# Modification of DDIM Encoding for Generating Counterfactual Pathology Images of Malignant Lymphoma

Ryoichi Koga[1], Mauricio Kugler[1], Tatsuya Yokota[1], Kouichi Ohshima[2,3], Hiroaki Miyoshi[2,3], Miharu Nagaishi[2], Noriaki Hashimoto[4], Ichiro Takeuchi[4,5] and Hidekata Hontani[1]

[1]*Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, Aichi, 466-8555, Japan*

[2]*Kurume University Department of Pathology, 67 Asahi-cho, Kurume-shi, Fukuoka, 830-0011, Japan*

[3]*The Japanese Society of Pathology, 1-2-5 Yushima, Bunkyo-ku, Tokyo, 113-0034, Japan*

[4]*RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan*

[5]*Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601, Japan*

Keywords: Counterfactual Images, Diffusion Models, Pathological Images, Malignant Lymphoma, Causal Inference.

Abstract: We propose a method that modifies encoding in DDIM (Denoising Diffusion Implicit Model) to improve the quality of counterfactual histopathological images of malignant lymphoma. Counterfactual medical images are widely employed for analyzing the changes in images accompanying disease. For the analysis of pathological images, it is desired to accurately represent the types of individual cells in the tissue. We employ DDIM because it can refer to exogenous variables in causal models and can generate counterfactual images. Here, one problem of DDIM is that it does not always generate accurate images due to approximations in the forward process. In this paper, we propose a method that reduces the errors in the encoded images obtained in the forward process. Since the computation in the backward process of DDIM does not include any approximation, the accurate encoding in the forward process can improve the accuracy of the image generation. Our proposed method improves the accuracy of encoding by explicitly referring to the given original image. Experiments demonstrate that our proposed method accurately reconstructs original images, including microstructures such as cell nuclei, and outperforms the conventional DDIM in several measures of image generation.

## 1 INTRODUCTION

Malignant lymphoma has more than 70 subtypes, and pathologists identify the subtype from a set of tissue slides of a specimen that is invasively extracted from a patient (Swerdlow SH et al., 2017). Some examples of tissue microscopic images of malignant lymphoma are shown in Fig.1. The top panel shows images of a non-cancerous tissue and the bottom panel images of a cancerous tissue. In the weakly magnified image (a-2) of Fig.1, a circular structure can be observed. This is a cross-section of a spherical microtissue structure called the follicle. On the other hand, the follicle cannot be observed in (b-2) because the degree of cell differentiation decreases in cancerous tissues and the structure of the follicle collapses. In the strong magnified image (a-3) of Fig.1, a greater variety of cells are observed in the non-cancerous tissue than in (b-3) of cancerous tissue. Non-cancerous tissues are composed of a wide variety of cells that differ from each other in the morphology and texture of their cell nuclei than the cancerous tissues. In the cancerous tissues, the ratio of self-replicated cancer cells increases, and the diversity of cell types constituting the tissues tends to decrease. Changes in the tissue structure in cancerous tissues can be observed both in the global tissue structures and in the local cell structures.

Pathologists identify the subtypes by observing the morphology of tissue and cell structures. Currently, the diagnosis is largely qualitative based on the pathologists' experience and intuition. This makes it difficult for pathologists to explain the basis for their diagnosis, and there is room for improvement in diagnostic reproducibility. To achieve the improvement, it is desired to quantitatively evaluate the morphology of tissue and cell structures. To construct quantitative



Koga, R., Kugler, M., Yokota, T., Ohshima, K., Miyoshi, H., Nagaishi, M., Hashimoto, N., Takeuchi, I. and Hontani, H.
Modification of DDIM Encoding for Generating Counterfactual Pathology Images of Malignant Lymphoma.
DOI: 10.5220/0012366100003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 519-527
ISBN: 978-989-758-679-8; ISSN: 2184-4321

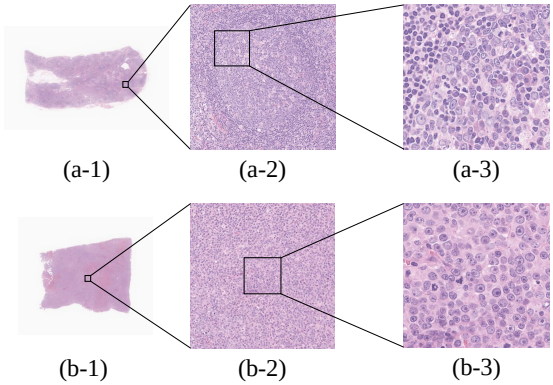



Figure 1: The examples of pathology images. In this figure, (a-1) and (b-1) are a non-cancerous tissue image and a cancerous one, respectively. (a-2) and (b-2) are weakly magnified images. (a-3) and (b-3) are strongly magnified images.

criteria for the changes in the morphology of these structures, we employ an approach of constructing a subtype classifier and then approximating the discriminant function by an explainable function at post-hoc. For example, decision trees are used to approximate the discriminant function of a neural network-based classifier for improving the interpretability of the classifier and constructing a quantitative criterion useful for the classification (Singla et al., 2021). In such approaches, counterfactual images are used to select image features that are interpretable and useful for classification. In this paper, we propose a method that generates counterfactual pathology images of malignant lymphoma.

A counterfactual image is a hypothetical image obtained when one factor changes in the causal model of a given image. A causal model consists of endogenous variables and exogenous ones. The causal model represents the causal relationships between factors represented by the endogenous variables, of which values are observable. The exogenous variables represent unobservable stochastic factors that are not affected by other ones. Here, we consider a simple causal model with only two endogenous variables: One represents the subtype and the other represents the pathological image. Fig.2 shows the causal model considered in this study. In this causal model, the pathological image $\boldsymbol{x}^{(2)}$ is modeled with the corresponding exogenous variable $\boldsymbol{u}^{(2)}$ and the subtype $\boldsymbol{x}^{(1)}$ as follows:

$$\boldsymbol{x}^{(2)} = f(\boldsymbol{x}^{(1)}, \boldsymbol{u}^{(2)}), \tag{1}$$

where the image $\boldsymbol{x}^{(2)}$ is deterministically computed by the function $f$ from $\boldsymbol{x}^{(1)}$ and $\boldsymbol{u}^{(2)}$. The counterfactual images generated in this study are the images obtained when only the endogenous variable $\boldsymbol{x}^{(1)}$ representing the subtype changes and the exogenous variable $\boldsymbol{u}^{(2)}$ is fixed. Counterfactual images are obtained by the disentanglement of tissue morphological features specific to the subtype difference from other features representing individual differences.

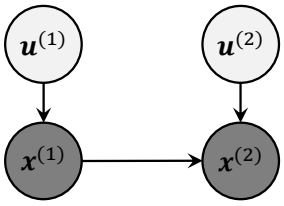



Figure 2: The causal model considered in this study. In this figure, $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ are endogenous variables, indicating the subtype and the pathology image, respectively. $\boldsymbol{u}^{(1)}$ and $\boldsymbol{u}^{(2)}$ are exogenous variables corresponding to $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, respectively.

Several methods for generating counterfactual images have been proposed (Singla et al., 2020), (Sanchez and Tsaftaris, 2022). In this study, we employ a method that uses a diffusion model. A diffusion model is one of the most popular generative models and is capable of generating higher-quality data than other methods. In addition, the generation of counterfactual images using denoising diffusion implicit models (DDIMs) (Song et al., 2021), which was proposed to alleviate the problem of high computational cost of denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020), one of the most popular diffusion models, is easier to interpret based on a causal model than other methods. In DDIM, the forward process to the noise image is deterministic, and the image obtained with the backward process is determined only by the initial noise image. This deterministic property is consistent with the causal model in Eq.(1) and the obtained noise image can be employed as a representation of the exogenous variables (Sanchez and Tsaftaris, 2022). We employ DDIM and generate images of the different subtype corresponding to the same exogenous variable from the noise image by guiding on the different subtype.

When a pathological microscopic image is first encoded into a noise image using DDIM and then the noise image is restored to the original image by the backward process, the details of the restored image may not match the original image. Fig.3 shows examples of the original image and the corresponding image reconstructed by a conventional DDIM. As shown in Fig.3, some cell nuclei are reconstructed with a different shape from those of the original image. This is because the computation of the forward process in DDIM includes some approximations, which degrades the accuracy of the encoding. As mentioned above, when one has malignant lymphoma, tissue

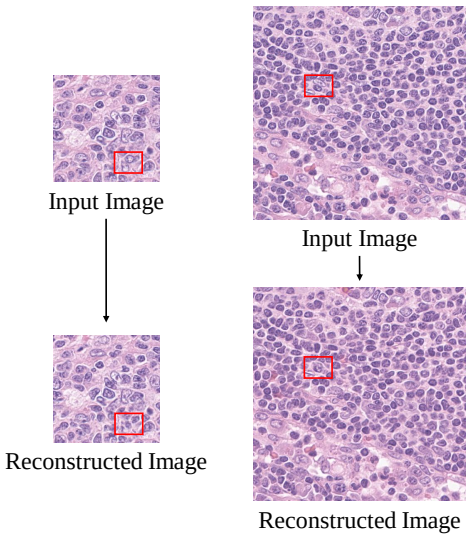



Figure 3: The result of reconstructed images using the conventional DDIM. The conventional DDIM fails to accurately reconstruct the input image in the region highlighted with a red rectangle.

structure changes not only in its global structure but also in the microcell structures. To quantitatively evaluate the changes in tissue structures, it is desired to be able to accurately reconstruct the microstructures such as individual cell nuclei.

In this study, we propose a method that removes errors added to the series of diffused images in the forward process of DDIM by referring to the original image. Since the computation of the backward process of DDIM does not include any approximation, making the encoding in the forward process accurate can improve the accuracy of the reconstruction. By utilizing the same noise estimator used in the backward process of a conventional DDIM, our method can accurately reconstruct the original image.

Our main contributions are as follows: (1) To improve the accuracy of encoding in the forward process of DDIM, we propose a method that determines optimal modification vectors to obtain a better noise image that accurately reconstructs the original input image, and (2) We evaluate the effectiveness of modified DDIM encoding and the quality of generated counterfactual images visually and quantitatively.

# 2 GENERATION OF COUNTERFACTUAL IMAGES USING DIFFUSION MODELS

In this section, we first describe DDPMs. Thereafter, we introduce a denoising diffusion implicit model (DDIM) that can deterministically encode an input image. Then we describe the generation of counterfactual images with the classifier-guidance.

## 2.1 Denoising Diffusion Probabilistic Models

Diffusion models are latent variable models of the form $p_\theta(\boldsymbol{x}_0) := \int p_\theta(\boldsymbol{x}_{0:T})d\boldsymbol{x}_{1:T}$, where $\boldsymbol{x}_0$ is an observed variable and $\boldsymbol{x}_1,...,\boldsymbol{x}_T$ are latent representation and indices of $\boldsymbol{x}$ are timesteps of forward process. The observed variable $\boldsymbol{x}_0$ follows the data distribution $q(\boldsymbol{x}_0)$ and the latent variables $\boldsymbol{x}_1,...,\boldsymbol{x}_T$ are the same dimensions as the observed variable $\boldsymbol{x}_0$. The joint distribution $p_\theta(\boldsymbol{x}_{0:T})$ is defined as the following equations:

$$p_\theta(\boldsymbol{x}_{0:T}) := p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \quad (2)$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t),\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t,t)), \quad (3)$$

where $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T;\mathbf{0},\mathbf{I})$ and $\theta$ is model parameters. Sampling from the distribution $p_\theta(\boldsymbol{x}_0)$ that is parametrized with $\theta$, we can compute the backward process of the diffusion model. For the forward process or diffusion process in DDPM, the posterior $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ comes from Markovian process that gradually adds Gaussian noise to the data according to noise schedulers $\beta_1,...,\beta_T$:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \quad (4)$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t;\sqrt{1-\beta_t}\boldsymbol{x}_{t-1},\beta_t\mathbf{I}). \quad (5)$$

When the distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ of Eq.(5) is Gaussian distribution, if $\beta_t$ is small, then the distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ of Eq.(3) is also Gaussian distribution (Sohl-Dickstein et al., 2015). Fig.4 illustrates the directed graphical model based on Eq.(3) and Eq.(5). This forward process have a notable property that admits sampling $\boldsymbol{x}_t$ at an arbitrary timestep $t$ in closed form:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\mathbf{I}), \quad (6)$$

where $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s$. Training of the DDPM is performed by optimizing the usual variational inference bound on negative log likelihood. Consequently, as described in (Ho et al., 2020), the objective function of DDPM is expressed as:

$$\min_\theta \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\varepsilon}}\|\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon},t)\|_2^2, \quad (7)$$

where $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ and $\boldsymbol{\varepsilon}_\theta$ is a function that predicts $\boldsymbol{\varepsilon}$ from $\boldsymbol{x}_t$ and $t$.

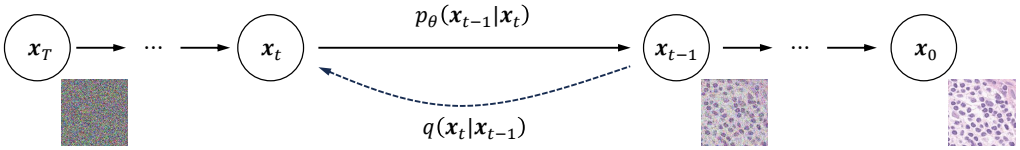



Figure 4: The directed graphical model considered in diffusion models.

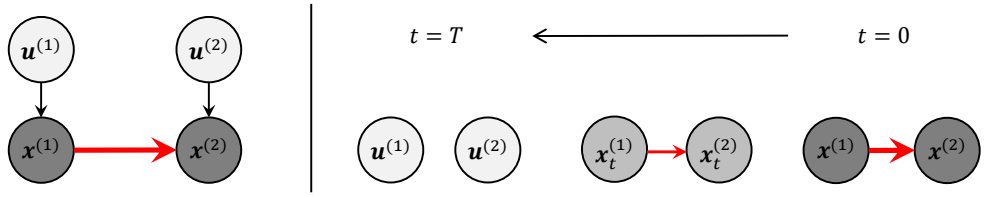



Figure 5: Illustration of the forward process of diffusion model as weakening of causal relationship considered in this study. Arrows in this figure indicate the causal relationships between variables and direction, and the thickness of red arrows express strength of the relation.

After training of the DDPM, a sample $\boldsymbol{x}_0$ is produced by repeating the sampling of $\boldsymbol{x}_{t-1} \sim p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ with $t = T, \ldots, 1$. The sampling of $\boldsymbol{x}_{t-1} \sim p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ can be realized by computing Eq.(8) as described in (Ho et al., 2020):

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t \boldsymbol{z}, \quad (8)$$

where $\sigma_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta t$ and $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since the DDPM is constructed under the small noise schedulers $\beta_t$ and the large timestep $T$, such as $T = 1{,}000$, it is known that the generation of samples with the DDPM takes much time.

## 2.2 Denoising Diffusion Implicit Models

In the DDPM, iterative noise addition in the forward process is formulated as Markovian process and an original image is encoded into a series of noise images. In the backward process, the estimation and removal of the noise must be repeated the same number of times as the number of the noise addition, which is computationally inefficient. DDIM can reduce the number of times to estimate and remove noise components in the backward process compared to the forward process (Song et al., 2021). This efficiency improvement is achieved by making the forward process non-Markovian while using the same objective function of DDPMs (Eq.(7)). The update equation in the backward process of DDIM is derived so that the marginal distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ at a timestep $t$ in the forward process matches that in the forward process of the DDPM, and is expressed as:

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1} - \{\hat{\sigma}_t(\eta)\}^2}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) + \hat{\sigma}_t(\eta)\boldsymbol{z}, \quad (9)$$

where

$$\hat{\sigma}_t(\eta) := \eta\sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}. \quad (10)$$

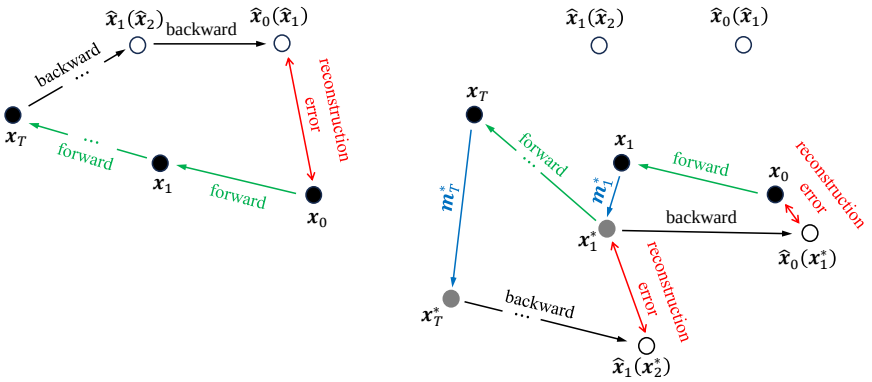
On the other hand, the forward process of DDIM is derived from Bayes' rule using Eq.(9). When $\eta = 1$ for all $t$, Eq.(9) reduces to Eq.(8). When $\eta = 0$ for all $t$, the coefficient of the random noise $\boldsymbol{z}$ in Eq.(9) becomes zero, and a sample is deterministically produced. When $\eta > 0$ at least one $t$, random noise $z$ in the Eq.(9) is added during sampling, and a sample stochastically produced.

DDIMs are utilized in order not only to accelerate the backward process but also to encode an input image $\boldsymbol{x}_0$. The authors of (Song et al., 2021) demonstrate that the original input image can be efficiently reconstructed from the corresponding final noise image, $\boldsymbol{x}_T$, encoded using the DDIM.

## 2.3 Generation with Classifier-Guidance

In our study, we generate counterfactual images using classifier-guidance. In the classifier-guidance, the backward process of the trained diffusion model is conditioned with a gradient of the classifier (Dhariwal and Nichol, 2021). The classifier $p_\phi(y|\boldsymbol{x}_t, t)$ is trained from noise images $\boldsymbol{x}_t$, where $\phi$ is the classifier's parameters and $y$ is a class label. After training of the classifier, we generate counterfactual images from an encoded representation by guiding the backward process of diffusion models based on the gradient $\nabla_{\boldsymbol{x}_t} p_\phi(y|\boldsymbol{x}_t, t)$.

Given a causal model, counterfactual images are generated by changing only the endogenous variable of interest under deleting the directed edges toward




(a) Conventional DDIM encoding (b) Modified DDIM encoding (Proposed)

Figure 6: The comparison of conventional DDIM encoding and modified DDIM encoding. The left panel illustrates the conventional DDIM encoding and the right panel illustrates our proposed modified DDIM encoding. Green arrows show the forward process using Eq.(14). Black arrows show the backward process using Eq.(11). Red arrows indicate the reconstruction error. Blue arrows show the modification vector.

the endogenous variable of interest and fixing all other variables except for that variable. According to (Sanchez and Tsaftaris, 2022), the forward process of the diffusion model weakens the causal relationships between variables, as illustrated in Fig.5, where $\boldsymbol{x}_t^{(k)}$ are $k$-th endogenous variables and $\boldsymbol{u}^{(k)}$ are respective exogenous variables. In this study, $\boldsymbol{x}^{(1)}$ denotes the subtype of malignant lymphoma and $\boldsymbol{x}^{(2)}$ denotes the pathological image. In the right panel of Fig.5, the forward process weakens the relationships between endogenous variables until these variables are completely independent at $t = T$. By computing the forward process of DDIM until $t = T$, the exogenous variables $\boldsymbol{u}^{(2)}$ of pathological image $\boldsymbol{x}^{(2)}$ can be inferred deterministically.

## 3 PROPOSED METHOD

In the generation of counterfactual images, it is desired that we can uniquely reconstruct the original images from the exogenous variables. This is one of the main reasons that we employ the DDIM. As mentioned in Sec.2.3, in counterfactual image generation using the DDIM, the noise image, $\boldsymbol{x}_T$, obtained by encoding the given image with the forward process is considered as an exogenous variable. For this reason, high accuracy is desired in the computation of the forward process. The computation of the forward process in the DDIM includes approximations, and there is room for the improvement of accuracy. The reason including the approximation is shown below. The forward process that computes $\boldsymbol{x}_t$ from $\boldsymbol{x}_{t-1}$ is obtained from Eq.(9). At first, Eq.(9) is rewritten as:

$$\boldsymbol{x}_{t-1} = \frac{1}{a_t}\boldsymbol{x}_t - \frac{b_t}{a_t}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t), \tag{11}$$

where $\eta = 0$ in Eq.(9) and

$$a_t = \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_{t-1}}}, \tag{12}$$

$$b_t = \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t} - \sqrt{\bar{\alpha}_t}\sqrt{1-\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_{t-1}}}. \tag{13}$$

By solving Eq.(11) for $\boldsymbol{x}_t$ under the assumption of $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) \approx \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_{t-1}, t)$ (Song et al., 2021), we obtain the equation that is used in a conventional DDIM:

$$\boldsymbol{x}_t \approx a_t \boldsymbol{x}_{t-1} + b_t \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_{t-1}, t). \tag{14}$$

Here, it should be noted that the approximation of $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) \approx \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_{t-1}, t)$ causes an encoded error for each $\boldsymbol{x}_t (t = 1, 2, ..., T)$ in the forward process. When one reconstructs the sample $\boldsymbol{x}_{t-1}$ from the $\boldsymbol{x}_t$ that includes noise, the reconstructed sample would have a non-negligible reconstruction error. This error would be added at each timestep in the backward process and it is known that the propagation of the error leads to incorrect image reconstruction (Wallace et al., 2023). This inaccuracy should be corrected for the generation of counterfactual pathology images. We propose a method that corrects the inaccuracy of the conventional DDIM.

### 3.1 Modified DDIM Encoding

Our proposed method modifies the series of the noise images, $\boldsymbol{x}_1, ..., \boldsymbol{x}_T$, obtained in the forward process so that the backward process accurately reconstructs the

Table 1: The settings for training the models. The diffusion models and the classifiers are trained with two image sizes. In this table, "DDPM" refers to the diffusion model, and "CLS" refers to the classifier.

| | $256 \times 256$ (DDPM) | $256 \times 256$ (CLS) | $512 \times 512$ (DDPM) | $512 \times 512$ (CLS) |
|---|---|---|---|---|
| Batch size | 16 | 128 | 4 | 32 |
| Epoch | 100 | 80 | 100 | 80 |
| Timesteps | $T = 1,000$ | $T = 1,000$ | $T = 2,000$ | $T = 2,000$ |

original image. The modification method is shown below. Fig.6 illustrates the comparison of conventional DDIM encoding and our proposed modified DDIM encoding. Let $\boldsymbol{x}_t$ denote a sample obtained by applying DDIM encoding to the sample $\boldsymbol{x}_{t-1}$. Let $\hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t)$ denote a sample reconstructed from the sample $\boldsymbol{x}_t$ using Eq.(11). The reconstructed sample $\hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t)$ would have the reconstruction error and the error strength at each timestep $t$ is evaluated as:

$$E_t := \|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t)\|^2. \tag{15}$$

This error comes from the inaccuracy of the encoding due to the approximation in the forward process. To reduce this error, we introduce a modification vector $\boldsymbol{m}_t$ for the compensation of reconstructed error as shown in Fig.6, that is, $\hat{\boldsymbol{x}}_{t-1}$ is not reconstructed from $\boldsymbol{x}_t$ but from $(\boldsymbol{x}_t + \boldsymbol{m}_t)$. This compensation by $\boldsymbol{m}_t$ makes the series of encodes, $\boldsymbol{x}_1, ..., \boldsymbol{x}_T$, more consistent with the theoretical non-Markovian forward process.

Let $\hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t + \boldsymbol{m}_t)$ denote a reconstructed sample from $(\boldsymbol{x}_t + \boldsymbol{m}_t)$ using Eq.(11). By adding a modification vector $\boldsymbol{m}_t$, the reconstruction error of Eq.(15) can be written as:

$$E_t(\boldsymbol{m}_t) = \|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t + \boldsymbol{m}_t)\|^2. \tag{16}$$

The objective here is to reduce the errors included in each $\boldsymbol{x}_t$ by inferring $\boldsymbol{m}_t$ for $t = 1, ..., T$. We start the inference of $\boldsymbol{m}_t$ from $t = 1$: We compute the optimal $\boldsymbol{m}_1^*$ by solving the optimization problem:

$$\boldsymbol{m}_1^* := \underset{\boldsymbol{m}_1}{\operatorname{argmin}} \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0(\boldsymbol{x}_1 + \boldsymbol{m}_1)\|^2, \tag{17}$$

where $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_1 + \boldsymbol{m}_1)$ is obtained by applying the backward process of the conventional DDIM. Once we obtain $\boldsymbol{m}_1^*$ that minimizes the reconstruction error (Eq.(17)), we update $\boldsymbol{x}_1$ as $\boldsymbol{x}_1^* = \boldsymbol{x}_1 + \boldsymbol{m}_1^*$ and apply the forward process of the conventional DDIM to obtain $\boldsymbol{x}_2$ from $\boldsymbol{x}_1^*$. Then, the $\boldsymbol{m}_2$ is obtained by minimizing $\|\boldsymbol{x}_1^* - \hat{\boldsymbol{x}}_1(\boldsymbol{x}_2 + \boldsymbol{m}_2)\|^2$. Incrementing $t$ from 1 to $T$, we estimate $\boldsymbol{m}_t^*$ for $t = 1, ..., T$ by minimizing the reconstruction error (Eq.(16)) at each timestep and obtain the series of encoded images, $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_T^*$.

The proposed method is summarized in Algorithm 1. Procedure FORWARD$(\cdot)$ refers to applying the forward process of the conventional DDIM and BACKWARD$(\cdot)$ refers to applying the backward process of the conventional DDIM. To determine the modification vectors, we utilize the trained diffusion model used in the conventional method and require no retraining of the diffusion model. We use the modified DDIM encoding denoted above to obtain the series of the noise images that can accurately reconstruct the input image $\boldsymbol{x}_0$.

**Data:** a given original image $\boldsymbol{x}_0$
**Result:** a series of modified noise images, $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_T^*$
$\boldsymbol{x}_0^* = \boldsymbol{x}_0$
**for** $t = 1, 2..., T$ **do**
  $\boldsymbol{x}_t$ = FORWARD$(\boldsymbol{x}_{t-1}^*)$
  $\boldsymbol{m}_t \leftarrow \boldsymbol{0}$
  $\hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t + \boldsymbol{m}_t)$ = BACKWARD$(\boldsymbol{x}_t + \boldsymbol{m}_t)$
  $\boldsymbol{m}_t^* = \underset{\boldsymbol{m}_t}{\operatorname{argmin}} \|\boldsymbol{x}_{t-1}^* - \hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t + \boldsymbol{m}_t)\|^2$
  $\boldsymbol{x}_t^* = \boldsymbol{x}_t + \boldsymbol{m}_t^*$
**end**

Algorithm 1: Modified DDIM encoding.

# 4 EXPERIMENTAL RESULTS

In this section, we first describe the training of diffusion models and classifiers for guidance. Thereafter, we demonstrate the performance of the modified DDIM encoding. Finally, we illustrate the result of generating counterfactual images.

## 4.1 Training of DDPMs and Classifiers

Our database for the experiments in this paper comprises the WSIs of 10 reactive cases, non-cancerous, and 10 DLBCL cases, one of the subtypes. DDPMs and classifiers for guidance are trained with the settings shown in Table 1 using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate $7.0 \times 10^{-4}$ from 128,000 patch images cropped in two type sizes, $256 \times 256$ and $512 \times 512$, from the WSIs.

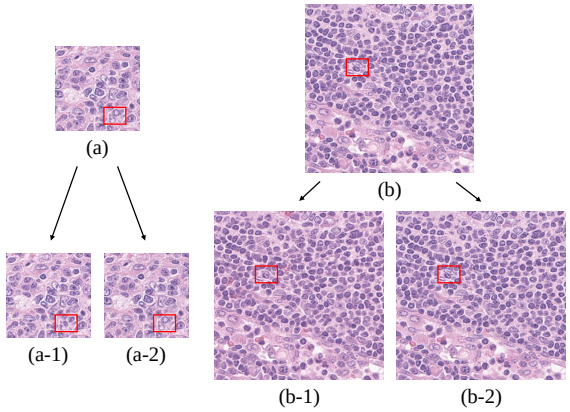



Figure 7: Visual comparison of the conventional method and the proposed one. The input images of (a) and (b) are the size of 256 × 256 and 512 × 512, respectively. (a-1) and (b-1) are the reconstructed images based on the conventional DDIM encoding. (a-2) and (b-2) are the reconstructed images based on the modified DDIM encoding.

Table 2: Quantitative comparison of the conventional and proposed methods. For each method, the reconstruction error between the input image $\boldsymbol{x}_0$ and reconstructed image $\hat{\boldsymbol{x}}_0$ is evaluated with the $l_1$ distance. The best result is marked in bold.

| Patch size | 256 × 256 | 512 × 512 |
|---|---|---|
| Conventional DDIM encoding | $0.025 \pm 0.012$ | $0.021 \pm 0.004$ |
| Modified DDIM encoding | $\mathbf{0.009 \pm 0.010}$ | $\mathbf{0.006 \pm 0.004}$ |

## 4.2 Performance of Modified DDIM Encoding

We evaluate the effect of introducing modification vectors in DDIM encoding. For the models constructed with two type patch sizes, the images that are reconstructed based on the conventional DDIM encoding and the modified DDIM encoding are shown in Fig.7 and Table 2. The number of iterative processes required to solve the optimization problem of $\boldsymbol{m}_t$ at each timestep is set to 10. Evidently from Fig.7, whereas the conventional method fails to accurately reconstruct the input images, our proposed method is successful in accurately reconstructing the input images. Specifically, our proposed method accurately reconstructs the input image in the region highlighted by the red rectangle in Fig.7. This visual evaluation is consistent with the results of quantitative evaluation, as shown in Table 2. This result demonstrates that our proposed method reduces the approximation error derived from the conventional DDIM encoding to obtain the series of noise images that can accurately reconstruct the input image.

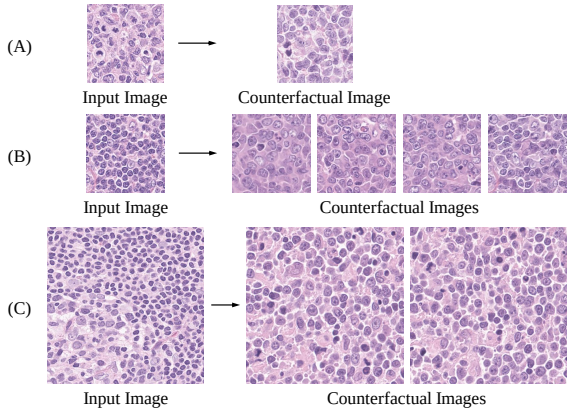



Figure 8: Result of generating counterfactual images. A row of (A) is the result with the existing method based on the cGAN. Rows of (B) and (C) are the results with the DDPM using the classifier-guidance.

## 4.3 Generation of Counterfactual Images

From the intermediate representation obtained by DDIM encoding, we generate counterfactual images when the patient changes from reactive to DLBCL. In image generation, $\eta$ in Eq.(10) is set to 0.5 and a guidance scale for the classifier-guidance is set to 20. The generated counterfactual images are shown in Fig.8. The row of (A) in Fig.8 includes the counterfactual images generated using the existing method based on the conditional GAN (Singla et al., 2020). This method deterministically generates a single counterfactual image from a given input image and cannot stochastically generate many counterfactual images. In addition, unfortunately, pathologists have commented that if the counterfactual image of a row of (A) was generated as DLBCL, the cell nuclei were too dense to be real. By contrast, counterfactual images in rows of (B) and (C) are stochastically generated from a given input image and are good in terms of the ability to render microstructures such as nucleoli. Moreover, whereas the cGAN-based method has failed to learn the counterfactual image generator of 512 × 512, the diffusion-based method is successful in generating images of 512 × 512.

We quantitatively evaluate the quality of generated counterfactual images. FID scores are known as indicators evaluating the quality of images generated using generative models (Heusel et al., 2017). Table 3 shows the computed FID scores. The cGAN-based method failed to learn the counterfactual image generator of 512 × 512. The diffusion-based method demonstrates better performance than the GAN-based one and this assessment is consistent with the visual evaluation in Fig.8.

Table 3: The FID scores of generated images. The best result is marked in bold.

| | 256 × 256 (cGAN) | 256 × 256 (DDPM) | 512 × 512 (DDPM) |
|---|---|---|---|
| FID scores (↓) | 15.061 | **14.304** | 7.264 |

Table 4: The quantitative evaluation of generated counterfactual images. Lower values for the composition and the reversibility measured with the $l_1$ distance indicate higher performance. Higher values for the effectiveness measured with the accuracy of the classifier indicate higher performance. The best result is marked in bold.

| | 256 × 256 (cGAN) | 256 × 256 (DDPM) | 512 × 512 (DDPM) |
|---|---|---|---|
| Composition (↓) | 0.209 | **0.049** | 0.041 |
| Reversibility (↓) | 0.215 | **0.087** | 0.069 |
| Effectiveness (↑) | 0.678 | **0.965** | 0.677 |

Furthermore, we evaluate the quality of generated images in terms of counterfactuals. The authors of (Monteiro et al., 2023) provide three indicators based on Pearl's axiomatic definition (Pearl, 2009) to evaluate the quality of counterfactual images; these indicators are composed of composition, reversibility, and effectiveness. Briefly, the composition implies that the generated image $\hat{\boldsymbol{x}}_0$ is consistent with the input image $\boldsymbol{x}_0$ under the case without any intervention, and this is often measured with the $l_1$ distance. The reversibility implies cycle-consistency in a cycle-backed transformation from the generated counterfactual image to the original input image, and this is also often measured with the $l_1$ distance. The effectiveness implies the effect of intervention on the generation of counterfactual images. For instance, when the generated counterfactual image is fed into a different subtype classifier from the classifier constructed for the classifier-guidance, its effectiveness is computed as whether the classifier can accurately classify it into the class specified in the generation of the counterfactual image.

We evaluate the quality of counterfactual images based on the three indicators. Fig.9 shows the reconstructed images to visually evaluate the composition and the reversibility of these indicators. Evidently from Fig.9, we can see that the diffusion-based method accurately reconstructs the input image than the cGAN-based one. Moreover, these three indicators are shown in Table 4. Since the diffusion-based method is superior in all the indicators, it is expected that the diffusion-based method is a better counterfactual image generator than the GAN-based one in most cases.

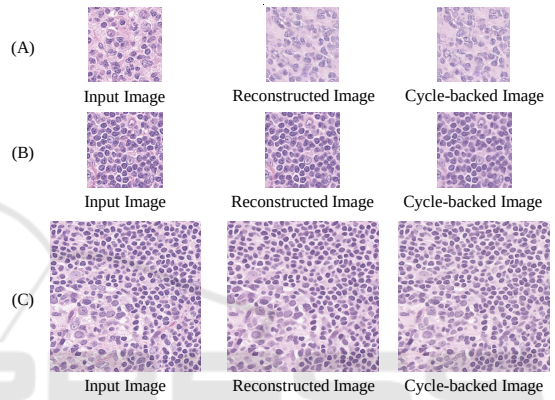



Figure 9: Result of reconstructed images and cycle-backed transformed ones. A row of (A) is the result with the existing method using the cGAN. Rows of (B) and (C) are the results with the DDPM using the classifier-guidance.

## 5 RELATED WORKS

There have been several studies that generate counterfactual images using diffusion models. The authors of (Jeanneret et al., 2022) proposed a method for generating counterfactual images using a DDPM and a perceptual loss (Johnson et al., 2016), and were successful in manipulating such as emotion and age of facial images. Since this method uses the DDPM, the original image cannot always be reconstructed from the noise image obtained with the forward process. Owing to this property, it is not easy to consider causal models for counterfactuals. Thus, we conduct the counterfactual image generation based on the DDIM encoding with the deterministic forward process, as proposed in (Sanchez and Tsaftaris, 2022).

## 6 SUMMARY AND FUTURE WORKS

In this paper, we propose a method that modifies encoding in DDIM to improve the quality of counterfactual histopathological images of malignant lymphoma. DDIM encoding is employed as an encoder for generating counterfactual images. DDIM encod-

ing generates non-negligible reconstruction error for pathological image analysis and it is not easy to obtain an intermediate representation that accurately reconstructs the original input image. To alleviate this problem, we propose a method that reduces the errors in DDIM encoding. Experimental results demonstrate that our proposed method is successful in obtaining better intermediate representations that accurately reconstruct the original input image. In addition, we generate multiple counterfactual images from the encoded representation and demonstrate that the quality of these images is good based on the visual and quantitative evaluation.

The final goal of our study is to construct quantitative criteria for the changes in the morphology of tissue structures for malignant lymphoma. To achieve this, we first generated counterfactual pathology images of DLBCL using diffusion models. Future works also include the construction of an explainable function that approximates a subtype classifier using the generated counterfactual images.

## ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP22H03613 to H.H. and JP23KJ1141 to R.K.

## REFERENCES

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Jeanneret, G., Simon, L., and Jurie, F. (2022). Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 858–876.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Monteiro, M., Ribeiro, F. D. S., Pawlowski, N., Castro, D. C., and Glocker, B. (2023). Measuring axiomatic soundness of counterfactual image models. In *The Eleventh International Conference on Learning Representations*.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Sanchez, P. and Tsaftaris, S. A. (2022). Diffusion causal models for counterfactual estimation. volume 177 of *Proceedings of Machine Learning Research*, pages 647–668. PMLR.

Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. (2020). Explanation by progressive exaggeration. In *International Conference on Learning Representations*.

Singla, S., Wallace, S., Triantafillou, S., and Batmanghelich, K. (2021). Using causal analysis for conceptual deep learning explanation. *Med Image Comput Comput Assist Interv*, 12903:pp. 519–528.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR.

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Swerdlow SH, Campo E, H. N., Jaffe ES, P. S., and Stein H, T. J., editors (2017). *World Health Organization classification of tumours of haematopoietic and lymphoid tissues. Revised 4th ed.* Lyon. IARC Press.

Wallace, B., Gokul, A., and Naik, N. (2023). Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22532–22541.