Diffusion-based generation of Histopathological Whole Slide Images at a Gigapixel scale

Robert Harb[1,2], Thomas Pock[1], Heimo Müller[2]
[1]Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[2]Diagnostic and Research Institute of Pathology, Medical University of Graz, Austria
{robert.harb, pock}@icg.tugraz.at, heimo.mueller@medunigraz.at

Abstract

We present a novel diffusion-based approach to generate synthetic histopathological Whole Slide Images (WSIs) at an unprecedented gigapixel scale. Synthetic WSIs have many potential applications: They can augment training datasets to enhance the performance of many computational pathology applications. They allow the creation of synthesized copies of datasets that can be shared without violating privacy regulations. Or they can facilitate learning representations of WSIs without requiring data annotations. Despite this variety of applications, no existing deep-learning-based method generates WSIs at their typically high resolutions. Mainly due to the high computational complexity. Therefore, we propose a novel coarse-to-fine sampling scheme to tackle image generation of high-resolution WSIs. In this scheme, we increase the resolution of an initial low-resolution image to a high-resolution WSI. Particularly, a diffusion model sequentially adds fine details to images and increases their resolution. In our experiments, we train our method with WSIs from the TCGA-BRCA dataset. Additionally to quantitative evaluations, we also performed a user study with pathologists. The study results suggest that our generated WSIs resemble the structure of real WSIs.

1. Introduction

Histopathology is the study of diseases through the inspection of tissue samples. It plays a vital role in clinical practice by providing information for accurate diagnosis. Furthermore, it is also essential in medical research for studying disease processes and contributing to developing new therapeutic strategies.

Histopathological analysis is preceded by a few preparatory steps. One first collects tissue samples, *e.g.* via biopsies, excisions, or endoscopies. Then, the samples are fixed, encased in paraffin, and thinly sliced. The resulting tissue slices are then mounted on glass slides. Followed by staining, *e.g.* using hematoxylin and eosin (H&E), to enhance the visibility of cellular components and highlight specific tissue features. After staining, slides can be scanned, resulting in high-resolution images, so-called Whole Slide Images (WSIs). Notably, a typical WSI has resolutions in the gigapixel range.

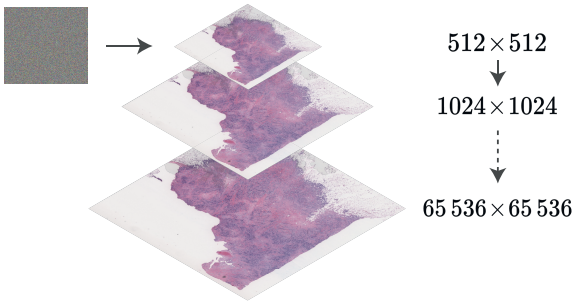

Figure 1. We sample a low-resolution image from noise using a diffusion-based generative image model. This low-resolution image is then sequentially upsampled in a coarse-to-fine scheme to generate a high-resolution Whole Slide Image.

A major challenge when developing algorithms that analyse WSIs is their high resolution. Many established methods are unsuitable since they are designed for much smaller resolutions. This also applies in the field of synthetic image generation with deep-learning. Although some methods exist, they all generated only small excerpts of WSIs, *i.e.* patches. However, such low-resolution patches contain far less detail than entire WSIs. Their high resolution offers a spectrum of detail, from a macroscopic overview of the tissue sample to fine details like individual cells at the highest magnification. Having this breadth of information is essential for many pathological applications. Consequently, to fully harness the potential of synthetic data in histopathology, generating WSIs at their full resolution is crucial.

There are many applications that could benefit from synthetic WSIs. For instance, using synthetic data to augment

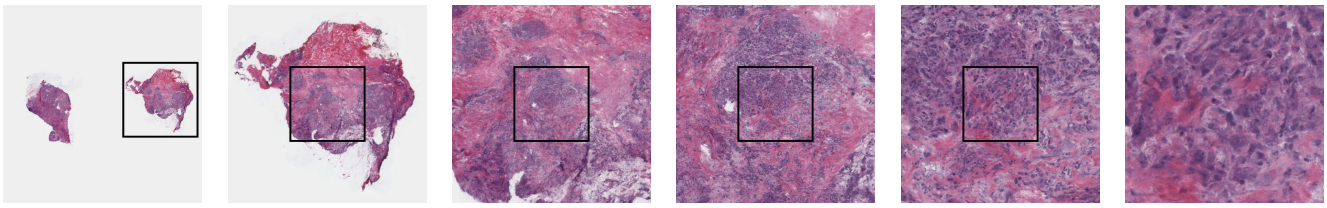

Figure 2. A synthetic WSI generated by our method with a resolution of $65\,536 \times 65\,361$ pixels. Our synthesized WSI covers the whole magnification spectrum of a WSI, starting from a macroscopic overview of tissue, down to structures at the cellular level. For visualization, we extracted patches at different magnifications, the black rectangle shows the location of the patch in the subsequent column.

datasets is common to improve the performance of deep-learning models, *e.g.*, in segmentation or classification [28]. Both of these tasks are essential in computational pathology. For example, to find new Biomarkers [47], make survival predictions [8], or for tumor segmentation [41].

Moreover, synthetic WSIs could unlock access to currently inaccessible datasets for broad audiences. Despite institutions like biobanks or hospitals collect vast amounts of human tissue samples, data protection laws often prevent publishing this data without restrictions. This limits the accessibility for research, hindering potential advancements in the field. One approach to circumvent this issue is to publish synthesised versions of real datasets [17]. Such synthesised datasets could maintain patient privacy while preserving the diagnostically relevant attributes of the original data.

Besides generating data, generative models can also be leveraged to learn data representations without requiring annotations [12, 30, 43]. This is of particular interest in histopathology. Annotating WSIs is time-consuming due to their high resolutions and can often only be done by pathologists that have the necessary domain knowledge.

Motivated by the multitude of potential applications, this work presents a novel diffusion-based method to generate synthetic WSIs. Most significantly, we generate WSIs at remarkably high resolutions up to $65\,536 \times 65\,536$ pixels. Fig. 2 shows such a high-resolution image generated by our approach.

The major challenge of our method is the computational infeasibility of training diffusion models for the high-resolution of WSIs. Instead, we are limited to a model that processes much lower-resolution images. We tackle this limitation through a novel coarse-to-fine diffusion-based sampling scheme. In this scheme, as illustrated in Fig. 1, we sample a low-resolution image and step-wise increase its resolution. Each step gradually adds finer details to an image while preserving its coarse structure. While the initial image entirely fits into our model, we do the refinement patch-wise at later steps. Even though patching limits the models' image context at later steps, the scheme has shown to be effective. This is because the coarse image structure is established in the first steps, where the context is still large. The refinement at later steps preserves this structure while gradually adding fine details that do not always require full-image context.

We describe our method in detail in Sec. 4. The main contributions of our work are as follows:

- To the best of our knowledge, we propose the first deep-learning-based method that creates synthetic histopathological WSIs at high resolutions up to $65\,536 \times 65\,536$ pixels.
- To this end, we propose a novel diffusion-based coarse-to-fine sampling scheme, where we guide the diffusion process with a relaxed super-resolution constraint.
- Even though our method involves patch-wise processing, we generate images without visible stitching artefacts. We achieve this through grid-shift, a novel technique where we interleave patching with diffusion iterations. In comparison with a related method, mask-shifting, grid-shift is computationally more efficient and simple to parallelize.
- We perform a user study with pathologists that suggests that our generated WSIs are not consistently distinguishable from real WSIs.

2. Related Work

In the following, we review related work in the areas of generating histopathological images and scaling diffusion models to high-resolutions.

Generation of Histopathology images. Several previously published methods tackle the generation of synthetic histopathological images. However, our approach stands out as the only one that is able to generate WSIs at gigapixel scale and is at the same time based on state-of-the-art generative deep-learning approaches.

A few methods were published before deep-learning-based image generation methods were widely adopted. Instead, these methods [1, 46] are based on texture-based image synthesis [9, 11, 29, 44], where the synthesis process is

based on the composition and modification of a small number of input patches. However, this approach lacks generalizability and, instead of producing diverse content, mainly replicates the features of the few provided input patches.

Contrarily to texture-based image synthesis, deep learning-based image generation methods can learn complex patterns from large training datasets that allow them to generate diverse and realistic images. This was demonstrated by several works [7, 21, 48] through the usage of Generative Adversarial Nets (GANs) [10]. However, all of them only generated low-resolution patches and not high-resolution WSIs.

Though GANs have been the dominant approach to generate histopathological images, diffusion models are becoming increasingly popular in other domains [5]. Mainly because GANs tend to be unstable at training [24], and suffer from mode collapse [23]. Moreover, in many domains diffusion models have shown to outperform GANs [6], including medical images [26]. Consequently, Moghadam *et al*. [25] used diffusion for histopathology image generation. However, in contrast to our work, only for small patches not for entire WSIs.

Diffusion for high-resolution images. Training diffusion models [35] is expensive, and the computational complexity grows with the image resolution. Consequently, early works operated on low-resolution images up to 256×256 pixels [13]. Since then, various approaches have been proposed to enable generation of images with higher resolutions. However, to the best of our knowledge, we are the first to demonstrate image generation with diffusion models at a gigapixel scale.

A common approach to scale diffusion models for higher resolutions are latent diffusion models (LDMs) [31]. In LDMs, the diffusion is not done directly in pixel space but in a lower-dimensional latent space, which reduces computational complexity. Despite LDMs provide remarkable results, demonstrated resolutions [3,31] go only up to about 1024×1024 pixels. Even though the latent space is more compact than the pixel space, increasing the resolution still requires a corresponding enlargement of the latent space. Therefore, LDMs cannot be scaled up arbitrarily.

Another line of methods [14, 32, 33] generates high-resolution images by passing an initial low-resolution image through a cascade of upscaling diffusion models. These methods train multiple diffusion models, one for each upscaling stage. Each of these models takes the full input image of the previous stage as input and predicts an upscaled output. However, this requires training multiple upscale models, one for each stage. Also, the last upscaling model must still process the full-resolution image, which is unfeasible for our gigapixel case.

3. Image generation with diffusion

Before describing our method in detail, we give the necessary preliminaries about image generation with diffusion. Diffusion models [35] generate novel images by pushing noise through a series of denoising steps. In particular, first, a noise image $\mathbf{x}_0$ is sampled from the Gaussian distribution $\mathcal{N}\left(\mathbf{0}, \sigma_{max}^2 \mathbf{I}\right)$ with variance σ_{max}^2. Then, $\mathbf{x}_0$ is sequentially denoised for N steps, producing the sequence $\{\mathbf{x}_i\}_{i\in[0,N]}$, where the noise level σ_i of each $\mathbf{x}_i$ decreases with each step

$$\sigma_0 = \sigma_{\text{max}} > \sigma_1 > \cdots > \sigma_{min} > \sigma_N = 0, \tag{1}$$

where σ_{min} is the minimum noise level. The last image $\mathbf{x}_N$ of this sequential denoising process is noise-free, and follows the data distribution p_{data} that was used to train the model.

The denoising process of diffusion models can be modelled with stochastic differential equations (SDEs). Additionally, Song *et al*. [38] proposed that every denoising SDE has a correspoding probability flow ordinary differential equation (ODE) with the same marginals. While SDEs typically converge to higher quality results after numerous steps, ODEs can still gives competitive results with significantly fewer steps [18, 38]. Since our method runs multiple diffusion processes to generate a single WSI, we use ODE-based denoising to keep the overall sampling time within a reasonable scope.

While various variations of the probability flow ODE exist, many of them can be expressed with one general equation [18]:

$$\mathrm{d}\mathbf{x} = \left[\frac{\dot{s}(t)}{s(t)}\boldsymbol{x} - s(t)^2 \dot{\sigma}(t)\sigma(t)\nabla_{\boldsymbol{x}} \log p\left(\frac{\boldsymbol{x}}{s(t)}; \sigma(t)\right)\right] \mathrm{d}t, \tag{2}$$

where for time t the function $\sigma(t)$ controls the amount of noise, $s(t)$ scales the image, and $\dot{\sigma}(t)$ and $\dot{s}(t)$ are the respective time derivatives. Setting $\sigma(t)$ and $s(t)$ accordingly, recovers various ODEs, *e.g*., variance preserving (VP) [38], variance exploding (VE) [38], DDIM [36], iDDPM [27] or EDM [18]. We use the EDM formulation, since it has shown to be favourable in terms of sampling speed and image quality [18]. The EDM ODE is obtained by setting $s(t) = 1$ and the noise-level as $\sigma(t) = t$ in Eq. (2). For clarity, we continue to denote the noise level as $\sigma(t)$, a function parametrized by time t, instead of replacing it directly with t, leading to the following EDM ODE

$$\mathrm{d}\mathbf{x} = \left[-\sigma(t)\, \nabla_{\boldsymbol{x}} \log p\left(\mathbf{x}; \sigma(t)\right)\right] \mathrm{d}t. \tag{3}$$

Following the empirical results and theoretical justifications of Karras *et al*. [18], we set time steps $t_{i\in[0,N]}$ as

$$t_i = \left(\sigma_{\max}{}^{\frac{1}{\rho}} + \frac{i}{N-1}\left(\sigma_{\min}{}^{\frac{1}{\rho}} - \sigma_{\max}{}^{\frac{1}{\rho}}\right)\right)^{\rho}, \tag{4}$$

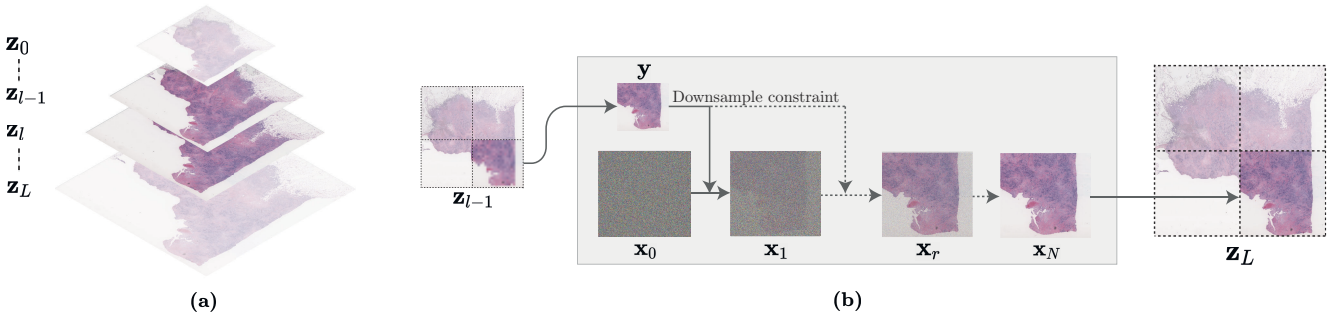

Figure 3. Overview of our method. **(a)** Shows how we upscale an initial low-resolution image $\mathbf{z}_0$ to a WSI $\mathbf{z}_L$ through L upscaling stages. **(b)** Shows how one stage upscales the image $\mathbf{z}_{l-1}$ to the image $\mathbf{z}_l$ using our diffusion-based approach. We split the image $\mathbf{z}_{l-1}$ into patches, each having a lower resolution than our diffusion model. We then provide each patch as a low-resolution guide $\mathbf{y}$ to a diffusion process. Throughout denoising, diffusion is pushed in a direction that satisfies a downsampling constraint with the guide $\mathbf{y}$. However, we stop enforcing this constraint after r iterations, which relaxes the constraint. Hence, the resulting images $\mathbf{x}_N$ follow the coarse structure of the guide $\mathbf{y}$, with increased resolution and added details. Finally, we stitch patches to the image $\mathbf{z}_L$.

where ρ adjusts between shortening steps near σ_{min} and lengthening those near σ_{max}.

To solve the ODE given in Eq. (3), one expresses the gradient of the log-likelihood w.r.t. input $\mathbf{x}$, *i.e.* the score function, as

$$\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}; \sigma) = \frac{D_\theta(\mathbf{x}; \sigma) - \mathbf{x}}{\sigma^2}, \tag{5}$$

where the function $D_\theta(\mathbf{x}; \sigma)$ parametrized by θ, takes a noisy image $\mathbf{x}$ and its noise level σ as input, and outputs a denoised image. After training the denoiser $D_\theta(\mathbf{x}; \sigma)$, any numerical ODE solver can be used to solve the ODE given by putting Eq. (5) into Eq. (3). Consequently, images can be generated by sampling noise, followed by sequential denoising using the ODE.

4. Method

Our method uses a diffusion model trained on histopathological images of size $M \times M$ to generate high-resolution WSIs of size $H \times H$, where $H \gg M$. To generate images of much larger resolution than the resolution of the diffusion model, we use a coarse-to-fine scheme. In this scheme, we first sample with the diffusion model an initial image $\mathbf{z}_0 \in \mathbb{R}^{M \times M}$. Then, we sequentially upscale it in L stages, producing the sequence $\{\mathbf{z}_l\}_{l \in [0,L]}$, where each image $\mathbf{z}_l$ has a k-times larger resolution compared to its predecessor $\mathbf{z}_{l-1}$, and the last image $\mathbf{z}_L \in \mathbb{R}^{M \times M}$ resembles a high-resolution WSI. Fig. 3 (a) illustrates this coarse-to-fine upscaling.

At each stage l of our coarse-to-fine scheme, we compute the higher-resolution image $\mathbf{z}_l$ through a diffusion process that is guided by the preceding lower-resolution image $\mathbf{z}_{l-1}$. Through this guidance, the image $\mathbf{z}_l$ is generated such that it follows the coarse structure of $\mathbf{z}_{l-1}$ while introducing novel details and having increased resolution. Due to the limited resolution of the diffusion model, we generate $\mathbf{z}_l$ patch-wise. Importantly, to prevent stitching artefacts in the image $\mathbf{z}_l$, despite patch-wise processing, we introduce a novel technique: grid-shift. Fig. 3 (b) summarizes the upscaling from $\mathbf{z}_{l-1}$ to $\mathbf{z}_l$.

In the following, we describe our method in detail. We start with the design of our diffusion denoising function in Sec. 4.1 and its training in Sec. 4.2. Followed by our guided denoising step for diffusion in Sec. 4.3 and the description of grid-shift in Sec. 4.4.

4.1. Diffusion denoiser

As discussed in Sec. 3, for diffusion, we need a denoiser function $D_\theta(\mathbf{x}; \sigma)$ that denoises images at each timestep. We propose to condition the denoiser $D_\theta(\mathbf{x}; \sigma)$ not only with noise level σ but also with the spatial image resolution s in $\mu\text{m/px}$. While in many applications the spatial resolution is unknown, it is consistently available in our case, as slide scanners usually save it in the metadata of WSIs. Conditioning allows us to control the spatial resolution of generated images. This is crucial for our coarse-to-fine scheme. Setting a high spatial resolution for the initial image ensures it depicts a macroscopic overview of a tissue sample. While decreasing spatial resolution accordingly at later refinement stages, conditions the network to introduce small details like cellular structures.

For denoising, we introduce a network $F_\theta(\mathbf{x}; \sigma, s)$, where we implement the conditioning on noise σ and spatial resolution s with a sinusoidal positional encoding [40]. However, we do not use the network $F_\theta(\mathbf{x}; \sigma, s)$ to directly denoise images, *i.e.* $D_\theta(\mathbf{x}; \sigma, s) = F_\theta(\cdot)$. Instead, we use the network preconditioning of Karras *et al.* [18]

$$D_\theta(\mathbf{x}; \sigma, s) = c_{\text{skip}}(\sigma)\,\mathbf{x} + c_{\text{out}}(\sigma)\,F_\theta\big(c_{\text{in}}(\sigma)\,\mathbf{x}; \sigma, s\big), \tag{6}$$

where the functions $c_{\text{in}}(\sigma)$ and $c_{\text{out}}(\sigma)$ scale the inputs and

outputs of the network $F_\theta(\mathbf{x}; \sigma, s)$, and $c_{\text{skip}}(\sigma)$ is a σ-dependent skip connection. These three functions scale network input and training targets to unit variance across all noise levels σ, which is beneficial for neural network training [15]. Additionally, $c_{\text{skip}}(\sigma)$ controls, if for denoising, the network has to predict the denoised image directly, only the noise component or a mixture of both. Empirically, it has been demonstrated that it depends on the noise level σ which of these cases is easier to learn, and $c_{\text{skip}}(\sigma)$ is set to adapt accordingly. We provide the full expressions of $c_{\text{in}}(\sigma)$, $c_{\text{out}}(\sigma)$ and $c_{\text{skip}}(\sigma)$ in the appendix.

4.2. Training

Using our denoiser function $D_\theta(\mathbf{x}; \sigma, s)$ given in Eq. (6), we can define the training loss for the diffusion model. In particular, we minimize the expected L_2 denoising error

$$\mathbb{E}_{s,\tilde{\mathbf{x}},\sigma,\mathbf{n}}\left[\lambda(\sigma)\|D_\theta(\tilde{\mathbf{x}} + \mathbf{n}; \sigma, s) - \tilde{\mathbf{x}}\|_2^2\right], \tag{7}$$

where the function $\lambda(\sigma)$ weights loss terms equally across all noise levels σ. At first, we sample the spatial resolution uniformly $s \sim \mathcal{U}(s_{\text{min}}, s_{\text{max}})$, where s_{min} and s_{max} refer to the smallest respectively largest spatial resolution of image patches in the training dataset. Then, we sample images from the distribution of training patches having spatial resolution s, *i.e.* $\tilde{\mathbf{x}} \sim p_{\text{train}|s}$. Finally, we sample noise levels σ from a log-normal distribution, and noise as $\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$.

4.3. Guided denoising step

Like a conventional diffusion denoising step, our guided denoising step removes noise from a noisy input image $\mathbf{x}_i$ with noise level $\sigma(t_i)$ such that the result $\mathbf{x}_{i+1}$ has noise level $\sigma(t_{i+1}) < \sigma(t_i)$. Additionally, we guide the denoising step with a low-resolution guidance patch $\mathbf{y} \in \mathbb{R}^{d\times 1}$ from the preceding layer $\mathbf{z}_{l-1}$. The goal of guidance is that the fully denoised image $\mathbf{x}_0$ follows the coarse structure of the guidance patch $\mathbf{y}$ while having additional details and a higher resolution. We implement this guidance through a relaxed super-resolution constraint.

For further derivations, we denote $\mathbf{u} \in \mathbb{R}^{D\times 1}$ as the output of the denoiser function $D_\theta(\mathbf{x}_i; \sigma(t_i), s)$ at step t_i. Notably, $\mathbf{u}$ gives at each denoising step an estimate of the *fully* denoised image $\mathbf{x}_0$. In our guided denoising step, we replace the initial estimation $\mathbf{u}$ of the denoised image with a guided estimate $\bar{\mathbf{u}}$, which is computed to be close to $\mathbf{u}$ while additionally satisfying a guidance constraint. This basically resembles the concept of projected gradient descent.

For guidance, we introduce the downsampling constraint $\mathbf{A}\mathbf{u} = \mathbf{y}$, where $\mathbf{A} \in \mathbb{R}^{d\times D}$ is a known linear downsampling operator. Therefore, downsampling the estimate $\mathbf{u}$ should equal the low-resolution guide $\mathbf{y}$. We can compute the guided estimate $\bar{\mathbf{u}}$ through the following optimization problem

$$\bar{\mathbf{u}} = \arg\min_{\bar{\mathbf{u}}} \frac{1}{2}\|\mathbf{u} - \bar{\mathbf{u}}\|^2 \quad \text{s.t. } \mathbf{A}\bar{\mathbf{u}} = \mathbf{y}, \tag{8}$$

that can solved using the method of Lagrangian multipliers. We provide a full derivation in the appendix and continue here with the solution

$$\bar{\mathbf{u}} = (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{u} + \mathbf{A}^\dagger\mathbf{y}, \tag{9}$$

where $\mathbf{A}^\dagger$ is the pseudoinverse for full row rank matrices

$$\mathbf{A}^\dagger = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}. \tag{10}$$

Notably, Eq. (9) resembles the proposed rectification equation of DDNM [42], a method to solve linear inverse problems with diffusion models. However, DDNM presents a different derivation based on a range-space null-space decomposition. Also, DDNM uses SDE-based diffusion processes, contrary to our ODE-based setting, leading to a different application of Eq. (9).

So far, our guidance resembles unrelaxed super-resolution. However, we do not strictly enforce the downsampling constraint, but relax it. Hence, we allow slight differences, between the downsampled fully denoised image $\mathbf{x}_0$ and the low-resolution guide $\mathbf{y}$. For relaxation, we stop replacing the estimate $\mathbf{u}$ with the guided estimate $\bar{\mathbf{u}}$ at iterations i where $i > r$. Consequently, in the last denoising steps, changes to the image are allowed that do not satisfy the downsample constraint. The strength of relaxation is controlled through r. If $r = 0$, the guidance constraint is enforced at all iterations, leading to no relaxation. Contrarily, if $r = N$, the constraint is never applied, leading to full relaxation. By setting r to values in between controls the amount of relaxation accordingly.

There are multiple reasons why we relax the downsample constraint. In our coarse-to-fine scheme, we do not pursue strict upsampling; instead, the diffusion model should add new details at every stage. Adhering strictly to the downsample constraint would restrict the flexibility to add new details. Furthermore, without relaxation, the downsampling constraint would be enforced across all upscaling stages. This is unreasonable due to the vast upscaling factors we face. For instance, if we have a diffusion model with input size 512×512 and generate a WSI with a resolution of $65\,536 \times 65\,536$, we have an upscaling factor of 128. Consequently, for a 512×512 area in the full-resolution WSI, the downsampling constraint would be enforced with a 4×4 patch in the lowest-resolution image. Clearly, this does not introduce any meaningful information. Moreover, without relaxation, even single-pixel errors at the lowest-resolution can distort large areas in the full-resolution image.

Finally, with Eq. (5) the score function, Eq. (3) the EDM ODE, and our guided estimation $\bar{\mathbf{u}}$, inplace of the denoiser

Algorithm 1 Guided denoising step

Input: Noisy image $\mathbf{x}_i$, guide $\mathbf{y}$, step i, spatial-resolution s
Output: Denoised image $\mathbf{x}_{i+1}$

1: $\mathbf{u} \leftarrow D_\theta(\mathbf{x}_i; \sigma(t_i), s)$
2: **if** $i < r$ **then**
3: $\quad \bar{\mathbf{u}} \leftarrow \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{u} + \mathbf{A}^\dagger\mathbf{y}$
4: $\quad \mathbf{d}_i \leftarrow \left(\mathbf{x}_i - \bar{\mathbf{u}}\right)/\sigma(t_i)$
5: **else**
6: $\quad \mathbf{d}_i \leftarrow \left(\mathbf{x}_i - \mathbf{u}\right)/\sigma(t_i)$
7: $\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + (t_{i+1} - t_i)\,\mathbf{d}_i$
8: **if** $t_{i+1} \neq 0$ **then** ▷ *Skip 2^{nd} order correction at last step*
9: $\quad \mathbf{u}' \leftarrow D_\theta(\mathbf{x}_{i+1}; \sigma(t_{i+1}), s)$
10: $\quad$ **if** $i < r$ **then**
11: $\quad\quad \bar{\mathbf{u}}' \leftarrow \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{u}' + \mathbf{A}^\dagger\mathbf{y}$
12: $\quad\quad \mathbf{d}_i' \leftarrow \left(\mathbf{x}_{i+1} - \bar{\mathbf{u}}'\right)/\sigma(t_{i+1})$
13: $\quad$ **else**
14: $\quad\quad \mathbf{d}_i' \leftarrow \left(\mathbf{x}_{i+1} - \mathbf{u}'\right)/\sigma(t_{i+1})$
15: $\quad \mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + (t_{i+1} - t_i)\left(\frac{1}{2}\mathbf{d}_i + \frac{1}{2}\mathbf{d}_i'\right)$
16: **return** $\mathbf{x}_{i+1}$

function $D_\theta(\mathbf{x}; \sigma, s)$, we get

$$\mathrm{d}\mathbf{x} = \frac{\mathbf{x} - \bar{\mathbf{u}}}{\sigma(t)}\,\mathrm{d}t. \tag{11}$$

In principle, we can solve Eq. (11) with any black-box ODE solver. Here, we use Heun's 2nd order solver [2], a predictor-corrector method, which has shown a good trade-off between truncation error and number of function evaluations in the context of diffusion models [16]. Algorithm 1 summarizes our guided denoising step. Note that skipping lines 8 to 15 simplifies Heun's 2nd order method to a simple Euler step.

4.4. Grid-shift

To avoid stitching artefacts in our patch-wise refinement scheme, we propose grid-shift. If we simply do patch-wise refinement and then stitch the refined patches back to a high-resolution image, the result could suffer from stitching artefacts. This is because there is no guarantee that the areas at the edges of neighbouring patches align such that they can be stitched seamlessly.

A recently proposed method to avoid stitching artefacts at patch-wise image processing with diffusion models is mask-shifting [42]. The idea of mask-shifting is to use overlapping patches. For each patch, areas that overlap with previously computed neighbouring patches are held constant during diffusion. This incorporates the content of a previously computed patch into the computation of its following patches. And consequently leads to smooth transitions between neighbouring patches.

We argue that mask-shifting has two drawbacks. At first, patches must be processed sequentially, which is not trivial to parallelize. And second, using overlapping patches increases the amount of total patches to process, increasing the computation time significantly depending on the amount of overlap between patches.

Algorithm 2 Coarse-to-fine scheme with grid-shift

Input: Low-resolution image $\mathbf{z}_0$, and its spatial-resolution s
Output: High-resolution WSI $\mathbf{z}_L$

for Stage l in $[1, L]$ **do**
  $s \leftarrow s/k$ ▷ *Adapt spatial-resolution to current stage*
  $\boldsymbol{x}_0 \sim \mathcal{N}\left(\mathbf{0}|\sigma_{\max}^2\mathbf{I}\right)$
  for i in $[1, \ldots, N]$ **do**
    shift_patch_grid()
    for $\mathbf{x}$, $\mathbf{y}$ in patch($\mathbf{x}_{i-1}$), patch($\mathbf{z}_{l-1}$) **do**
      $\mathbf{x}_{i,p} \leftarrow$ Algorithm 1$(\mathbf{x}, \mathbf{y}, i, s)$
    $\mathbf{x}_i \leftarrow$ stitch_patches($[\mathbf{x}_{i,0}, \ldots, \mathbf{x}_{i,P}]$)
  $\mathbf{z}_l \leftarrow \mathbf{x}_N$
return $\mathbf{z}_L$

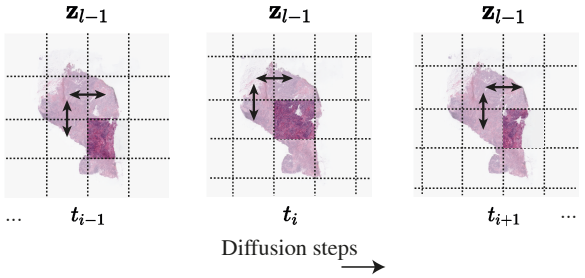

Figure 4. Visualization of grid-shift. After each diffusion step, we shift the patch grid that is used to extract guidance patches from the preceding image $\mathbf{z}_{l-1}$.

With grid-shift, we address both discussed drawbacks of mask-shifting by interleaving diffusion iterations with patching. Instead of using a fixed grid to extract patches, we shift the patch grid after each diffusion step. This makes patch boundaries temporary since they change after each diffusion step. Consequently, information between neighbouring patches is continuously transferred, resulting in a more coherent result without visible seams. Fig. 4 illustrates grid-shift. In our experiments, we shifted the patch-grid with random translations and padded boundary patches with the background colour.

Grid-shift has two computational advantages over mask-shifting. First, it does not increase the total amount of patches to process. And second, during one diffusion step, all patches are processed independently. Therefore, grid-shift is trivial to parallelize, *e.g.*, for a multi-GPU implementation. Algorithm 2 shows our full coarse-to-fine scheme with grid-shift.

5. Experiments

To evaluate our method, we performed a user study with pathologists, and quantitative evaluations. Particularly, quantitative evaluation is challenging due to the lack of a

suitable standardized metric. Common metrics for generative models, such as FID [39], or improved precision (IP) and improved recall (IR) [19] require features from pre-trained networks. These standardized networks have a fixed input size of 224×224. Downscaling WSIs to this resolution would discard most information, making the metrics inconclusive. Also, these metrics require large sample sizes of $50\,000$ images to provide consistent results. Generating that many WSIs is infeasible, considering that we need ~ 40 minutes for a single WSI. Moreover, the metrics utilize feature spaces strongly influenced by ImageNet classes [20]. Using these feature spaces to evaluate images from entirely different domains than ImageNet, such as histopathology images might be problematic.

Due to the discussed limitations, our quantitative evaluations are restricted to isolated evaluations of our diffusion model without the coarse-to-fine scheme. In terms of metrics, we use IP and IR, following Moghadam *et al*. [25]. However, we add that these metrics should be taken with reservations due to their ImageNet-related feature spaces.

Additionally, to the experiments presented in this section, we compare our method with multiple super-resolution methods in the appendix.

Data. For all experiments, we used the The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) dataset [45]. The dataset contains 1 978 high-resolution WSIs stained using various protocols showing diverse tissue types, *e.g*., epithelium, muscle, and connective tissue. For training, we extract patches from the dataset with spatial resolutions ranging from $s_{\text{min}} = 0.3\,\mu\text{m/px}$ to $s_{\text{max}} = 150\,\mu\text{m/px}$.

Setup. We generate WSIs at a resolution of $65\,536 \times 65\,536$ pixels. For the diffusion model, we use a resolution of 512×512 pixels. In our coarse-to-fine scheme, we use an upscaling factor of $k = 2$ at each stage, resulting in $L = 7$ stages in total. Initial images $\mathbf{z}_0$ are generated with a spatial-resolution randomly between 80 $\mu\text{m/px}$ and 150 $\mu\text{m/px}$. We set the number of diffusion denoising steps to $N = 40$ based on the results of Sec. 5.1. The relaxation parameter of our relaxed super-resolution constraint is set to $r = 28$, which was manually tuned towards a good tradeoff between consistency and novelty. For the downsampling operator $\mathbf{A}$ we use average-pooling.

We train for five days on four NVIDIA Quadro RTX 8000 GPUs with 48 GB of memory each. It took on average $\sim$40 minutes on one GPU to sample a single WSIs with a resolution of 65 536×65 536 pixels. For the diffusion-related hyperparameters, we use, if not otherwise stated, the proposed settings of Karras *et al*. [18]. Likely, an extensive hyperparameter search could further improve our results, but given the extensive training cost, it is beyond our computational capacities.

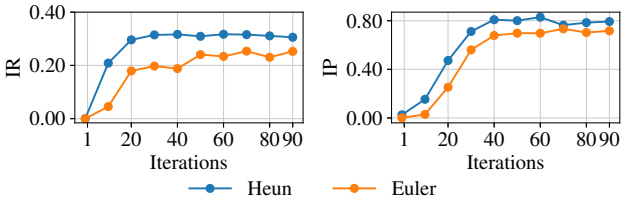

Figure 5. IR and IP values for 512×512 patches with $1\mu\text{m/px}$ for different numbers of denoising iterations using Heun and Euler as ODE solver.

	Spatial Resolution [µm / px]						
	0.3	1.0	25	50	100	150	$\mathcal{U}(0.3, 150)$
IP	0.81	0.82	0.82	0.85	0.82	0.84	0.86
IR	0.32	0.33	0.36	0.38	0.37	0.36	0.34

Table 1. IP and IR for 512×512 patches at varying spatial resolutions. The last column shows results for uniformly sampled spatial resolutions between 0.3 and 150.

5.1. Number of diffusion iterations

An important hyperparameter we must choose is the number of diffusion iterations N. Too few iterations degrade image quality, while too many might increase runtime unnecessarily. Finding the right balance is crucial for us since we have to run many diffusion processes to sample a single WSI. To this end, we compute IR and IP scores for generating 512×512 patches with a spatial resolution of $1\mu\text{m/px}$ across different iteration numbers N. We also validated if using Heun's 2nd order method is beneficial over a plain Euler solver. Fig. 5 shows the results. According to the metrics, the Heun solver showed preferable performance. After an additional manual inspection, we chose $N = 40$ as a good tradeoff between image quality and runtime for further experiments.

5.2. Image quality across spatial resolutions

To evalute our spatial resolution conditioning of the model, we compute IP and IR metrics across a variety of different spatial resolutions. We obtained all results from a *single* model trained with uniformly sampled spatial resolutions as described in our training setup. We then conditionally sampled $50\,000$ images for each spatial resolution, and compared them with images of identical spatial resolution from the training dataset. Tab. 1 shows the result. According to the metrics, performance is relatively consistent across the full range of spatial resolutions without any major outliers.

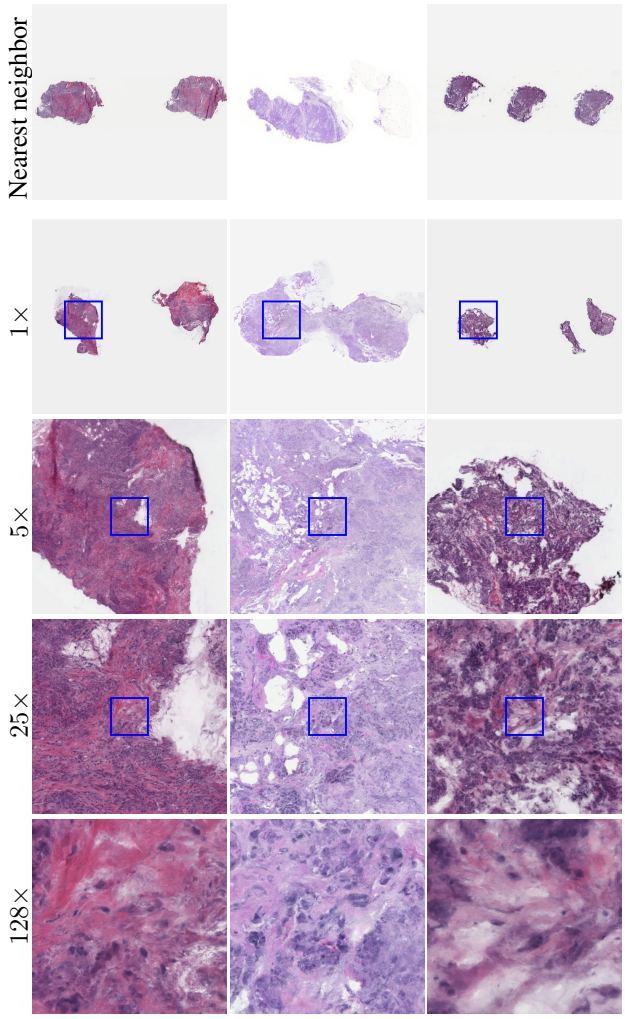

Figure 6. WSIs with a resolution of $65\,536 \times 65\,536$ pixels generated by our method. We show 512×512 patches extracted at different magnifications, the blue rectangle shows the location of the patch in the subsequent row. The top row shows for each WSI the nearest neighbor in the training data. To find neighbors, we resized WSIs to 512×512 and compared WSIs in the feature space of Inception-v3.

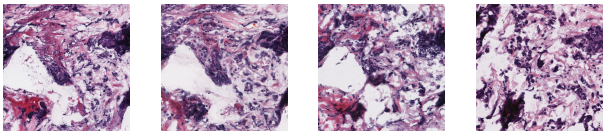

Figure 7. Decreasing the parameter r relaxes the super-resolution constraint.

5.3. Relaxation parameter

To evaluate the influence of our super-resolution relaxation parameter r, we perform a simple experiment. We sample a 512×512 sized image with our diffusion model, downsample it to 256×256, and provide it as a guide $\mathbf{y}$ to a diffusion process guided by our relaxed super-resolution constraint. We repeat this for multiple values of the relaxation parameter r. Fig. 7 shows the result, it can be clearly seen how consistency with the initial images decreases with decreasing relaxation parameter r.

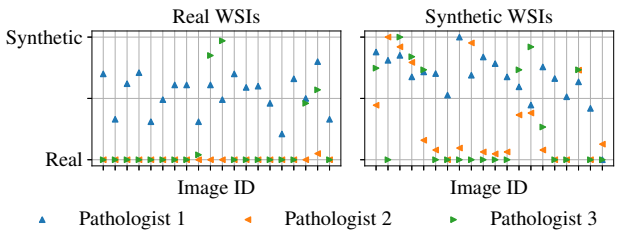

Figure 8. Result of our user-study. The plots show how three pathologists rated the realness of 20 real versus 20 synthetic WSIs.

5.4. User study

To evaluate the quality of our synthetic WSIs, we conducted a user study with three pathologists. For the study, we used 20 synthetic WSIs, and 20 real WSIs randomly chosen from the training data. We presented the WSIs to the pathologists in random order and asked them to identify whether each WSI was synthetic or real. To this end, they were given a slider to select values between 0 = "I believe the slide is real." and 100 = "I believe the slide is synthetic.". Values in between represented corresponding gradations between the two extremes. Fig. 6 shows three WSIs that were part of the study.

Fig. 8 shows the ratings for all individual images. Despite our study's limited sample size, our primary goal was to assess whether our method could generate plausible-looking WSIs. The results of our study indicate that this is the case, as pathologists could not consistently distinguish our synthetic WSIs from real ones.

6. Conclusion

We presented a method that generates synthetic histopathological WSIs at resolutions up to $65\,536 \times 65\,536$. We evaluated parts of our method quantitatively and also performed a user study with pathologists. Our study's results showed that pathologists could not consistently differentiate the WSIs generated by our method from real ones. In the future, the duration of WSI generation could be further reduced by incorporating distillation-based diffusion models [34, 37].

Acknowledgement This work has been co-funded by the Austrian Science Fund (FWF), Project: P-32554 explainable Artificial Intelligence.

References

[1] Grégory Apou, Friedrich Feuerhake, Germain Forestier, Benoît Naegel, and Cédric Wemmert. Synthesizing whole slide images. *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2015.

[2] Uri M Ascher and Linda R Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*, volume 61. Siam, 1998.

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*.

[4] Daniel Bug, Friedrich Feuerhake, and Dorit Merhof. Foreground extraction for histopathological whole slide imaging. *Bildverarbeitung für die Medizin 2015: Algorithmen-Systeme-Anwendungen. Proceedings des Workshops vom 15. bis 17. März 2015 in Lübeck*, 2015.

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE TPAMI*, 2023.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

[7] James M Dolezal, Rachelle Wolk, Hanna M Hieromnimon, Frederick M Howard, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Jung Woo Kwon, Meghana Agni, et al. Deep learning generates synthetic cancer histology for explainability and education. *NPJ Precision Oncology*, 2023.

[8] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4):686–696, 2021.

[9] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *ICCV*, 1999.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014.

[11] David J Heeger and James R Bergen. Pyramid-based texture analysis/synthesis. 1995.

[12] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[14] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 2022.

[15] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Normalization techniques in training dnns: Methodology, analysis and application. *IEEE TPAMI*, 2023.

[16] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

[17] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[18] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022.

[19] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 32, 2019.

[20] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. 2023.

[21] Adrian B Levine, Jason Peng, David Farnell, Mitchell Nursey, Yiping Wang, Julia R Naso, Hezhen Ren, Hossein Farahani, Colin Chen, Derek Chiu, et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *The Journal of pathology*, 2020.

[22] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017.

[23] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[24] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *NeurIPS*, 31, 2018.

[25] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *WACV*, 2023.

[26] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Sven Nebelung, Jakob Nikolas Kather, et al. Diffusion probabilistic models beat gans on medical images. *arXiv preprint arXiv:2212.07501*, 2022.

[27] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. 2021.

[28] Sergey I Nikolenko. *Synthetic data for deep learning*. Springer, 2021.

[29] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. 2000.

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

[33] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022.

[34] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *ICLR*, 2022.

[35] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015.

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021.

[37] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[41] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*.

[42] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *ICLR*, 2023.

[43] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. *arXiv preprint arXiv:2304.03283*, 2023.

[44] Li-Yi Wei, Sylvain Lefebvre, Vivek Kwatra, and Greg Turk. State of the art in example-based texture synthesis. *Eurographics 2009, State of the Art Report, EG-STAR*, pages 93–117, 2009.

[45] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 2013.

[46] Li-Yi Wie, Sylvain Lefebvre, Vivek Kwatra, and Greg Turk. State of the Art in Example-based Texture Synthesis. *Eurographics, 2009*.

[47] Ellery Wulczyn, David F Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ digital medicine*, 2021.

[48] Yuan Xue, Jiarong Ye, Qianying Zhou, L Rodney Long, Sameer Antani, Zhiyun Xue, Carl Cornwell, Richard Zaino, Keith C Cheng, and Xiaolei Huang. Selective synthetic augmentation with histogan for improved histopathology image classification. *Medical image analysis*, 2021.