Hallucination Index: An Image Quality Metric for Generative Reconstruction Models

Matthew Tivnan[1][0000−0002−8326−2991], Siyeop Yoon[1][0000−−0002−9083−4089], Zhennong Chen[1][0000−−0001−6133−8716], Xiang Li[1][0000−−0002−9851−6376], Dufan Wu[1][0000−−0002−3204−3502], and Quanzheng Li[1][0000−−0002−9651−5820]

Massachusetts General Hospital and Harvard Medical School, 55 Fruit Street, Boston, MA, 02114. li.quanzheng@mgh.harvard.edu

Abstract. Generative image reconstruction algorithms such as measurement conditioned diffusion models are increasingly popular in the field of medical imaging. These powerful models can transform low signal-to-noise ratio (SNR) inputs into outputs with the appearance of high SNR. However, the outputs can have a new type of error called hallucinations. In medical imaging, these hallucinations may not be obvious to a Radiologist but could cause diagnostic errors. Generally, hallucination refers to error in estimation of object structure caused by a machine learning model, but there is no widely accepted method to evaluate hallucination magnitude. In this work, we propose a new image quality metric called the hallucination index. Our approach is to compute the Hellinger distance from the distribution of reconstructed images to a zero hallucination reference distribution. To evaluate our approach, we conducted a numerical experiment with electron microscopy images, simulated noisy measurements, and applied diffusion based reconstructions. We sampled the measurements and the generative reconstructions repeatedly to compute the sample mean and covariance. For the zero hallucination reference, we used the forward diffusion process applied to ground truth. Our results show that higher measurement SNR leads to lower hallucination index for the same apparent image quality. We also evaluated the impact of early stopping in the reverse diffusion process and found that more modest denoising strengths can reduce hallucination. We believe this metric could be useful for evaluation of generative image reconstructions or as a warning label to inform radiologists about the degree of hallucinations in medical images.

Keywords: Hallucinations, Uncertainty, Diffusion Models, Generative Models, Medical Image Reconstruction

1 Introduction

In the rapidly evolving landscape of medical image reconstruction algorithms, generative models have become increasingly popular. Examples include generative adversarial networks [15, 20, 21], normalizing flow models [7] [8], and diffusion models [9] [10][13][19]. These generative models have shown exceptional

capability in transforming low signal-to-noise ratio (SNR) inputs into higher quality outputs that appear to have high SNR with realistic structural details. However, the advancement of generative reconstruction methods introduces a new type of error called hallucinations. Hallucinations are commonly referenced in research literature [1] [2] [20] [11] [5] but definitions vary. Generally, hallucinations are structural uncertainty in the output of machine learning models, especially generative models. It is distinct from measurement noise which does not have the appearance of varied object structure.

Hallucinations in the context of medical imaging are particularly concerning and several members of the field are sounding the alarm [4] [16]. These errors, while not immediately obvious to radiologists, could lead to diagnostic inaccuracies, compromising patient safety. Despite the recognition of hallucinations as a potential issue, the field lacks a standard methodology for quantitative evaluation of such generative artifacts. To address this gap, we propose the *hallucination index*, a novel image quality metric designed to evaluate the extent of hallucinations in the outputs of generative algorithms. Our approach employs the Hellinger distance [12] between the distribution of reconstructed images and a zero-hallucination reference distribution. The reference distribution should be generated without a machine learning model, and it should have the same apparent SNR as the reconstructions without the structural variability. For example, in diffusion models, we can apply the forward diffusion process to ground truth.

One noteworthy alternative for evaluating hallucinations in medical images is the *Hallucination Map* proposed by [2]. This method involves projecting the reconstructions to the imaging system null space to identify features which provably did not arise from the measurements. While this method could be extremely useful, especially for ill-conditioned inverse problems, we argue it does not capture the possibility of hallucinations within the measurement space. For example, if an imaging system is unbiased with additive noise, it technically has no null space, but a generative denoising model may still be affected by hallucinations. By comparison, our hallucination index is a distribution-to-distribution distance metric, which could measure the hallucinations in this scenario.

The image reconstruction task we simulated is electron microscopy (EM) of cortical neurons using the Machine Intelligence from Cortical Networks (MICrONS) dataset [6] which is openly available under a Creative Commons Public License. This modality offers the high spatial resolution necessary for neuron body segmentation and connectome mapping. However, the time it takes to scan hundreds of terabytes of imaging data per cubic millimeter represents a significant bottleneck. One proposed solution is to decrease the dwell time [3]. While effective in reducing scan time, this method decreases the SNR of the resulting images [18]. Generative image restoration with diffusion models emerges as a promising solution for low dwell time EM reconstruction, but one concern is that hallucinated artifacts could interfere with segmentation, leading to incorrect interpretations of neural structures. Our experiment aims to evaluate the tendency of diffusion based reconstructions to hallucinate and how those hal-

lucinations vary with respect to other parameters and metrics like input SNR, reverse process stopping time, and mean squared error.
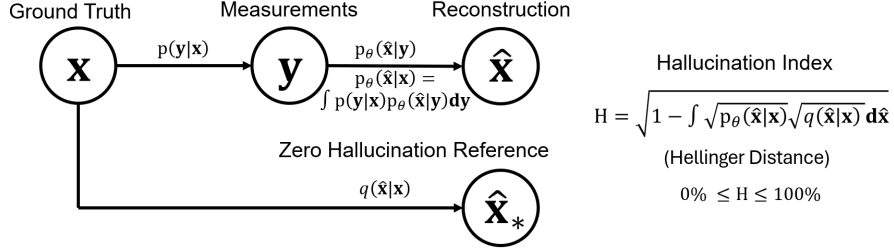
2   Theoretical Methods

Fig. 1. Probabilistic graphical model of true images, measurements, and reconstructions. The hallucination index is defined as the Hellinger distance between the reconstruction distribution and a zero-hallucination reference distribution. In diffusion models, the forward process can be used as the zero hallucination reference.

2.1   Hallucination Index

A probabilistic model of image reconstruction is shown in Figure 1. We consider the true images, $\mathbf{x}$, the measurements, $\mathbf{y}$, and the reconstructed images, $\hat{\mathbf{x}}$, as flattened random vectors. The measurements are a stochastic function of the true images, related by the conditional probability density function $p(\mathbf{y}|\mathbf{x})$. Classical reconstruction methods are often deterministic functions of the measurements, including linear deterministic methods like pseudo-inverse reconstruction or nonlinear deterministic methods like pre-trained convolutional neural networks. However, there is increasing interest in generative reconstruction methods, such as diffusion models, which are a stochastic function of the measurements. Therefore, we model the reconstruction process with the conditional probability distribution $p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{y})$, dependent on some parameter vector, $\boldsymbol{\theta}$. The joint probability density function is

$$p(\mathbf{x}, \mathbf{y}, \hat{\mathbf{x}}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{y}). \tag{1}$$

Therefore, an end-to-end conditional probability density function of reconstructions given true images is given by

$$p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x})p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{y})\mathbf{dy} \tag{2}$$

This distribution of reconstructions can contain various types of error such as bias (e.g. spatial blur), noise, and hallucinations. In the context of image reconstruction, hallucinations are generally understood to refer to variability or uncertainty in the structure of an object arising from the application of a machine learning based image processing model. For example, if the measurement signal-to-noise ratio (SNR) is low, and a generative image reconstruction method is repeatedly applied to produce multiple samples of output images with the appearance of high SNR, this will typically result in highly varied estimates of the object underlying the noise.

The key to our approach is to define what hallucinations are not. We claim that when no machine learning model has been applied, there are no hallucinations. For example, if white noise is added to the input, there will be increased variance in the output, but that error is not a hallucination by definition because no machine learning model has been applied. Therefore, we introduce a new distribution, $q(\hat{\mathbf{x}}|\mathbf{x})$ which we define as a zero hallucination reference with the same apparent signal-to-noise ratio as the generative reconstructions $p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x})$. In the case of diffusion models, this zero hallucination reference is available via the forward diffusion process. We define the *hallucination index* as the Hellinger distance between the distribution of reconstructed images and the zero hallucination reference distribution as shown

$$H\big(p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x}), q(\hat{\mathbf{x}}|\mathbf{x})\big) = \sqrt{1 - \int \sqrt{p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x})}\sqrt{q(\hat{\mathbf{x}}|\mathbf{x})}\mathbf{d}\hat{\mathbf{x}}} \qquad (3)$$

This metric has several desirable mathematical properties. First of all, it is bounded by $0 \leq H \leq 1$ with equality to zero when the two distributions are identical and equality to one when the two distributions have no overlap. Loosely, the hallucination index can be interpreted as the percentage of error due to hallucinations as opposed to noise. The Hellinger distance is a metric that satisfies the triangle inequality which could be useful for analyzing a cascade of multiple data processing steps. Rearranging terms in (3) results in

$$H\big(p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x}), q(\hat{\mathbf{x}}|\mathbf{x})\big) = \sqrt{1 - \int p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x})\frac{\sqrt{q(\hat{\mathbf{x}}|\mathbf{x})}}{\sqrt{p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x})}}\mathbf{d}\hat{\mathbf{x}}} \qquad (4)$$

The form of (4) now includes an expectation over $p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x})$. This mean can be approximated by the sample mean, and the probability densities can be approximated by various methods such as kernel density estimation. In this work, we chose to use a simpler approach where the hallucination index is approximated by the Hellinger distance between two multivariate normal distributions with the same mean and covariance as $p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}|\mathbf{x})$ and $q(\hat{\mathbf{x}}|\mathbf{x})$ using the closed form solution

$$H \approx \sqrt{1 - \exp\Big(-\frac{1}{4}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q)^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) + c\Big)} \qquad (5)$$

$$c = \frac{1}{4}\log|\boldsymbol{\Sigma}_p| + \frac{1}{4}\log|\boldsymbol{\Sigma}_q| - \frac{1}{4}\log|\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q|$$

We approximate the covariance matrices by a circulant covariance matrix with the same noise power spectrum. So, in practice we compute the hallucination index by 1) taking multiple samples of the measurements 2) applying the generative reconstruction model 3) computing the sample mean of the reconstructions 4) compute sample noise power spectrum of the reconstruction distribution and the zero hallucination reference distribution 5) compute the hallucination index using the formula above.

2.2   Fourier Diffusion Models

We followed the methods in [17] to implement Fourier diffusion models. The main difference with respect to most diffusion models is that the scalar multiplication and additive white Gaussian noise in the forward stochastic process is generalized to include convolutional operators and additive stationary Gaussian noise. The practical benefit of this method is that we can construct a diffusion bridge model starting from ground truth images at $t = 0$ and converging to the same noise power spectrum as the measurements. At inference time, the reverse process can be initialized at $t = t_{\text{start}}$ directly from the measurements. Fourier diffusion models are based on a forward stochastic process defined by

$$\mathbf{x}_t = \mathbf{H}_t \mathbf{x}_0 + \boldsymbol{\Sigma}_t^{1/2} \frac{1}{\sqrt{t}} \mathbf{w}_t, \quad t > 0 \tag{6}$$

where $\mathbf{H}_t$ is a invertible circulant matrix operator representing the modulation transfer function in the spatial frequency domain, $\boldsymbol{\Sigma}_t$ is a circulant covariance matrix of spatially stationary Gaussian noise, and $\mathbf{w}_t$ is a standard Wiener process. The corresponding forward diffusion stochastic differential equation is

$$\mathbf{dx}_t = \mathbf{F}_t \mathbf{x}_t dt + \mathbf{G}_t \mathbf{dw}_t, \ \ \mathbf{F}_t = \mathbf{H}_t' \mathbf{H}_t^{-1}, \ \ \mathbf{G}_t = [\boldsymbol{\Sigma}_t' - \mathbf{F}_t \boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t \mathbf{F}_t^T]^{1/2} \tag{7}$$

Following the framework for more general stochastic differential equations in Appendix A of [14], we can define the reverse stochastic differential equation as

$$\mathbf{dx}_t = [\mathbf{F}_t \mathbf{x}_t - \frac{1}{2} \mathbf{G}_t \mathbf{G}_t^T \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)] dt + \mathbf{G}_t \mathbf{d}\bar{\mathbf{w}}_t. \tag{8}$$

The generative model is implemented by approximating the score function with a score matching neural network, $\mathbf{s}_\theta(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$. We can sample from the reverse process, $p_\theta(\mathbf{x}_t | \mathbf{x}_{t_{\text{start}}} = \mathbf{y})$ using stochastic integral sampling methods such as the Euler-Maruyama method applied to (8). For hallucination analysis, this will be compared to the forward process, $q(\mathbf{x}_t | \mathbf{x}_0) \sim \mathcal{N}(\mathbf{H}_t \mathbf{x}_0, \boldsymbol{\Sigma}_t)$, which is a zero-hallucination reference with the same image quality in terms of modulation transfer function and noise power spectrum.

3   Experimental Methods

We conducted a large scale numerical experiment to evaluate hallucinations in Fourier diffusion models with various configurations. The purpose of this experiment is twofold. First, we seek to validate our proposed metric to verify that it corresponds to what is commonly understood as hallucinations by visual inspection. Second, we seek to quantify the relationship between hallucination index and other measurable parameters and metrics such as input SNR, reverse process early stop time, noise power spectrum, and mean squared error.

The experiment simulates the task of imaging cortical neurons through electron microscopy, particularly under conditions where the images are affected by low SNR due to faster scanning. We used the Machine Intelligence from Cortical Networks (MICrONS) dataset to model the ground truth images [6]. We used 10,000 image samples of image size 256x256 for training, 1,000 samples for validation, and 1,000 samples for evaluation.

For each evaluation case, we simulated 64 measurement noise realizations. The measurement model is the ground truth images plus additive stationary Gaussian noise with noise power spectrum, $\boldsymbol{\Sigma}_t$, as shown in Figure 3. For the Fourier diffusion model, we did not include any spatial blur, so $\mathbf{H}_t = \mathbf{I}$, and the noise covariance increases linearly from zero noise to the same noise power spectrum as the measurements. This is similar to the variance-exploding stochastic process defined in [14] but with spatially correlated noise rather than white noise.

For training the Fourier diffusion model, we used the exact same methods described in [17] including score matching neural network architecture and score matching loss function. We used 1,000 training epochs, 100 iterations per epoch, 32 images per batch, and the Adam optimizer with learning rate 10^{-3}. After initializing with measurements, we sample the reverse process with the Euler-Maruyama method with 100 equally spaced time steps. We repeated the reverse diffusion process sampling for each evaluation sample. An example of multiple reverse process samples is shown in Figure 3. We also repeated all experiments for 5 different measurement signal-to-noise ratio levels. This was accomplished by initializing the reverse process with measurement simulations corresponding to the intermediate noise power spectrum at t_{start} values of 1.0, 0.8, 0.6, 0.4, and 0.2. When running the reverse diffusion sampling, we saved all the intermediate time step results. All together, this forms a large scale dataset that we used for image quality evaluation. Each reconstruction distribution in the dataset
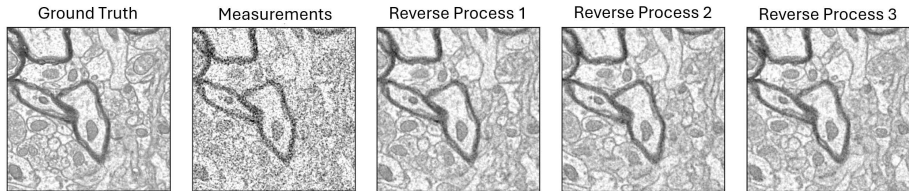

| Ground Truth | Measurements | Reverse Process 1 | Reverse Process 2 | Reverse Process 3 |

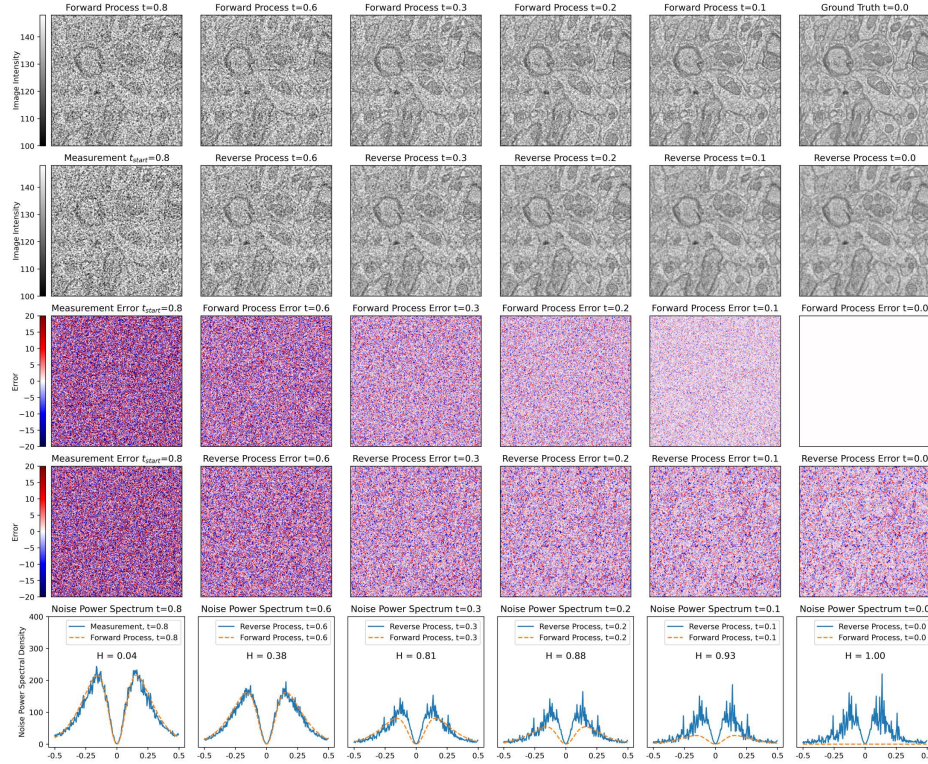**Fig. 2.** Ground truth, measurements, and reverse process samples with hallucinations.

Fig. 3. Comparison between forward and reverse process samples. Hallucination index increases for more reverse diffusion steps because there is increased Hellinger distance between forward and reverse process.

represents 64 samples from $p_\theta(\hat{\mathbf{x}}|\mathbf{x})$. We computed the mean, bias, variance, mean squared error, and noise power spectrum. We computed the hallucination index as a function of bias and noise power spectrum using (5).

4 Results

Figure 3 shows a comparison of the forward and reverse process. The forward process does not have hallucinations, it is ground truth plus noise. The reverse process has hallucinations introduced by the diffusion model. The error images show increasing divergence between the reverse process and the forward process with more reverse diffusion steps. Visually, the error images for the reverse process error near $t = 0$ show increasing intensity of structural patterns. Hallucination is increasing with more reverse diffusion steps because there is increased Hellinger distance between the forward and reverse process distributions.

Figure 4 shows a quantitative analysis of mean squared error and hallucination index as a function of measurement SNR and reverse diffusion time steps.

Both mean squared error and hallucination index are reduced if measurement SNR is increased. However, this is not always an option as it requires improvements to the imaging data acquisition. Mean squared error can also be reduced by running more iterations of a reverse diffusion process. However, we see that hallucination index increases with more reverse diffusion time steps. Overall there is a tradeoff between mean squared error and hallucination index. That is, the more aggressive the denoising applied by a generative reconstruction model, the more it hallucinates.
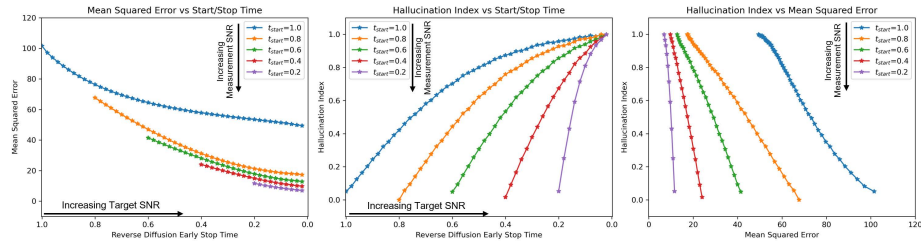

Fig. 4. Image quality metrics as a function of input SNR and reverse process time step. Increasing measurement SNR decreases both mean squared error and hallucinations. Increasing reverse process time steps reduces mean squared error but increases hallucinations. Third plot shows tradeoff between mean squared error and hallucinations.

5   Conclusion

Our results demonstrate how the proposed hallucination index can be useful for identifying and quantifying generative artifacts that are not sufficiently described by mean squared error. However, our approach is not without its limitations. Notably, we assumed that the reconstruction distribution is Gaussian and shift-invariant. While these approximations facilitate computation, they may not fully capture the complexity of the error distributions. We also acknowledge that we have only tested our method on Fourier diffusion models, which form a direct diffusion bridge from measured images to noiseless images. One unique aspect of these models is the capacity for early stopping. This property was critical for our work because we used the forward process at the same intermediate time as the zero hallucination reference. This highlights a limitation of our approach that for generative models purporting to synthesize noiseless images given noisy inputs, the hallucination index will invariably evaluate to 100%, due to the fact that noiseless images are given by a deterministic Dirac delta distribution. However, we note that 100% hallucination index does accurately reflect the percent of covariance that is attributable to hallucinations as opposed to noise, for cases purporting to be a noiseless reconstruction.

Our approach relies on a well-defined zero hallucination reference distribution, and we use the Hellinger distance to measure the distance to the distribution

of reconstructions. In the future, we are interested to investigate other options for the zero hallucination reference distribution for cases when the forward diffusion process may not be available. We can also investigate alternative distance metrics such as KL divergence or other f-divergences in the future.

Looking forward, our goal is to refine and apply hallucination analysis methods. We want to explore optimized generative reconstructions for minimal hallucinations. We are also interested to compare various diffusion-based reconstruction methods and evaluate their relative susceptibility to hallucinations.

In conclusion, the analysis of hallucinations in medical imaging is of paramount importance, especially as generative reconstruction models become more prevalent. The proposed Hallucination Index represents one option for an quantitative metric for image quality assessment, offering a tool to navigate the complexities of enhancing image quality while minimizing the introduction of potentially misleading generative artifacts. As we continue to refine this metric and apply it to a broader range of models and applications, we anticipate valuable insights that will contribute to the ongoing improvement of medical image reconstruction.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Barbano, R., Denker, A., Chung, H., Roh, T.H., Arrdige, S., Maass, P., Jin, B., Ye, J.C.: Steerable conditional diffusion for out-of-distribution adaptation in imaging inverse problems. arXiv preprint arXiv:2308.14409 (2023)
2. Bhadra, S., Kelkar, V.A., Brooks, F.J., Anastasio, M.A.: On hallucinations in tomographic image reconstruction. IEEE transactions on medical imaging **40**(11), 3249–3260 (2021)
3. Buban, J.P., Ramasse, Q., Gipson, B., Browning, N.D., Stahlberg, H.: High-resolution low-dose scanning transmission electron microscopy. Journal of electron microscopy **59**(2), 103–112 (2010)
4. Chu, L.C., Anandkumar, A., Shin, H.C., Fishman, E.K.: The potential dangers of artificial intelligence for radiology and radiologists. Journal of the American College of Radiology **17**(10), 1309–1311 (2020)
5. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. pp. 529–536. Springer (2018)
6. Consortium, M., Bae, J.A., Baptiste, M., Bishop, C.A., Bodor, A.L., Brittain, D., Buchanan, J., Bumbarger, D.J., Castro, M.A., Celii, B., et al.: Functional connectomics spanning multiple areas of mouse visual cortex. BioRxiv pp. 2021–07 (2021)
7. Denker, A., Schmidt, M., Leuschner, J., Maass, P., Behrmann, J.: Conditional normalizing flows for low-dose computed tomography image reconstruction. arXiv preprint arXiv:2006.06270 (2020)
8. Hajij, M., Zamzmi, G., Paul, R., Thukar, L.: Normalizing flow for synthetic medical images generation. In: 2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT). pp. 46–49. IEEE (2022)

9. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis p. 102846 (2023)
10. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., et al.: Denoising diffusion probabilistic models for 3d medical image generation. Scientific Reports **13**(1), 7303 (2023)
11. Mardani, M., Gong, E., Cheng, J.Y., Vasanawala, S.S., Zaharchuk, G., Xing, L., Pauly, J.M.: Deep generative adversarial neural networks for compressive sensing mri. IEEE transactions on medical imaging **38**(1), 167–179 (2018)
12. Nikulin, M.S., et al.: Hellinger distance. Encyclopedia of mathematics **78** (2001)
13. Song, Y., Shen, L., Xing, L., Ermon, S.: Solving inverse problems in medical imaging with score-based generative models. arXiv preprint arXiv:2111.08005 (2021)
14. 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)
15. Suganthi, K., et al.: Review of medical image synthesis using gan techniques. In: ITM Web of Conferences. vol. 37, p. 01005. EDP Sciences (2021)
16. Teneggi, J., Tivnan, M., Stayman, W., Sulam, J.: How to trust your diffusion model: A convex optimization approach to conformal risk control. In: International Conference on Machine Learning. pp. 33940–33960. PMLR (2023)
17. Tivnan, M., Teneggi, J., Lee, T.C., Zhang, R., Boedeker, K., Cai, L., Gang, G.J., Sulam, J., Stayman, J.W.: Fourier diffusion models: A method to control mtf and nps in score-based stochastic image generation. arXiv preprint arXiv:2303.13285 (2023)
18. Trampert, P., Bourghorbel, F., Potocek, P., Peemen, M., Schlinkmann, C., Dahmen, T., Slusallek, P.: How should a fixed budget of dwell time be spent in scanning electron microscopy to optimize image quality? Ultramicroscopy **191**, 11–17 (2018)
19. Xie, Y., Li, Q.: Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–664. Springer (2022)
20. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical image analysis **58**, 101552 (2019)
21. Zhou, T., Li, Q., Lu, H., Cheng, Q., Zhang, X.: Gan review: Models and medical image fusion applications. Information Fusion **91**, 134–148 (2023)