	Stms	1.10 ± 0.56	1.19 ± 0.86

Maxilla 1 crown	1.28 ± 0.68	1.36 ± 0.82	Lower Embrasure	1.39 ± 0.85	1.52 ± 1.04
Maxilla 1 root	1.25 ± 0.69	1.43 ± 0.88	Stmi	1.23 ± 0.74	1.44 ± 0.74
Mandible 1 crown	1.30 ± 0.76	1.39 ± 0.80	Lower Lip	1.18 ± 0.80	1.72 ± 1.00
Mandible 1 root	1.53 ± 0.92	1.55 ± 0.88	Soft tissue B	1.95 ± 1.22	2.37 ± 2.01
Occlusal plane point	1.17 ± 0.63	1.60 ± 1.00	Soft tissue Pogonion	1.56 ± 0.84	2.10 ± 1.94
Maxilla 6 distal	1.54 ± 0.09	1.68 ± 1.11	Soft tissue Mention	1.34 ± 0.89	1.64 ± 1.01
Maxilla 6 root	1.59 ± 0.92	1.65 ± 1.21			

*Note: Unit:mm, The landmarks for hard tissue range from **A-point** to **Mandible 6 root**, while the landmarks for soft tissue range from **Point on Upper profile** to **soft tissue Mention**.*

**Table 9** represents the differences between the generated images and the real post surgery images. The landmarks of the generated images were measured by a DDS with over 10 years of experience. A total of 50 images were measured for both the Internal and External datasets. The distance error for the 33 landmarks in the Internal dataset was measured to be 1.41, while for the External dataset, it was 1.65. Furthermore, in the Internal dataset, the distance error for landmarks in the hard tissue was measured to be 1.57, and for landmarks in the soft tissue, it was 1.25. In the External dataset, the distance error for landmarks in the hard tissue was 1.78, and for landmarks in the soft tissue, it was 1.47.

Almost all landmarks were measured within a medically significant range of 2mm or less, indicating plausible results. Furthermore, for the external aspect, although there are no significant differences observed in the hard tissue, there is an average error of approximately 0.2mm measured for each soft tissue landmark. This can be considered a reasonably acceptable error range, as it can vary significantly depending on the imaging equipment and surgical techniques employed by different institutions.

## ***8. Discussion***

Diffusion model is introducing a new paradigm in generative models and has shown remarkable progress by utilizing synthesis data in various applications. While its application in natural images has been extensively explored, its application in the medical domain is still emerging but anticipated. Therefore, we provide guidelines for the application of diffusion model in the medical imaging by performing data augmentation, image-to-image translation, and slice-based 3D generation. Additionally, we demonstrate its value in the medical domain through experiments applicable in medical settings, such as post-surgery image generation.

Diffusion model estimates the likelihood (score) based on Parzen-windowing kernel density estimation, specifically approximating Gaussian distribution, which allows for precise modeling of data distribution. As a result, diffusion model can be capable of generating high-depth images, not only in high signal-to-noise regions but also in low signal-to-noise regions. This is particularly advantageous in medical image generation, where complex structures like blood vessels and tissues in fundus photo images or mammography can be adequately generated, even challenging for other generative models. Furthermore, diffusion model demonstrates its capability in medical imaging by achieving indistinguishable or similar scores to real images in visual turing test on MRI and anatomical continuity tests on CT in axial axis (shown in Table 8). It can generate medically plausible images that maintain the structural integrity of pathological structures.

Due to the explicit density approximation employed in training and sampling, diffusion model is higher capacity conditional sampling compared to other models. It has a high potential in those other conditions, such as label, image, and text, can be created through prompts or condition sampling through classifier-guidance or classifier-free. Moreover, we showcase the pioneering experiment of utilizing the topological structure of graphs as prompts for generating post-operation images, incorporating a graph-based module prompt.

However, it is not always the case that the diffusion model outperforms GANs. Since the

diffusion model is trained based on the distribution of the data, it requires a large amount of data for diversity and may struggle to generate unseen data. Additionally, we observed that the single-model colorization of the diffusion model can pose challenges in clinical applications due to its outperformed generative capability. Therefore, it is crucial to carefully select the appropriate generative model according to the task or objective at hand.

## **9. Conclusion**

We conducted research to enhance deep learning models and aid surgical planning through the application of the diffusion model in medical imaging. Additionally, we demonstrated the clinical utility of the diffusion model through various medical images, providing guidelines for its utilization.

## **Reference**

- [1] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.
- [2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [3] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," *Advances in neural information processing systems*, vol. 33, pp. 19667-19679, 2020.
- [4] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, "Neural autoregressive distribution estimation," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 7184-7220, 2016.
- [5] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International conference on machine learning*, 2016: PMLR, pp. 1747-1756.
- [6] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," *Advances in neural information processing systems*, vol. 29, 2016.

- [7] Y. Song, C. Meng, and S. Ermon, "Mintnet: Building invertible neural networks with masked convolutions," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [8] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.
- [9] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*, 2015: PMLR, pp. 1530-1538.
- [10] S. Hong *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*: Springer, 2021, pp. 24-34.
- [11] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [12] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321-331, 2018.
- [13] C. Bowles *et al.*, "Gan augmentation: Augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018.
- [14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110-8119.
- [15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401-4410.
- [16] T. Karras *et al.*, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852-863, 2021.
- [17] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12104-12114, 2020.
- [18] P. Vincent, "A connection between score matching and denoising autoencoders,"

- Neural computation*, vol. 23, no. 7, pp. 1661-1674, 2011.
- [19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.
- [20] Y. Song, S. Garg, J. Shi, and S. Ermon, "Sliced score matching: A scalable approach to density and score estimation," in *Uncertainty in Artificial Intelligence, 2020*: PMLR, pp. 574-584.
- [21] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in neural information processing systems*, vol. 33, pp. 12438-12448, 2020.
- [22] D. Kim, S. Shin, K. Song, W. Kang, and I.-C. Moon, "Score matching model for unbounded data score," *arXiv preprint arXiv:2106.05527*, 2021.
- [23] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.
- [24] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
- [25] C. Saharia *et al.*, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *arXiv preprint arXiv:2205.11487*, 2022.
- [26] 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. 10684-10695.
- [27] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning, 2021*: PMLR, pp. 8162-8171.
- [28] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," *arXiv preprint arXiv:2103.16091*, 2021.
- [29] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021*, pp. 2837-2845.
- [30] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

- [31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840-6851, 2020.
- [32] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780-8794, 2021.
- [33] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [34] W. H. Organization, *WHO position paper on mammography screening*. World Health Organization, 2014.
- [35] N. Panwar *et al.*, "Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare," *Telemedicine and e-Health*, vol. 22, no. 3, pp. 198-208, 2016.
- [36] H. Phillips, S. Soffer, and E. Klang, "Oncological Applications of Deep Learning Generative Adversarial Networks," *JAMA oncology*, vol. 8, no. 5, pp. 677-678, 2022.
- [37] F. J. Moreno-Barea, J. M. Jerez, and L. Franco, "Improving classification accuracy using data augmentation on small data sets," *Expert Systems with Applications*, vol. 161, p. 113696, 2020.
- [38] J. E. Park, "Artificial Intelligence in Neuro-Oncologic Imaging: A Brief Review for Clinical Use Cases and Future Perspectives," *Brain Tumor Research and Treatment*, vol. 10, no. 2, p. 69, 2022.
- [39] J. E. Park, D. Eun, H. S. Kim, D. H. Lee, R. W. Jang, and N. Kim, "Generative adversarial network for glioblastoma ensures morphologic variations and improves diagnostic model for isocitrate dehydrogenase mutant type," *Scientific reports*, vol. 11, no. 1, pp. 1-11, 2021.
- [40] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, 2015: PMLR, pp. 2256-2265.
- [41] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794-2802.
- [42] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network,"



*arXiv preprint arXiv:1609.03126*, 2016.

- [43] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, 2017: PMLR, pp. 214-223.
- [44] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.
- [45] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
- [46] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.
- [47] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501-1510.
- [48] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
- [49] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798-8807.
- [50] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125-1134.
- [51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223-2232.
- [52] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337-2346.
- [53] A. Van Den Oord and O. Vinyals, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

- [54] I. Higgins *et al.*, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.
- [55] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *arXiv preprint arXiv:2104.07636*, 2021.
- [56] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.
- [57] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.
- [58] E. Platen and N. Bruti-Liberati, *Numerical solution of stochastic differential equations with jumps in finance*. Springer Science & Business Media, 2010.
- [59] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *Journal of computational and applied mathematics*, vol. 6, no. 1, pp. 19-26, 1980.
- [60] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.
- [61] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096-1103.
- [62] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," *Advances in neural information processing systems*, vol. 26, 2013.
- [63] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021: PMLR, pp. 8748-8763.
- [64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [65] C. Zerna, G. Thomalla, B. C. Campbell, J.-H. Rha, and M. D. Hill, "Current practice and future directions in the diagnosis and acute treatment of ischaemic stroke," *The Lancet*, vol. 392, no. 10154, pp. 1247-1256, 2018.

- [66] S. E. Gerard *et al.*, "CT image segmentation for inflamed and fibrotic lungs using a multi-resolution convolutional neural network," *Scientific reports*, vol. 11, no. 1, pp. 1-12, 2021.
- [67] N. Lassau *et al.*, "Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients," *Nature communications*, vol. 12, no. 1, pp. 1-11, 2021.
- [68] A. Arab *et al.*, "A fast and fully-automated deep-learning approach for accurate hemorrhage segmentation and volume quantification in non-contrast whole-head CT," *Scientific Reports*, vol. 10, no. 1, p. 19389, 2020/11/09 2020, doi: 10.1038/s41598-020-76459-7.
- [69] W. Kuo, C. Häne, P. Mukherjee, J. Malik, and E. L. Yuh, "Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 45, pp. 22737-22745, 2019.
- [70] N. Park and S. Kim, "How Do Vision Transformers Work?," *arXiv preprint arXiv:2202.06709*, 2022.
- [71] N. Brunel and D. Hansel, "How noise affects the synchronization properties of recurrent networks of inhibitory neurons," *Neural computation*, vol. 18, no. 5, pp. 1066-1110, 2006.
- [72] D. N. Louis *et al.*, "The 2021 WHO classification of tumors of the central nervous system: a summary," *Neuro-oncology*, vol. 23, no. 8, pp. 1231-1251, 2021.
- [73] P. Y. Wen and R. J. Packer, "The 2021 WHO classification of tumors of the central nervous system: clinical implications," vol. 23, ed: Oxford University Press US, 2021, pp. 1215-1217.
- [74] P. McConville *et al.*, "Magnetic resonance imaging determination of tumor grade and early response to temozolomide in a genetically engineered mouse model of glioma," *Clinical Cancer Research*, vol. 13, no. 10, pp. 2897-2904, 2007.
- [75] A. Pierallini *et al.*, "Supratentorial diffuse astrocytic tumours: proposal of an MRI classification," *European radiology*, vol. 7, pp. 395-399, 1997.
- [76] K. Chang *et al.*, "Residual Convolutional Neural Network for the Determination of

- IDH Status in Low-and High-Grade Gliomas from MR Imaging Neural Network for Determination of IDH Status in Gliomas," *Clinical Cancer Research*, vol. 24, no. 5, pp. 1073-1081, 2018.
- [77] Y. S. Choi *et al.*, "Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics," *Neuro-oncology*, vol. 23, no. 2, pp. 304-313, 2021.
- [78] P. Chang *et al.*, "Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas," *American Journal of Neuroradiology*, vol. 39, no. 7, pp. 1201-1207, 2018.
- [79] C. Kwon, S. Park, S. Ko, and J. Ahn, "Increasing prediction accuracy of pathogenic staging by sample augmentation with a GAN," *Plos one*, vol. 16, no. 4, p. e0250458, 2021.
- [80] J. F. Cohen *et al.*, "STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration," *BMJ open*, vol. 6, no. 11, p. e012799, 2016.
- [81] D. N. Louis *et al.*, "The 2016 World Health Organization classification of tumors of the central nervous system: a summary," *Acta neuropathologica*, vol. 131, no. 6, pp. 803-820, 2016.
- [82] M. Weller *et al.*, "European Association for Neuro-Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas," *The lancet oncology*, vol. 18, no. 6, pp. e315-e329, 2017.
- [83] J. Ashburner and K. J. Friston, "Unified segmentation," *Neuroimage*, vol. 26, no. 3, pp. 839-51, Jul 1 2005, doi: 10.1016/j.neuroimage.2005.02.018.
- [84] F. Isensee *et al.*, "Automated brain extraction of multisequence MRI using artificial neural networks," *Human brain mapping*, vol. 40, no. 17, pp. 4952-4964, 2019.
- [85] N. C. Institute. "Wiki for the VASARI feature set. <https://wiki.nci.nih.gov/display/CIP/VASARI> ." (accessed.
- [86] Y. K. Nam *et al.*, "Reproducible imaging-based prediction of molecular subtype and risk stratification of gliomas across different experience levels using a structured reporting system," *European Radiology*, vol. 31, no. 10, pp. 7374-7385, 2021.
- [87] A. Saco, J. A. Bombi, A. Garcia, J. Ramírez, and J. Ordi, "Current status of whole-

- slide imaging in education," *Pathobiology*, vol. 83, no. 2-3, pp. 79-88, 2016.
- [88] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," *Annual Review of Pathology: Mechanisms of Disease*, vol. 8, pp. 331-359, 2013.
- [89] T. C. Cornish, R. E. Swapp, and K. J. Kaplan, "Whole-slide imaging: routine pathologic diagnosis," *Advances in anatomic pathology*, vol. 19, no. 3, pp. 152-159, 2012.
- [90] T. W. Bauer, L. Schoenfield, R. J. Slaw, L. Yerian, Z. Sun, and W. H. Henricks, "Validation of whole slide imaging for primary diagnosis in surgical pathology," *Archives of pathology & laboratory medicine*, vol. 137, no. 4, pp. 518-524, 2013.
- [91] L. Pantanowitz *et al.*, "Review of the current state of whole slide imaging in pathology," *Journal of pathology informatics*, vol. 2, no. 1, p. 36, 2011.
- [92] J. Barker, A. Hoogi, A. Depeursinge, and D. L. Rubin, "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles," *Medical image analysis*, vol. 30, pp. 60-71, 2016.
- [93] 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*, 2015: Springer, pp. 234-241.
- [94] A. Janowczyk, A. Basavanahally, and A. Madabhushi, "Stain normalization using sparse autoencoders (StaNOSA): application to digital pathology," *Computerized Medical Imaging and Graphics*, vol. 57, pp. 50-61, 2017.
- [95] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. van der Laak, and P. H. de With, "Stain normalization of histopathology images using generative adversarial networks," in *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, 2018: IEEE, pp. 573-577.
- [96] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE international symposium on biomedical imaging: from nano to macro*, 2009: IEEE, pp. 1107-1110.
- [97] A. Vahadane *et al.*, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35,

- no. 8, pp. 1962-1971, 2016.
- [98] T. de Bel, M. Hermsen, J. Kers, J. van der Laak, and G. Litjens, "Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology," 2019.
- [99] J. Jeong, K. D. Kim, Y. Nam, C. E. Cho, H. Go, and N. Kim, "Stain normalization using score-based diffusion model through stain separation and overlapped moving window patch strategies," *Computers in Biology and Medicine*, vol. 152, p. 106335, 2023.
- [100] B. E. Bejnordi *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199-2210, 2017.
- [101] Y. J. Kim *et al.*, "PAIP 2019: Liver cancer segmentation challenge," *Medical Image Analysis*, vol. 67, p. 101854, 2021.
- [102] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology*, vol. 23, no. 4, pp. 291-299, 2001.
- [103] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, 1999.
- [104] C. Saharia *et al.*, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1-10.
- [105] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81-84, 2002.
- [106] D. Renza, E. Martinez, and A. Arquero, "A new approach to change detection in multispectral images by means of ERGAS index," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 1, pp. 76-80, 2012.
- [107] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2003, vol. 2: Ieee, pp. 1398-1402.
- [108] P. Salehi and A. Chalechale, "Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis," in *2020 International*

- Conference on Machine Vision and Image Processing (MVIP)*, 2020: IEEE, pp. 1-7.
- [109] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.
- [110] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.
- [111] H. Nam and H.-E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [112] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *arXiv preprint arXiv:1907.10830*, 2019.
- [113] A. Volokitin *et al.*, "Modelling the distribution of 3D brain MRI using a 2D slice VAE," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020: Springer, pp. 657-666.
- [114] S. Pieper, M. Halle, and R. Kikinis, "3D Slicer," in *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*, 2004: IEEE, pp. 632-635.
- [115] S. M. Rivera, J. P. Hatch, C. Dolce, R. A. Bays, J. E. Van Sickels, and J. D. Rugh, "Patients' own reasons and patient-perceived recommendations for orthognathic surgery," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 118, no. 2, pp. 134-140, 2000.
- [116] V. Narayanan, S. Guhan, K. Sreekumar, and A. Ramadorai, "Self-assessment of facial form oral function and psychosocial function before and after orthognathic surgery: a retrospective study," *Indian Journal of Dental Research*, vol. 19, no. 1, p. 12, 2008.
- [117] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349-3364, 2020.
- [118] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

## 국문 요약

생성 모델은 의료 영상 분야에서 매우 유용하게 사용될 수 있습니다. 이러한 모델은 데이터 불균형 문제를 해결하거나 다른 모달리티로 변환하는데 사용될 수 있습니다. 또한, 3D 생성은 임상 연구, 분포 분석, 이상 감지 등에 적용할 수 있습니다. 그러나 의료 영상은 자연 이미지보다 복잡하기 때문에 생성이 어렵습니다. 따라서 의료적으로 타당한 이미지를 생성하기 위해서는 많은 노력이 필요하며, 생성 모델이 의료 영상 분야에서 뛰어날 수 있는 것은 어렵습니다. 그러나 최근 확산 모델의 발전으로 고품질 이미지를 생성하는 것이 가능해졌고, 잠재적 확산 모델의 사용으로 생성 속도 문제도 해결되었습니다. 따라서 본 논문은 확산 모델을 사용한 생성, 생성을 통한 데이터 보강, 이미지 간 변환, 3D 생성 및 예측 생성에 대한 실험을 제안합니다. 이 연구의 결과는 의료 전문가를 위한 더 정확하고 종합적인 진단 도구를 제공하여 의료 영상 분야에 큰 영향을 미칠 수 있습니다. 확산 모델의 사용은 의료 영상 생성에 필요한 시간과 노력을 줄이고 의료 이미지의 전반적인 품질을 향상시켜 환자의 치료 결과를 개선할 수 있습니다. 본 논문은 의료 영상 진단 분야에 확률 기반 확산 모델의 기술적 구현과 임상 응용에 대한 포괄적인 개요를 제공하며, 의료 영상 진단 분야를 혁신할 잠재력을 강조합니다.