Diffusion Models for Generative Histopathology

Niranjan Sridhar[1(✉)], Michael Elad[2], Carson McNeil[1], Ehud Rivlin[2], and Daniel Freedman[2]

[1] Verily Research, South San Francisco 94080, USA
nirsd@verily.com
[2] Verily Research, Tel Aviv, Israel

Abstract. Conventional histopathology requires chemical staining to make tissue samples usable by pathologists for diagnosis. This introduces cost and variability and does not conserve the tissue for advanced molecular analysis of the sample. We demonstrate the use of conditional denoising diffusion models applied to non-destructive autofluorescence images of tissue samples in order to generate virtually stained images. To demonstrate the power of this technique, we would like to measure the perceptual quality of the generated images; however, standard measures like the Frechet Inception Distance (FID) are inappropriate for this task, as they have been trained on natural images. We therefore introduce a new perceptual measure, the Frechet StainNet Distance (FSD), and show that our model attains significantly higher FSD than competing pix2pix models. Finally, we also present a method of quantifying uncertain regions of the image using the variations produced by diffusion models.

Keywords: Diffusion · Pathology

1 Introduction

Conventional histopathology involves obtaining tissue sections from patient biopsies and applying chemical staining protocols which highlight different biological features of the tissue. This stained tissue can then be assessed and diagnosed by pathologist using a brightfield (BF) microscope. There are many chemical stains corresponding to different features to be highlighted. However, the process of staining can be destructive. A given stained tissue sample often cannot be used again for other analyses. Therefore, the cost of advanced testing, research or second opinions, which are often required for newer/rarer diseases, can be prohibitive. Additional drawbacks of histochemical staining include expensive laboratory infrastructure, slow processing times and the inherent variability in equipment and expertise.

Virtual staining [3,15,18,19] is an AI-enabled alternative which removes the need for chemical staining. Tissue samples are imaged using a non-destructive auto-fluorescence (AF) scanner. The AF image records the spatial distribution

A. Mukhopadhyay et al. (Eds.): DGM4MICCAI 2023, LNCS 14533, pp. 154–163, 2024.
https://doi.org/10.1007/978-3-031-53767-7_15

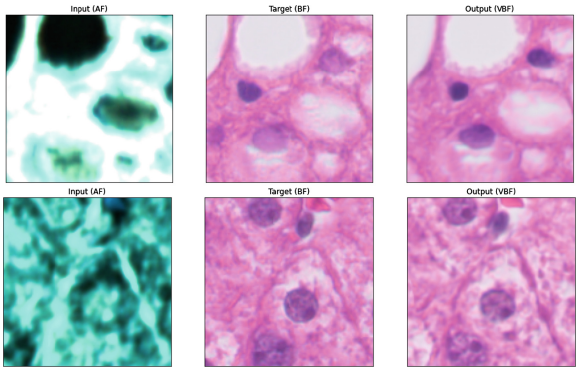

Fig. 1. Diffusion models are able to generate virtual stained outputs with both high fidelity to the target image and high perceptual quality.

emission spectra of the tissue after exposing it to excitation lasers and therefore contains information about both the condition and location of different biological features of the tissue. An image-to-image translation model can then be used to learn the mapping from the AF image of the tissue to its stained BF image. If this virtually stained image can capture all of the clinical features of real stained tissues, the pathologist can use the translated image for clinical diagnosis. Since the same AF image can be used for any number of stain types and the original tissue is preserved, virtual staining can greatly reduce the cost and effort of clinical pathology. The crucial step in this process is the image-to-image translation algorithm. In this paper, we apply conditional diffusion models to this task. We make the following key contributions:

1. **Diffusion Models for Staining.** We present a conditional diffusion model for virtual staining, which maps AF images to chemically stained BF images.
2. **Frechet StainNet Distance (FSD).** We develop a new technique for evaluating the perceptual quality of our output, referred to as the Frechet StainNet Distance. As compared to FID, FSD is much more appropriate for evaluating stained microscopy images. We note that FSD may be applied to other scenarios beyond that described in this paper.
3. **Significantly Improved Perceptual Quality.** We show empirically that as compared to conditional GANs, the diffusion models perform significantly better on perceptual quality as measured by FSD, while remaining comparable on distortion measures.
4. **Uncertainty Quantification.** We use the capabilities of our diffusion method to provide a reliable approximation of the uncertainty associated with the stained estimate per each pixel (Fig. 1).

2 Related Work

Virtual Staining: Until recently, the state-of-the-art in image-to-image translation were conditional GANs such as pix2pix [11] for paired datasets, and

CycleGANs [26] for unpaired ones. To prevent the GAN generator from hallucinating realistic images unrelated to the input, pixel-wise losses such as L_p distance between the virtual images and their corresponding ground truth are also used in addition to adversarial losses. A major drawback of GANs is that they are hard to train due to loss instability and mode collapse, e.g. see [2]. Previous efforts in virtual staining have used both pix2pix and CycleGANs [3,15,18,19]. These prior work often only report distortion measures such as L_p norms. However, for virtual stains to replace chemical stains in a clinical workflow, they must also look similar to human pathologists. Therefore it is important to benchmark these models on perceptual quality. In this work, we benchmark our models against a pix2pix model inspired by Rivenson et al. [19] with 128×128 resolution inputs, two discriminator losses (conditional and unconditional) and two pixel-wise losses (L1 and rotated L1).

Diffusion: Diffusion models [22] have recently emerged with impressive results on the task of unconditional image generation, beating GANs for generating images with high diversity and perceptual quality [6,8]. An important variation of these techniques is the conditional diffusion model, see e.g. [9,21,23,25] which is the basis of our current work. Saharia et al. [9,21] show that diffusion models can produce images with high perceptual quality without losing the structural and semantic information of the input image on a number of image-to-image translation tasks.

Perceptual Measures: The FID score [7] is commonly used to quantify perceptual quality. However, since the standard InceptionV3 model [24] used in FID has been trained on natural images, the measure is likely to have difficulty differentiating between varying distributions of histological images, which can be close to each other in the space of natural images. This has been documented previously for other data types such as audio and molecular data [12,16]. The tradeoff between the distortion between the expected and the predicted images, and the perceptual quality of the predicted image has been well studied [5]. Regression models that minimize the distortion between the labels and the prediction cannot produce outputs belonging to the expected output distribution. Therefore such models have low perceptual quality, i.e. they do not produce images that look like real images to humans. In contrast, GANs and diffusion models have the ability to generate images of high perceptual quality.

Uncertainty Quantification: Quantifying uncertainty in deep learning is difficult due to the lack of a closed form expression for the density. Deterministic models that produce a single output per input require complex interrogation to extract such information. Perturbative methods, such as LIME [17], involve repeated inference with varying data augmentations to estimate the effect of input variations on the output. In contrast, integrated methods, such as quantile regression [13], involve adding credible interval bound estimation as an additional training objective, either during the original training or after it. Generative models provide a new opportunity as they can sample different outputs from the target distribution upon repeated inference. Following [10], in conditional diffusion models

we apply a series of inference rounds and generate multiple results to approximate the distribution of the output conditioned on the input.

3 Methodology

3.1 Denoising Diffusion Probabilistic Models

A denoising diffusion probabilistic model [8] can be described as a parameterized Markov chain. The forward diffusion process is a series of steps that add small amounts of Gaussian noise to the data until the signal is destroyed. Given data x_0 which we consider a sampling of the distribution $q(x_0)$, we can create T vectors $\{x_1, ..., x_T\}$ of the same dimensions as x_0 defined by the forward diffusion process:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \quad \text{for} \quad t = 1, 2, \ \dots \ , T,$$

i.e. x_t is constructed as a mixture of x_{t-1} with a Gaussian noise, with the scaling variance parameter $\beta_t \in (0, 1)$. It immediately follows from the above:

$$q(x_t|x_0) = \prod_{i=1}^{t} q(x_i|x_{i-1}) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}) \quad \text{for} \quad t = 1, 2, \ \dots \ , T,$$

where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$. The number of steps T and the variance schedule β_t are chosen such that x_T is pure Gaussian noise, while at the same time the variances β_t of the forward process are small. Under these conditions, we can learn a reverse process p_θ which can be defined as

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \sigma_\theta(x_t, t)) \quad \text{for} \quad t = T, T-1, \ \dots \ , 1. \tag{1}$$

Note that chaining these probabilities leads to a sampled outcome x_0 following the probability density function

$$p_\theta(x_0) = p(x_T) \prod_{t=T}^{1} p_\theta(x_{t-1}|x_t).$$

Returning to our goal of reversing the diffusion process, we can leverage the following relationship:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}x_0 + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}x_t, \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t\mathbf{I}\right) \tag{2}$$

Observe the similarity between Eq. (1) and the above expression, where the later adds the knowledge of x_0. Thus, we can approximate p_θ by aligning the two moments of these Gaussians, which imply that we use a learned denoiser neural network $T_\theta(x_t, t)$ for estimating x_0 from x_t and t:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}T_\theta(x_t, t) + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}x_t. \tag{3}$$

During inference, this denoising neural network is recursively applied, starting from pure Gaussian noise, to produce samples from the data distribution.

In conditional diffusion, the denoiser is modified to include AF images of tissue sample, y, concatenated to the input, both during training and inference. As a result, with the modified denoiser $T_\theta(x_t, y, t)$, the output of the diffusion model is a sample from the posterior data distribution $q(x_0|y)$. This allows conditional diffusion models to be used for image-to-image translation where the input image is used as the condition.

3.2 Architecture

As is common practice in the diffusion literature [21], rather than learning $T_\theta(x_t, t)$ which returns the clean signal, one learns the noise itself (which is trivially related to the clean signal). To learn this noise estimator, we adopt the UNet [20] denoiser architecture, as proposed by Ho et al. [8] for diffusion models, and the improvements proposed by Saharia et al. [21]. The UNet model uses a stack of 6 blocks, each made of 2 residual layers and 1 downsampling convolution, followed by a stack of 6 blocks of 2 residual layers and 1 upsampling convolution. Skip connections connect the layers with the same spatial size. In addition, we use a global attention layer with 2 heads at each downsampled resolution and add the time-step embedding into each residual block.

3.3 Perceptual Quality Measures

For each model, we run inference on 20,000 128×128 tiles and evaluate the virtual stain results against real stain patches. In addition to standard L_p-based distortion measures, we consider the following two measures of perceptual quality:

FID: The Frechet Inception Distance (FID) score is a perceptual measure shown to correlate well with human perception of realism [7]. FID measures the Frechet distance between two multivariate Gaussians fit to features of generated and real images extracted by the pre-trained InceptionV3 model [24].

FSD: We created a new custom measure to characterize the perceptual quality of stained images, which we dub the Frechet StainNet Distance (FSD). We create a dataset where each training example is a 128×128 patch of a stained BF images with a corresponding label representing the slide-level clinical Non-Alcoholic SteatoHepatitis (NASH) steatosis score (for more details on this score, see Sect. 4.1). We then train a classification model, StainNet, on this dataset. The features from StainNet are then taken to be the outputs of the penultimate layer of the StainNet network. Analogously to FID, FSD then measures the Frechet distance between two multivariate Gaussians fit to the StainNet features: the first Gaussian for the generated images and the second Gaussian for the real images. We note that FSD may be applied to other scenarios beyond that described in this paper.

Table 1. Quantitative evaluation results of different methods. All evals are done on 20,000 test image patches.

Model	FID	FSD	L1	L2
Regression	356.4	624.8	**10.5**	**13.8**
pix2pix	115.7	54.1	12.7	18.1
Diffusion-B	82.2	15.2	15.1	21.1
Diffusion-B/R	81.7	34.5	13.5	19.3
Diffusion-L	**69.1**	**4.5**	13.4	19.3

3.4 Sample Diversity and Uncertainty

To calculate pixel-wise 90% credible intervals (i.e. we expect 10% of samples to fall outside the bounds), we follow the approach proposed by Hoshen et al. [10]. We sample 20 outputs for every input image, and use these to approximate the output image distribution and its 5^{th} and 95^{th} quantiles as the bounds. The credible interval size is then the difference between the upper and the lower bound values for each pixel. This well-motivated but heuristic notion of uncertainty is then properly calibrated using a calibration factor λ to the interval bounds, which is determined using our validation set [1,4,10].

4 Experiments

4.1 Experimental Setup

Dataset. We use a proprietary dataset collected from a clinical study of patients diagnosed with Non-Alcoholic SteatoHepatitis (NASH). The dataset contains 192 co-registered pairs of images of whole slides of liver tissue: one AF image (26 spectral channels) and one H&E chemically stained BF image (3 RGB channels). The whole slides are captured at 40x resolution yielding large gigapixel images of variable shapes and sizes. We split the slides into train/val/test data in 0.5:0.2:0.3 ratio. Finally, we extract paired patches of size 128×128 from both AF and BF images, and all of the training and evaluation is done at the patch level. Each slide is between 1000 to 10000 pixels height and width and corresponds to between 700 and 3000 patches; thus the combined dataset is approximately 200,000 patches. In addition, for each slide we also have a clinical steatosis score. This score is an ordinal class between 0–3 assigned by human expert hepatopathologists quantifying the amount of liver disease features they observe in the whole slide.

Training. The diffusion model is trained on 16 TPUv3 cores in parallel. We use a batch size of 16 and a learning rate of 1e−5 throughout the entire training for 1.5 million steps or 120 epochs. We choose the number of diffusion steps $T = 1000$ and set the forward diffusion variances β_t to increase following a cosine function from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$, in accordance with the findings of Nichol and Dhariwal [14].

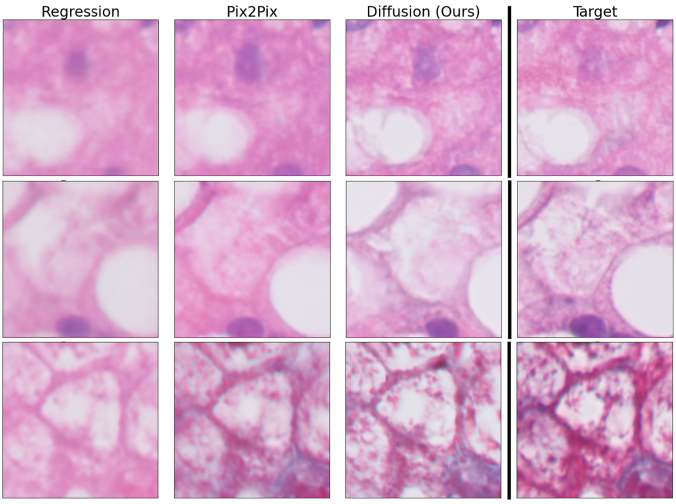

Fig. 2. We compare our Diffusion-L model with the target and benchmark it against regression and pix2pix models. Images generated by our diffusion model are closer to the target images in both texture and color.

Model Variants. In addition to our large diffusion model *Diffusion-L*, which has already been described, we train a number of variants. *Diffusion-B* is our base model which is similar to Diffusion-L but with only one single head attention layer at the 16×16 layer. The *Diffusion-B/R* model is the base model trained with an additional feature, in which a random part of the target image is masked and used as a prior; during inference, however, the prior is completely masked so that the generation is comparable to the other models.

4.2 Image Quality

We compare our diffusion model described above with a naive regression model, as well as a pix2pix (conditional GAN) model. Both models also use a UNet architecture, and the pix2pix model has additional unconditional and conditional adversarial losses. The results are presented in Table 1, which shows both the FID and FSD scores which measure perceptual quality, as well as L1 and L2 norms which measure distortion. Qualitative examples are shown in Fig. 2.

All of the diffusion models do score better (lower) in terms of FID scores; nevertheless, as previously noted, FID is not a very discriminative perceptual measure for stained pathology images, as it has been trained on natural images. For example, the Diffusion-B and Diffusion-B/R models attain almost identical FIDs. By contrast, FSD is much more discriminative and clearly shows that Diffusion-B/R has worse perceptual quality. Overall, the best result is attained by the Diffusion-L model, which receives an FSD score of 4.5; this is considerably better than the scores attained by the regression and pix2pix models, which are

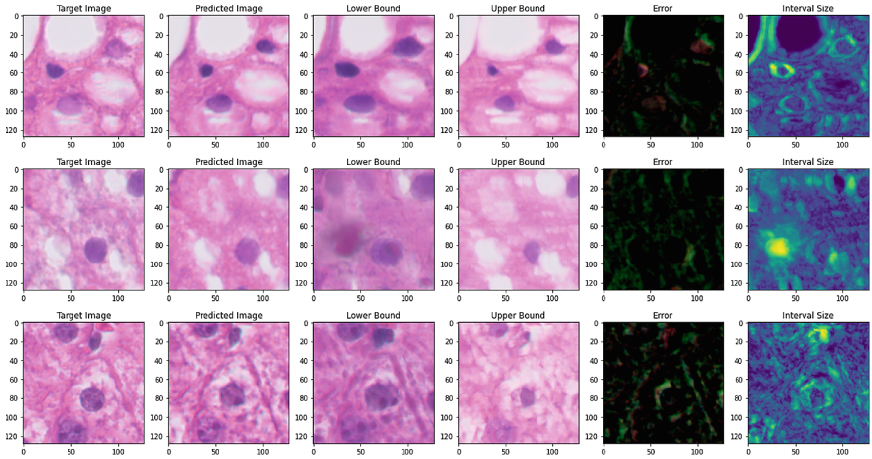

Fig. 3. Examples of per-pixel credible interval bound estimation using generative sampling. The 5^{th} and 95^{th} percentile for each pixel is used as the lower and upper bounds of the credible interval.

624.8 and 54.1, respectively. This perceptual advantage is demonstrated qualitatively in Fig. 2: images generated by Diffusion-L are closer to the target images in both texture and color than pix2pix and the regression model.

It has been theoretically established that attaining a better perceptual score leads to worse performance on distortion [5]. It is thus not surprising that the regression model attains the best distortion measures, as its loss is completely focused on the distortion; as a consequence, its FSD is very poor. Both pix2pix and the diffusion models aim at optimizing a combination of distortion and perceptual measures. Comparing the Diffusion-L and pix2pix models, we note that they have comparable distortion scores, despite the Diffusion-L model's significant performance advantage on perceptual scores.

4.3 Uncertainty Estimation

Figure 3 shows examples of our per-pixel uncertainty estimation. The interval size is the difference between the lower bound and the upper bound of the credible interval, thus larger intervals indicate greater uncertainty. Using our validation set, we observe a calibration factor $\lambda = 1.32$. As we can see in Fig. 3, nuclei are an important source of uncertainty in stains. This finding might motivate the development of future methods which focus on nuclei, e.g. through the use of manual annotation of some nuclei and weighted losses emphasizing these regions.

5 Conclusion

In this work, we demonstrate conditional diffusion models for synthesizing highly realistic histopathology images. We test the perceptual quality of these models

using a custom Frechet distance measure. The lack of resolution of the standard Frechet distance FID and the increased discrimination using our custom Frechet distance FSD, indicates that embeddings trained on natural image datasets are not general enough to capture perceptual quality for pathology images. More work is needed to determine whether new quality measures can generalize across a variety of medical image type or must being tailored to each specific image type such as the measure for NASH pathology images in this work. Our results suggest that conditional diffusion models are a promising approach for image-to-image translation tasks, even when we expect outputs with high fidelity and low sample diversity. The observed sample diversity itself can be usefully employed to compute an empirical measure of uncertainty.

Acknowledgements. We thank Shek Azizi, Saurabh Saxena and David Fleet for helpful discussions and suggestions to improve diffusion models. We also thank Yang Wang, Jessica Loo and Peter Cimermancic for their expertise and assistance with the clinical dataset, data processing and digital pathology.

References

1. Angelopoulos, A.N., et al.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: International Conference on Machine Learning, pp. 717–730. PMLR (2022)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223. PMLR (2017)
3. Bai, B., Yang, X., Li, Y., Zhang, Y., Pillar, N., Ozcan, A.: Deep learning-enabled virtual histological staining of biological samples (2023). https://doi.org/10.1038/s41377-023-01104-7. https://www.nature.com/articles/s41377-023-01104-7
4. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.: Distribution-free, risk-controlling prediction sets. J. ACM (JACM) **68**(6), 1–34 (2021)
5. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, June 2018. https://doi.org/10.1109/cvpr.2018.00652
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
9. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**(47), 1–33 (2022)
10. Horwitz, E., Hoshen, Y.: Conffusion: Confidence intervals for diffusion models (2022). https://doi.org/10.48550/ARXIV.2211.09795. https://arxiv.org/abs/2211.09795
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)

12. Kilgour, K., Zuluaga, M., Roblek, D., Sharifi, M.: Fréchet audio distance: a reference-free metric for evaluating music enhancement algorithms. In: INTERSPEECH, pp. 2350–2354 (2019)
13. Koenker, R., Bassett Jr, G.: Regression quantiles. Econometrica: journal of the Econometric Society, pp. 33–50 (1978)
14. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
15. Picon, A., et al.: Autofluorescence image reconstruction and virtual staining for in-vivo optical biopsying. IEEE Access **9**, 32081–32093 (2021). https://doi.org/10.1109/ACCESS.2021.3060926
16. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., Klambauer, G.: Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. J. Chem. Inf. Model. **58**(9), 1736–1741 (2018)
17. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
18. Rivenson, Y., Liu, T., Wei, Z., Zhang, Y., de Haan, K., Ozcan, A.: Phasestain: the digital staining of label-free quantitative phase microscopy images using deep learning. Light: Science and Applications (2019). https://doi.org/10.1038/s41377-019-0129-y. https://doi.org/10.1038/s41377-019-0129-y
19. Rivenson, Y., et al.: Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nature Biomed. Eng. **3**(6), 466–477 (2019). https://doi.org/10.1038/s41551-019-0362-y
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Saharia, C., et al.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10 (2022)
22. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
23. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
25. Tashiro, Y., Song, J., Song, Y., Ermon, S.: Csdi: conditional score-based diffusion models for probabilistic time series imputation. Adv. Neural. Inf. Process. Syst. **34**, 24804–24816 (2021)
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)