GSURE-Based Diffusion Model Training with Corrupted Data

Bahjat Kawar [*1] **Noam Elata** [*2] **Tomer Michaeli** [2] **Michael Elad** [1]

Abstract

Diffusion models have demonstrated impressive results in both data generation and downstream tasks such as inverse problems, text-based editing, classification, and more. However, training such models usually requires large amounts of clean signals which are often difficult or impossible to obtain. In this work, we propose a novel training technique for generative diffusion models based only on corrupted data. We introduce a loss function based on the Generalized Stein's Unbiased Risk Estimator (GSURE), and prove that under some conditions, it is equivalent to the training objective used in fully supervised diffusion models. We demonstrate our technique on face images as well as Magnetic Resonance Imaging (MRI), where the use of undersampled data significantly alleviates data collection costs. Our approach achieves generative performance comparable to its fully supervised counterpart without training on any clean signals. In addition, we deploy the resulting diffusion model in various downstream tasks beyond the degradation present in the training set, showcasing promising results.

1. Introduction

Denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019), or diffusion models for short, are a family of generative models that has recently risen to prominence. They have achieved state-of-the-art performance in image generation (Song et al., 2020b; Vahdat et al., 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022; Kim et al., 2022), as well as impressive generative modeling capabilities in other modalities (Singer et al., 2023; Kong et al., 2021; Gong et al., 2023; Tevet et al., 2022; Watson et al., 2022; Song et al., 2023; Chung & Ye, 2022; Jalal et al., 2021).

Training a diffusion model to learn an unknown data distribution is a complex task. It usually requires training parameter-heavy neural networks on large amounts of pristine data. For instance, diffusion models' success in image generation was in part enabled by large curated datasets, containing millions or even billions of images (Deng et al., 2009; Schuhmann et al., 2022). However, such large-scale datasets of pristine samples may often be expensive, difficult, or even impossible to obtain, especially in the medical domain (Mullainathan & Obermeyer, 2022). In this work, we present GSURE-Diffusion, a method for training generative diffusion models based on data corrupted by linear degradations and Gaussian noise. This can make data collection for deep learning significantly faster and cheaper.

GSURE-Diffusion operates on a datasets of noisy linear measurements of signals, and assumes the signal acquisition process is randomized within a fixed general structure, which is the case in many real-world applications. In this setting, we present a novel loss function to learn the underlying data distribution. First, we use the Singular Value Decomposition (SVD) of the degradation operators to decouple the measurement equation, following DDRM (Kawar et al., 2022). Then, we add synthetic noise to the SVD-transformed measurements, likening them to the noisy samples used in the DDPM (Ho et al., 2020) framework. Finally, we use the ensemble version of the Generalized Stein's Unbiased Risk Estimator (GSURE) (Aggarwal et al., 2022; Eldar, 2008) to learn to denoise samples without access to ground-truth clean signals. Our proposed GSURE-based loss function is general to all randomized linear measurement settings, and we prove its equivalence to the fully supervised denoising diffusion loss under some conditions.

We apply our technique on face images as well as Magnetic Resonance Imaging (MRI). After training solely on corrupted data, our GSURE-Diffusion models exhibit comparable generative performance to oracle model counterparts, which train on pristine data. Furthermore, we demonstrate the capabilities of our trained models by deploying them in various downstream tasks.

We hope that GSURE-Diffusion will facilitate future work on generative modelling for challenging settings, generaliz-

[*]Equal contribution [1]Department of Computer Science, Technion, Haifa, Israel [2]Department of Electrical and Computer Engineering, Technion, Haifa, Israel. Correspondence to: Bahjat Kawar <bahjat.kawar@cs.technion.ac.il>, Noam Elata <noamelata@campus.technion.ac.il>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling 2023, Honolulu, Hawaii, USA.

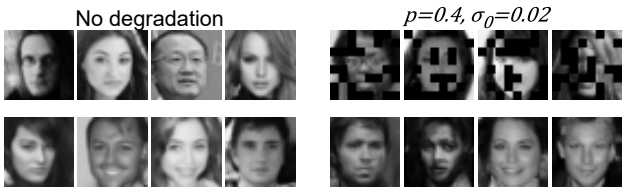

Figure 1. Training set (top) and generated (bottom) samples of different degradation settings in CelebA experiments.

ing for more complex scenarios and various modalities.

2. Background

2.1. Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) are a family of generative models that learn a distribution $p_\theta(\mathbf{x})$, approximating a data distribution $q(\mathbf{x})$ from a dataset $\mathcal{D}$ of samples. DDPMs follow a Markov chain structure $\mathbf{x}_T \to \mathbf{x}_{T-1} \to \cdots \to \mathbf{x}_1 \to \mathbf{x}_0$ that reverses a forward noising process from $\mathbf{x}_0$ to $\mathbf{x}_T$. In the forward process, $\mathbf{x}_0$ is set to be $\mathbf{x}$, and the intermediate variables $\mathbf{x}_t$ are defined by $q^{(t)}(\mathbf{x}_t|\mathbf{x}_{t-1})$, usually chosen to be Gaussian $\mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right)$. This leads to a useful property, $q^*(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\,\mathbf{I}\right)$ with $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_t)$, which facilitates model training. In the reverse process, the learned distribution $p_\theta^{(t)}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is also modeled as Gaussian, with a learned mean dependent on a neural network $f_\theta^{(t)}(\mathbf{x}_t)$ and a fixed (Ho et al., 2020) or learned (Nichol & Dhariwal, 2021) covariance.

The diffusion model $f_\theta^{(t)}(\mathbf{x}_t)$ is trained to optimize an evidence lower bound (ELBO) on the likelihood objective (Sohl-Dickstein et al., 2015). The ELBO can be simplified into the following denoising objective:

$$\sum_{t=1}^{T}\gamma_t\mathbb{E}\left[\left\|f_\theta^{(t)}(\mathbf{x}_t)-\mathbf{x}_0\right\|_2^2\right], \tag{1}$$

where the γ_t assign weights to different t (different noise levels), and the expectation is over $\mathbf{x}_t \sim q^*(\mathbf{x}_t|\mathbf{x}_0)$, $\mathbf{x}_0 \sim q(\mathbf{x})$. Please refer to (Ho et al., 2020; Song et al., 2020a) for derivations. After training, diffusion models synthesize data by starting with a sample $\mathbf{x}_T \sim \mathcal{N}(0,\mathbf{I})$, following the learned distributions $p_\theta^{(t)}$ along the Markov chain, sampling from each, and outputting $\mathbf{x}_0$ as the final sample. In this work, we seek a way to train DDPMs using only corrupted data.

2.2. Generalized Stein's Unbiased Risk Estimator

Given noisy measurements $\mathbf{y} = \mathbf{x}+\mathbf{z}$ (where $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathbb{R}^n$) with noise $\mathbf{z} \sim \mathcal{N}(0,\sigma^2\mathbf{I})$, and a function $f(\mathbf{y})$ aiming to estimate $\mathbf{x}$ from $\mathbf{y}$, Stein's unbiased risk estimator (SURE) (Stein, 1981) is an unbiased estimator for the mean squared error (MSE) of $f(\mathbf{y})$. Crucially, SURE provides the ability to estimate the MSE of a denoiser $f(\mathbf{y})$ without access to clean signals $\mathbf{x}$.

In the context of inverse problems, SURE has been generalized for corrupted measurements beyond additive white Gaussian noise (Eldar, 2008). The Generalized SURE (GSURE) considers $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{z}$ (where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{H} \in \mathbb{R}^{m\times n}$, $\mathbf{y},\mathbf{z} \in \mathbb{R}^m$, and $\mathbf{z} \sim \mathcal{N}(0,\mathbf{C})$), and a function $f(\mathbf{y})$ estimating $\mathbf{x}$. In this case, GSURE provides an unbiased estimate for the projected MSE $\mathbb{E}\left[\|\mathbf{P}\left(f(\mathbf{y})-\mathbf{x}\right)\|_2^2\right]$:

$$\mathbb{E}\left[\|\mathbf{P}\left(f(\mathbf{y})-\mathbf{x}_{\mathrm{ML}}\right)\|_2^2\right]+2\mathbb{E}\left[\nabla_{\mathbf{H}^\top\mathbf{C}^{-1}\mathbf{y}}\cdot\mathbf{P}f(\mathbf{y})\right]+c, \tag{2}$$

where $\mathbf{H}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{H}$, $\mathbf{P} = \mathbf{H}^\dagger\mathbf{H}$ is a projection matrix onto the range-space of $\mathbf{H}$, $\mathbf{x}_{\mathrm{ML}} = \left(\mathbf{H}^\top\mathbf{C}^{-1}\mathbf{H}\right)^\dagger\mathbf{H}^\top\mathbf{C}^{-1}\mathbf{y}$, and c is a constant that does not depend on $f(\mathbf{y})$. GSURE has been utilized for solving inverse problems by training only on corrupted measurements (Metzler et al., 2018). However, when $\mathbf{H}$ causes significant information loss, the projected MSE stops being a good proxy for the full MSE. Ensemble SURE (ENSURE) (Aggarwal et al., 2022) learns from a dataset of measurements, each corrupted by a different operator $\mathbf{H}$. Therefore, the expectation over the projected MSE is taken over $\mathbf{H}$ as well as the data and noise. This constitutes a more accurate proxy for the full MSE without relying on clean signals. In this work, we extend ENSURE for training a diffusion model using corrupted data.

3. GSURE-Diffusion Formulation

3.1. Problem Formulation

We are interested in training a generative diffusion model that can sample from an unknown data distribution $q(\mathbf{x})$. However, we only have access to a dataset $\mathcal{D}$ of corrupted measurements $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{z}$, where $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{H} \in \mathbb{R}^{m\times n}$, and $\mathbf{z} \sim \mathcal{N}(0,\sigma_0^2\mathbf{I})$ is additive white Gaussian noise (AWGN).[1] $\mathbf{x}$ and $\mathbf{y}$ represent a single instance of an ideal image and its corresponding measurement. More generally, different measurements $\mathbf{y}$ in the dataset may relate to different signals $\mathbf{x}$, different degradation procedures $\mathbf{H}$, and different noise realizations $\mathbf{z}$. We assume $\mathbf{x}$, $\mathbf{z}$, and $\mathbf{H}$ are independently sampled from their respective distributions.

To decouple the relationship between the observed measurements and the underlying data, we follow (Kawar et al., 2022) and utilize the singular value decomposition (SVD) $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ is a rectangular diagonal matrix containing the singular values of $\mathbf{H}$. We define

[1] Our method can also handle anisotropic uncorrelated noise. We only consider AWGN to simplify notations.

$\bar{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$, $\bar{\mathbf{y}} = \mathbf{\Sigma}^\dagger \mathbf{U}^\top \mathbf{y}$, and $\bar{\mathbf{z}} = \mathbf{\Sigma}^\dagger \mathbf{U}^\top \mathbf{z}$. Using these definitions, the measurement equation becomes

$$\bar{\mathbf{y}} = \mathbf{P}\bar{\mathbf{x}} + \bar{\mathbf{z}}, \tag{3}$$

where $\mathbf{P} = \mathbf{\Sigma}^\dagger \mathbf{\Sigma}$ is a diagonal subsampling matrix with zeroes and ones, and $\bar{\mathbf{z}} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{\Sigma}^\dagger \mathbf{\Sigma}^{\dagger\top})$ constitutes anisotropic uncorrelated Gaussian noise.

We make the following assumptions on the training dataset: (i) The sampling matrices $\mathbf{H}$ and noise levels σ_0 are known; (ii) All matrices $\mathbf{H}$ share the same left-singular vectors $\mathbf{V}^\top$; and (iii) The different $\mathbf{H}$ across the dataset jointly cover the signal space $\mathbb{R}^n$, *i.e.*, $\mathbb{E}[\mathbf{P}]$ is positive definite.[2] These assumptions are satisfied in many real-world applications such as Magnetic Resonance Imaging (MRI).

Under the transformed measurement equation presented in Equation 3, we aim to train a generative model for $\bar{\mathbf{x}}$, which can easily translate to $\mathbf{x}$ using $\mathbf{x} = \mathbf{V}\bar{\mathbf{x}}$.

3.2. GSURE-Based Denoising Diffusion Loss Function

To train a diffusion model for $\bar{\mathbf{x}}$, we aim to obtain noisy training samples $\bar{\mathbf{x}}_t$ that satisfy the marginal distribution $q^*(\bar{\mathbf{x}}_t|\bar{\mathbf{x}}) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\bar{\mathbf{x}}, (1-\bar{\alpha}_t)\,\mathbf{I}\right)$, as in traditional diffusion models. However, we only have access to corrupted measurements $\bar{\mathbf{y}}$ as in Equation 3. For a given t, we perturb these measurements with additional noise according to

$$\bar{\mathbf{x}}_t = \sqrt{\bar{\alpha}_t}\bar{\mathbf{y}} + \left((1-\bar{\alpha}_t)\,\mathbf{I} - \bar{\alpha}_t \sigma_0^2 \mathbf{\Sigma}^\dagger \mathbf{\Sigma}^{\dagger\top}\right)^{\frac{1}{2}} \boldsymbol{\epsilon}_t, \tag{4}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ is independently sampled. Intuitively, $\sqrt{\bar{\alpha}_t}\bar{\mathbf{y}}$ includes noise with a diagonal covariance $\bar{\alpha}_t \sigma_0^2 \mathbf{\Sigma}^\dagger \mathbf{\Sigma}^{\dagger\top}$. We increase the noise level in each entry by an appropriate amount to reach a variance of $1-\bar{\alpha}_t$ in all entries. This way, we obtain samples $\bar{\mathbf{x}}_t$ suitable for training a diffusion model, as they follow the marginal distribution

$$q(\bar{\mathbf{x}}_t|\bar{\mathbf{x}}, \mathbf{P}) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{P}\bar{\mathbf{x}}, (1-\bar{\alpha}_t)\,\mathbf{I}\right). \tag{5}$$

This resembles the ideal distribution of training samples $q^*(\bar{\mathbf{x}}_t|\bar{\mathbf{x}})$, differing only in the mean value for entries dropped by $\mathbf{P}$. In the following, we derive a loss function that uses $\bar{\mathbf{x}}_t$ satisfying Equation 5, and utilizes an expectation over $\bar{\mathbf{x}}$, $\bar{\mathbf{z}}$, and $\mathbf{P}$. This results in an estimate for denoising ideal samples from $q^*(\bar{\mathbf{x}}_t|\bar{\mathbf{x}})$. The estimate assumes the ability of the trained neural network to infer $\mathbf{P}$ from $\bar{\mathbf{x}}_t$, and to generalize for $\mathbf{P} = \mathbf{I}$ despite having trained only on signals with an undersampled $\mathbf{P}$.

Ideally, we would like to train a diffusion model $f_\theta^{(t)}(\bar{\mathbf{x}}_t)$ using the traditional denoising diffusion loss function in Equation 1. However, as we only have access to undersampled measurements, we consider a weighted expected projected MSE objective (similar to Aggarwal et al. (2022)):

$$\sum_{t=1}^{T} \gamma_t \mathbb{E}\left[\left\|\mathbf{W}\mathbf{P}\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\right\|_2^2\right], \tag{6}$$

where the expectation is taken over $\bar{\mathbf{x}}_t \sim q(\bar{\mathbf{x}}_t|\bar{\mathbf{x}}, \mathbf{P})$, $\bar{\mathbf{x}} \sim q(\bar{\mathbf{x}})$, and $\mathbf{P}$ is independently sampled. In practice, this expectation is realized through $\bar{\mathbf{y}}$ sampled from the dataset $\mathcal{D}$ and Equation 4.

Proposition 3.1. *For $\mathbf{x}$ sampled from an unknown $q(\mathbf{x})$, $\bar{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$, $\bar{\mathbf{x}}_t$ sampled from Equation 5, and the diagonal weight matrix $\mathbf{W} = \mathbb{E}[\mathbf{P}]^{-\frac{1}{2}} \succ 0$ (positive definite), it holds that Equation 6 approximately equals Equation 1.*

We place the proof in Appendix A. The proof relies on the aforementioned assumptions on the neural network's generalization abilities, as well as the assumption that $\mathbf{P}$ and the network's MSE are statistically independent, as assumed in ENSURE (Aggarwal et al., 2022). The expected projected MSE term in Equation 6 measures the squared error of the denoiser $f_\theta^{(t)}(\bar{\mathbf{x}}_t)$ only in entries kept by $\mathbf{P}$. This fact makes the loss easier to measure, as we do not have access to the entries dropped by $\mathbf{P}$. However, we still cannot accurately measure this loss, because we lack access to noiseless signals $\mathbf{P}\bar{\mathbf{x}}$. To mitigate this, we utilize GSURE to estimate Equation 6 using only $\bar{\mathbf{x}}_t$ with the loss

$$\sum_{t=1}^{T} \gamma_t \mathbb{E}\left[\left\|\mathbf{W}\mathbf{P}\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}\bar{\mathbf{x}}_t\right)\right\|_2^2 + 2\lambda_t\left(\nabla_{\bar{\mathbf{x}}_t} \cdot \mathbf{P}\mathbf{W}^2 f_\theta^{(t)}(\bar{\mathbf{x}}_t)\right) + c\right], \tag{7}$$

where c is a constant that does not depend on θ, and λ_t is a scalar hyperparameter. The expectation is over the same random variables from Equation 6.

Proposition 3.2. *For $\mathbf{x}$ sampled from an unknown $q(\mathbf{x})$, $\bar{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$, $\bar{\mathbf{x}}_t$ sampled from Equation 5, and $\lambda_t = 1 - \bar{\alpha}_t$, it holds that Equation 7 equals Equation 6.*

We place the proof in Appendix A. Proposition 3.1 and Proposition 3.2 present a principled method to train a denoising diffusion model based only on corrupted data $\bar{\mathbf{y}}$. By minimizing the loss function in Equation 7, we obtain a trained diffusion model that can be utilized in the same fashion as a fully supervised one.

When our proposed training scheme is applied in practice, the expectation in Equation 7 is replaced by an average over training batches, $\frac{1}{\sqrt{\bar{\alpha}_t}}\bar{\mathbf{x}}_t$ is replaced by $\bar{\mathbf{y}}$ to alleviate the high variance of the loss function, and the divergence term is calculated using an unbiased Monte Carlo estimator (Ramani et al., 2008; Hutchinson, 1989). We defer these pragmatic implementation details to Appendix E.

[2] Under these notations, $\mathbb{E}[\mathbf{P}]$ is measured for a fixed $\mathbf{V}^\top$, and the values in $\mathbf{\Sigma}$ are not necessarily ordered.

Table 1. FID results for diffusion models trained on increasing levels of degradation for CelebA, with different DDIM steps.

DEGRADATION / STEPS	10	20	50	100
NONE (ORACLE)	21.99	13.09	08.15	06.84
$p = 0.2$, $\sigma_0 = 0.01$	18.77	12.25	08.84	08.82
$p = 0.4$, $\sigma_0 = 0.02$	19.26	14.98	14.03	15.14
$p = 0.6$, $\sigma_0 = 0.03$	34.51	27.74	26.42	28.31

4. Experiments

In the following, we demonstrate the capabilities of our method for training a diffusion model using corrupted data. To obtain corrupted data, we simulate several corruptions on datasets containing clean images. Then, we train a diffusion model based only on the corrupted data, and compare its results against an *oracle* model (with identical training hyperparameters) which is trained on the pristine data with the traditional diffusion loss function from Equation 1. We provide architecture and training details in Appendix D.

Face Images. We apply GSURE-Diffusion on a 32×32-pixel grayscale variant of CelebA (Liu et al., 2015). We simulate a corrupted measurement process by splitting the images into 4×4-pixel non-overlapping patches, randomly erasing each patch with probability p, and adding AWGN with standard deviation σ_0. This degradation matches our assumptions in subsection 3.1 (see Appendix C). We train diffusion models for increasing levels of degradation. After training, we generate images from the models using DDIM (Song et al., 2020a). We measure the generative performance using the FID (Heusel et al., 2017) between 10000 generated images and the CelebA validation set. As can be seen in Table 1 and Figure 1, our GSURE-Diffusion models achieve generative performance comparable to the oracle model, despite having trained only on corrupted data.

Magnetic Resonance Imaging (MRI). MRI is a ubiquitous non-invasive medical imaging modality that can provide life-saving diagnostic information. MRI measurements are obtained in the Fourier spectrum (also called k-space) of an object with magnetic fields. However, since measuring the entire k-space can be time-consuming and expensive, MRI scans are often accelerated, resulting in randomly partial and possibly noisy k-space measurements. The accelerated MRI procedure satisfies the assumptions we introduced in subsection 3.1 (see Appendix C). We use this fact and train an MRI diffusion model based solely on accelerated scans. We train on the fastMRI (Knoll et al., 2020; Zbontar et al., 2019) single-coil knee MRI dataset, center-cropped to 320×320. The randomized accelerated MRI subsampling process is simulated following (Jalal et al., 2021) with an acceleration factor $R = 4$ and AWGN with $\sigma_0 = 0.01$.

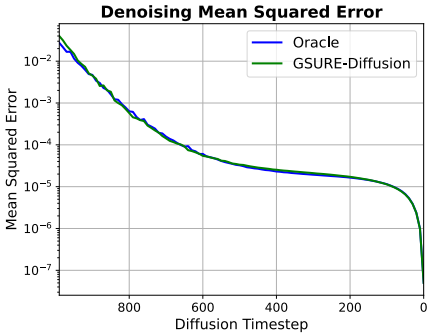

Figure 2. Denoising MSE (on fully sampled noisy images) for the GSURE-Diffusion and oracle models across diffusion timesteps.

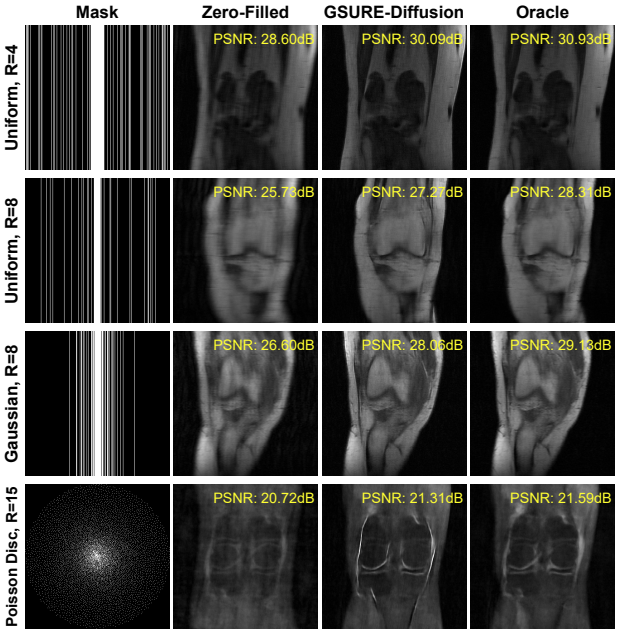

Figure 3. Accelerated MRI reconstruction results with different subsampling masks and $\sigma_0 = 0.01$.

We measure the denoising MSE of GSURE-Diffusion and the oracle model for fully sampled images. Both models perform very similarly (see Figure 2) despite the fact that GSURE-Diffusion trained solely on undersampled noisy data. We then use our model for MRI reconstruction (using DDRM (Kawar et al., 2022) with $\eta = 0$ and 100 steps) for different subsampling masks and acceleration factors, and show the results in Figure 3. We further show additional experiments in Appendix F. These results constitute evidence that a generative model trained on corrupted data can be deployed in various applications. By loosening the requirements on the quality of the training data, we significantly reduce the cost of data acquisition for model training.

Acknowledgements

Data used in the preparation of this article were obtained from the NYU fastMRI Initiative database (Zbontar et al., 2019; Knoll et al., 2020). As such, NYU fastMRI investigators provided data but did not participate in analysis or writing of this report. A listing of NYU fastMRI investigators, subject to updates, can be found at fastmri.med.nyu.edu. The primary goal of fastMRI is to test whether machine learning can aid in the reconstruction of medical images.

References

Aali, A., Arvinte, M., Kumar, S., and Tamir, J. I. Solving inverse problems with score-based generative priors learned from noisy data. *arXiv preprint arXiv:2305.01166*, 2023.

Abu-Hussein, S., Tirer, T., Chun, S. Y., Eldar, Y. C., and Giryes, R. Image restoration by deep projected GSURE. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3602–3611, 2022.

Aggarwal, H. K., Pramanik, A., John, M., and Jacob, M. ENSURE: A general approach for unsupervised training of deep image reconstruction algorithms. *IEEE Transactions on Medical Imaging*, 2022.

Batson, J. and Royer, L. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pp. 524–533. PMLR, 2019.

Blau, T., Ganz, R., Kawar, B., Bronstein, A., and Elad, M. Threat model-agnostic adversarial defense using diffusion models. *arXiv preprint arXiv:2207.08089*, 2022.

Blu, T. and Luisier, F. The sure-let approach to image denoising. *IEEE Transactions on Image Processing*, 16 (11):2778–2786, 2007.

Chen, D., Tachella, J., and Davies, M. E. Robust equivariant imaging: a fully unsupervised framework for learning to image from noisy and partial measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5647–5656, 2022.

Chung, H. and Ye, J. C. Score-based diffusion models for accelerated MRI. *Medical Image Analysis*, 80:102479, 2022.

Chung, H., Ryu, D., McCann, M. T., Klasky, M. L., and Ye, J. C. Solving 3D inverse problems using pre-trained 2D diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Eldar, Y. C. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2008.

Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. DiffuSeq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations*, 2023.

Hendriksen, A. A., Pelt, D. M., and Batenburg, K. J. Noise2inverse: Self-supervised deep convolutional denoising for tomography. *IEEE Transactions on Computational Imaging*, 6:1320–1335, 2020.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Hsieh, Y.-G., Kasiviswanathan, S. P., Kveton, B., and Blöbaum, P. Thompson sampling with diffusion generative prior. *arXiv preprint arXiv:2301.05182*, 2023.

Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A. G., and Tamir, J. Robust compressed sensing mri with deep generative priors. *Advances in Neural Information Processing Systems*, 34:14938–14954, 2021.

Jo, Y., Chun, S. Y., and Choi, J. Rethinking deep image prior for denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5087–5096, 2021.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023.

Kim, D., Kim, Y., Kang, W., and Moon, I.-C. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022.

Knoll, F., Zbontar, J., Sriram, A., Muckley, M. J., Bruno, M., Defazio, A., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzalv, M., Romero, A., Rabbat, M., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. Noise2noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pp. 2965–2974. PMLR, 2018.

Li, X. L., Thickstun, J., Gulrajani, I., Liang, P., and Hashimoto, T. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022.

Liu, J., Sun, Y., Eldeniz, C., Gan, W., An, H., and Kamilov, U. S. RARE: Image reconstruction using deep priors learned without groundtruth. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1088–1099, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Metzler, C. A., Mousavi, A., Heckel, R., and Baraniuk, R. G. Unsupervised learning with stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

Mullainathan, S. and Obermeyer, Z. Solving medicine's data bottleneck: Nightingale open science. *Nature Medicine*, 28(5):897–899, May 2022.

Nguyen, H. V., Ulfarsson, M. O., and Sveinsson, J. R. Hyperspectral image denoising using sure-based unsupervised convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3369–3382, 2020.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Pinaya, W. H., Graham, M. S., Gray, R., Da Costa, P. F., Tudosiu, P.-D., Wright, P., Mah, Y. H., MacKinnon, A. D., Teo, J. T., Jager, R., et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pp. 705–714. Springer, 2022.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Ramani, S., Blu, T., and Unser, M. Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms. *IEEE Transactions on image processing*, 17(9):1540–1554, 2008.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., and Taigman, Y. Make-A-Video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Soltanayev, S. and Chun, S. Y. Training deep learning based denoisers without ground truth data. *Advances in neural information processing systems*, 31, 2018.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020b.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2023.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A. H., and Cohen-Or, D. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Theis, L., Salimans, T., Hoffman, M. D., and Mentzer, F. Lossy compression with Gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, pp. 2022–12, 2022.

Wyatt, J., Leach, A., Schmon, S. M., and Willcocks, C. G. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 650–656, 2022.

Xiang, T., Yurt, M., Syed, A. B., Setsompop, K., and Chaudhari, A. DDM^2: Self-supervised diffusion MRI denoising with generative diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.

Xie, Y. and Li, Q. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pp. 655–664. Springer, 2022.

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., Stern, R., Johnson, P., Bruno, M., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Yakubova, N., Pinkerton, J., Wang, D., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. fastMRI: An open dataset and benchmarks for accelerated MRI, 2019.

Zhang, X.-P. and Desai, M. D. Adaptive denoising based on sure risk. *IEEE signal processing letters*, 5(10):265–267, 1998.

Zhussip, M., Soltanayev, S., and Chun, S. Y. Training deep learning based image denoisers from undersampled measurements without ground truth and without image prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10255–10264, 2019.

Zimmermann, R. S., Schott, L., Song, Y., Dunn, B. A., and Klindt, D. A. Score-based generative classifiers. *arXiv preprint arXiv:2110.00473*, 2021.

A. Proposition Proofs

Proposition 3.1. *For* $\mathbf{x}$ *sampled from an unknown* $q(\mathbf{x})$, $\bar{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$, $\bar{\mathbf{x}}_t$ *sampled from Equation 5, and the diagonal weight matrix* $\mathbf{W} = \mathbb{E}[\mathbf{P}]^{-\frac{1}{2}} \succ 0$ *(positive definite), it holds that Equation 6 approximately equals Equation 1.*

Proof. We focus on the expectation term from Equation 6, which is taken over $\bar{\mathbf{x}}_t \sim q(\bar{\mathbf{x}}_t|\bar{\mathbf{x}}, \mathbf{P}),\ \bar{\mathbf{x}} \sim q(\bar{\mathbf{x}}),\ \mathbf{P} \sim q_P(\mathbf{P})$ with unknown $q(\bar{\mathbf{x}}), q_P(\mathbf{P})$. Namely,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{W}\mathbf{P}\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\right\|_2^2\right] \\
&\overset{1}{=} \mathbb{E}\left[\mathrm{Trace}\left(\mathbf{W}\mathbf{P}\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top \mathbf{P}\mathbf{W}\right)\right] \\
&\overset{2}{=} \mathbb{E}\left[\mathrm{Trace}\left(\mathbf{P}\mathbf{W}^2\mathbf{P}\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top\right)\right] \\
&\overset{3}{=} \mathbb{E}\left[\mathrm{Trace}\left(\mathbf{W}^2\mathbf{P}^2\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top\right)\right] \\
&\overset{4}{=} \mathbb{E}\left[\mathrm{Trace}\left(\mathbf{W}^2\mathbf{P}\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top\right)\right] \\
&\overset{5}{=} \mathrm{Trace}\left(\mathbb{E}\left[\mathbf{W}^2\mathbf{P}\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top\right]\right) \\
&\overset{6}{=} \mathrm{Trace}\left(\mathbf{W}^2\mathbb{E}\left[\mathbf{P}\right]\mathbb{E}\left[\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top\right]\right) \\
&\overset{7}{=} \mathrm{Trace}\left(\mathbb{E}\left[\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top\right]\right) \\
&\overset{5}{=} \mathbb{E}\left[\mathrm{Trace}\left(\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)^\top\right)\right] \\
&\overset{8}{=} \mathbb{E}\left[\left\|f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right\|_2^2\right].
\end{aligned} \tag{8}
$$

Justifications:

1. Using the linear algebra property $\|\mathbf{v}\|_2^2 = \mathrm{Trace}\left(\mathbf{v}\mathbf{v}^\top\right)$ for any vector $\mathbf{v}$, and $\mathbf{P} = \mathbf{P}^\top, \mathbf{W} = \mathbf{W}^\top$ because they are diagonal matrices.
2. Using the cyclical shift invariance of the trace operator, $\mathrm{Trace}\left(\mathbf{ABC}\right) = \mathrm{Trace}\left(\mathbf{CAB}\right)$.
3. Diagonal matrices (such as $\mathbf{P}$ and $\mathbf{W}$) commute with one another.
4. Since $\mathbf{P}$ is a diagonal matrix whose values are either zeroes or ones, it holds $\mathbf{P}^2 = \mathbf{P}$.
5. For any random matrix $\mathbf{A}$, it holds that $\mathbb{E}\left[\mathrm{Trace}\left(\mathbf{A}\right)\right] = \mathrm{Trace}\left(\mathbb{E}\left[\mathbf{A}\right]\right)$.
6. $\mathbf{W}^2$ is a constant and can therefore be taken out of the expectation. Additionally, we use the assumption that the denoiser's mean squared error $\left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)$ is independent of $\mathbf{P}$.
7. $\mathbf{W}$ is defined as $\mathbb{E}[\mathbf{P}]^{-\frac{1}{2}}$. This results in $\mathbf{W}^2\mathbf{P} = \mathbf{I}$.
8. Using the linear algebra property $\|\mathbf{v}\|_2^2 = \mathrm{Trace}\left(\mathbf{v}\mathbf{v}^\top\right)$ for any vector $\mathbf{v}$.

Equation 8 is identical to the expectation term in Equation 1, except for the distribution of $\bar{\mathbf{x}}_t$ considered in the expectation. Equation 8 considers $\bar{\mathbf{x}}_t \sim q(\bar{\mathbf{x}}_t|\bar{\mathbf{x}}, \mathbf{P}) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{P}\bar{\mathbf{x}}, (1-\bar{\alpha}_t)\mathbf{I})$, whereas Equation 1 considers $\bar{\mathbf{x}}_t \sim q^*(\bar{\mathbf{x}}_t|\bar{\mathbf{x}}) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\bar{\mathbf{x}}, (1-\bar{\alpha}_t)\mathbf{I})$. We assume that the neural network $f_\theta^{(t)}(\bar{\mathbf{x}}_t)$ is able to infer $\mathbf{P}$ from $\bar{\mathbf{x}}_t$, and can also tailor its output for each $\mathbf{P}$ including $\mathbf{P} = \mathbf{I}$, matching $q^*(\bar{\mathbf{x}}_t|\bar{\mathbf{x}})$. Under these assumptions, both expectations share the same minimizer, thereby completing the proof. A similar proof is presented in ENSURE (Aggarwal et al., 2022). □

Proposition 3.2. *For* $\mathbf{x}$ *sampled from an unknown* $q(\mathbf{x})$, $\bar{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$, $\bar{\mathbf{x}}_t$ *sampled from Equation 5, and* $\lambda_t = 1 - \bar{\alpha}_t$, *it holds that Equation 7 equals Equation 6.*

Proof. We utilize a weighted version of the generalized SURE (Eldar, 2008; Aggarwal et al., 2022) presented in Equation 2,

$$\mathbb{E}\left[\left\|\mathbf{WP}\left(f(\mathbf{y}) - \mathbf{x}\right)\right\|_2^2\right] = \mathbb{E}\left[\left\|\mathbf{WP}\left(f(\mathbf{y}) - \mathbf{x}_{\mathrm{ML}}\right)\right\|_2^2\right] + 2\mathbb{E}\left[\nabla_{\mathbf{H}^\top \mathbf{C}^{-1}\mathbf{y}} \cdot \mathbf{W}^2\mathbf{P}f(\mathbf{y})\right] + c. \tag{9}$$

This weighted GSURE considers the measurement equation $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(0, \mathbf{C})$, with $\mathbf{P} = \mathbf{H}^\dagger\mathbf{H}$, $\mathbf{x}_{\mathrm{ML}} = \left(\mathbf{H}^\top\mathbf{C}^{-1}\mathbf{H}\right)^\dagger \mathbf{H}^\top\mathbf{C}^{-1}\mathbf{y}$, and a constant c. We consider the measurement equation matching Equation 5, namely $\bar{\mathbf{x}}_t = \sqrt{\bar{\alpha}_t}\mathbf{P}\bar{\mathbf{x}} + \bar{\mathbf{z}}_t$ with $\bar{\mathbf{z}}_t \sim \mathcal{N}(0, (1-\bar{\alpha}_t)\mathbf{I})$. For these measurements, the left-hand-side in Equation 9 becomes

$$\mathbb{E}\left[\left\|\mathbf{W}\left(\sqrt{\bar{\alpha}_t}\mathbf{P}\right)^\dagger\left(\sqrt{\bar{\alpha}_t}\mathbf{P}\right)\left(f(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\right\|_2^2\right] = \mathbb{E}\left[\left\|\mathbf{WP}\left(f(\bar{\mathbf{x}}_t) - \bar{\mathbf{x}}\right)\right\|_2^2\right],$$

which is identical to the expectation term in Equation 6. This equation holds because $\mathbf{P}$ is a diagonal matrix with ones and zeroes, resulting in $\mathbf{P}^\dagger = \mathbf{P} = \mathbf{P}^2$. Meanwhile, by substituting $\mathbf{H} = \sqrt{\bar{\alpha}_t}\mathbf{P}$, $\mathbf{C} = (1-\bar{\alpha}_t)\mathbf{I}$, and $\mathbf{y} = \bar{\mathbf{x}}_t$, $\mathbf{x}_{\mathrm{ML}}$ becomes

$$\begin{aligned}
\mathbf{x}_{\mathrm{ML}} &= \left(\sqrt{\bar{\alpha}_t}\mathbf{P}^\top\left((1-\bar{\alpha}_t)\mathbf{I}\right)^{-1}\sqrt{\bar{\alpha}_t}\mathbf{P}\right)^\dagger \sqrt{\bar{\alpha}_t}\mathbf{P}^\top\left((1-\bar{\alpha}_t)\mathbf{I}\right)^{-1}\bar{\mathbf{x}}_t \\
&= \left(\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\mathbf{P}^\top\mathbf{P}\right)^\dagger \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{P}^\top\bar{\mathbf{x}}_t \\
&= \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\mathbf{P}^\dagger\left(\mathbf{P}^\top\right)^\dagger \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\mathbf{P}^\top\bar{\mathbf{x}}_t \\
&= \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{PPP}\bar{\mathbf{x}}_t = \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{P}\bar{\mathbf{x}}_t.
\end{aligned}$$

The last two equalities hold because $\mathbf{P}^\dagger = \mathbf{P}^\top = \mathbf{P} = \mathbf{P}^2$. Finally, the right-hand-side in Equation 9 becomes

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{WP}^\dagger\mathbf{P}\left(f(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{P}\bar{\mathbf{x}}_t\right)\right\|_2^2\right] + 2\mathbb{E}\left[\nabla_{\mathbf{P}^\top((1-\bar{\alpha}_t)\mathbf{I})^{-1}\bar{\mathbf{x}}_t} \cdot \mathbf{W}^2\mathbf{P}^\dagger\mathbf{P}f(\bar{\mathbf{x}}_t)\right] + c \\
&\overset{1}{=} \mathbb{E}\left[\left\|\mathbf{WP}\left(f(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}\bar{\mathbf{x}}_t\right)\right\|_2^2\right] + 2\mathbb{E}\left[\nabla_{\mathbf{P}((1-\bar{\alpha}_t)\mathbf{I})^{-1}\bar{\mathbf{x}}_t} \cdot \mathbf{W}^2\mathbf{P}f(\bar{\mathbf{x}}_t)\right] + c \\
&= \mathbb{E}\left[\left\|\mathbf{WP}\left(f(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}\bar{\mathbf{x}}_t\right)\right\|_2^2\right] + 2\mathbb{E}\left[\nabla_{(1/(1-\bar{\alpha}_t))\mathbf{P}\bar{\mathbf{x}}_t} \cdot \mathbf{W}^2\mathbf{P}f(\bar{\mathbf{x}}_t)\right] + c \\
&\overset{2}{=} \mathbb{E}\left[\left\|\mathbf{WP}\left(f(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}\bar{\mathbf{x}}_t\right)\right\|_2^2\right] + 2\mathbb{E}\left[(1-\bar{\alpha}_t)\nabla_{\mathbf{P}\bar{\mathbf{x}}_t} \cdot \mathbf{W}^2\mathbf{P}f(\bar{\mathbf{x}}_t)\right] + c \\
&\overset{3}{=} \mathbb{E}\left[\left\|\mathbf{WP}\left(f(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}\bar{\mathbf{x}}_t\right)\right\|_2^2\right] + 2\mathbb{E}\left[(1-\bar{\alpha}_t)\nabla_{\bar{\mathbf{x}}_t} \cdot \mathbf{PW}^2 f(\bar{\mathbf{x}}_t)\right] + c \\
&\overset{4}{=} \mathbb{E}\left[\left\|\mathbf{WP}\left(f(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}\bar{\mathbf{x}}_t\right)\right\|_2^2 + 2(1-\bar{\alpha}_t)\nabla_{\bar{\mathbf{x}}_t} \cdot \mathbf{PW}^2 f(\bar{\mathbf{x}}_t) + c\right],
\end{aligned}$$

which is identical to the expectation term in Equation 7 with $\lambda_t = 1 - \alpha_t$. Justifications:

1. $\mathbf{P}^\dagger = \mathbf{P}^\top = \mathbf{P} = \mathbf{P}^2$.
2. Using the change of variables formula.
3. Diagonal matrices (such as $\mathbf{P}$ and $\mathbf{W}$) commute with one another. Additionally, the divergences w.r.t. $\bar{\mathbf{x}}_t$ and w.r.t. $\mathbf{P}\bar{\mathbf{x}}_t$ are identical, because $\mathbf{PW}^2 f(\bar{\mathbf{x}}_t)$ equals zero in entries where $\mathbf{P}$ is zero, and $\mathbf{P}\bar{\mathbf{x}}_t$ and $\bar{\mathbf{x}}_t$ are identical in entries where $\mathbf{P}$ is non-zero.
4. Using the linearity of the expectation operator.

By rewriting both sides of Equation 9, we obtain that Equation 7 equals Equation 6. □

B. Related Work

There has been a rich vein of works on unsupervised learning from datasets of corrupted data (Lehtinen et al., 2018; Batson & Royer, 2019; Hendriksen et al., 2020; Chen et al., 2022), including several SURE- and GSURE-based approaches (Zhang & Desai, 1998; Blu & Luisier, 2007; Soltanayev & Chun, 2018; Nguyen et al., 2020; Jo et al., 2021; Metzler et al., 2018; Zhussip et al., 2019; Aggarwal et al., 2022; Abu-Hussein et al., 2022; Liu et al., 2020). While they achieve impressive results, these efforts are mostly focused on learning a specific task. In contrast, our approach learns a foundational generative model, making it suitable for a wide range of applications.

Diffusion models have had incredible success in image generation (Song et al., 2020b; Vahdat et al., 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022; Kim et al., 2022), as well as generation of other modalities (Ho et al., 2022; Singer et al., 2023; Kong et al., 2021; Popov et al., 2021; Gong et al., 2023; Li et al., 2022; Tevet et al., 2022). They have also been deployed in a myriad of related tasks (Kawar et al., 2022; Theis et al., 2022; Blau et al., 2022; Pinaya et al., 2022; Wyatt et al., 2022; Kawar et al., 2023; Zimmermann et al., 2021). In particular, diffusion models have been adapted for medical imaging (*e.g.*, MRI) (Song et al., 2023; Chung & Ye, 2022; Jalal et al., 2021; Xie & Li, 2022; Chung et al., 2023). These models can serve a multitude of tasks, but can be expensive to train as they require fully sampled noiseless training data. Notably, DDM^2 (Xiang et al., 2023) and the concurrent work of SURE-Score (Aali et al., 2023) offer ways to train a diffusion model based on noisy data, which is often the case in practical settings. However, collected data can also be undersampled or corrupted by other transformations. Our proposed framework is more general, as it can handle both Gaussian noise and linear corruptions.

Diffusion models have permeated various research areas, including online decision making (Hsieh et al., 2023). In that context, obtaining full noiseless data may often be impossible. Hsieh et al. (2023) suggest a diffusion loss function that learns a diffusion-based prior from noisy data with missing elements, similar to an image inpainting problem. They note that their proposed loss function could be of independent interest in future work. Our loss function closely resembles theirs, albeit generalizing for linear corruptions beyond inpainting by utilizing the singular value decomposition (SVD).

C. Detailed Data Descriptions

C.1. Dataset Collection

Here, we detail the collection process for the training and testing data in our experiments. Note that the data described here is what we consider pristine uncorrupted data. The corruption process for training GSURE-Diffusion is detailed in subsection C.2.

CelebA. In our experiments on human face images, we use images from the CelebA (Liu et al., 2015) dataset. The original CelebA images were center-cropped to 128×128 pixels, then resized to 32×32 pixels, and finally turned into grayscale. The images were converted to grayscale by averaging all color channels. Overall, the dataset includes 162770 training set images, and 19867 validation set images (which we use for FID (Heusel et al., 2017) evaluations).

FastMRI. We consider all single-coil knee MRI scans from the fastMRI (Zbontar et al., 2019; Knoll et al., 2020) dataset, excluding slices with indices below 10 or above 40 as they generally contain less interpretable information. This yields a training set size of 24853. For the validation set we only use the first 1024 valid slices (which we use for all our post-training experiments). We treat each slice as a 2-channel image, separating the complex values into real and imaginary channels. We center-crop the images to a spatial size of 320×320 following (Jalal et al., 2021), and normalize the images by $7e-5$ to obtain better neural network performance. When displaying MR images, we take the absolute value of the complex number in each pixel, and then use min-max normalization to view the resulting values as a grayscale image.

C.2. Data Corruptions for GSURE-Diffusion

CelebA. In our CelebA (Liu et al., 2015) experiments, we consider a degradation operator $\mathbf{H}$ that randomly drops each 4×4-pixel patch with probability p. This operator can be mathematically defined as a diagonal matrix $\mathbf{H}$ with zeroes in pixels that are dropped and ones in pixels that are kept. The singular value decomposition (SVD) is trivially and efficiently obtained by

$$\mathbf{H} = \mathbf{IHI}. \tag{10}$$

Note that the singular values in $\mathbf{H}$ are not ordered. Since the SVD of $\mathbf{H}$ has $\mathbf{V}^\top = \mathbf{I}$ regardless of the randomness of dropping patches, this family of random operators $\mathbf{H}$ matches our assumption that all $\mathbf{H}$ share the same left-singular vectors $\mathbf{V}^\top$.

Additionally, the projection matrix $\mathbf{P} = \mathbf{H}^\dagger\mathbf{H}$ is simply $\mathbf{H}$ (as $\mathbf{H}$ is diagonal with zeroes and ones). Because each patch is dropped randomly with probability p, it follows that $\mathbb{E}[\mathbf{P}] = (1-p)\mathbf{I} \succ 0$ is positive definite, matching our assumption.

Finally, we assume $\mathbf{H}$ and the additive white Gaussian noise standard deviation σ_0 to be known for all measurements in the dataset. For simplicity, we assume a uniform σ_0 for all measurements.

FastMRI. For MRI slices from fastMRI (Zbontar et al., 2019; Knoll et al., 2020), the degradation operator we use is the horizontal frequency subsampling operator used in (Jalal et al., 2021). For an acceleration factor R, the degradation operator $\mathbf{H}$ keeps the central $120/R$ frequencies, and then uniformly samples an additional $200/R$ frequencies. This results in a sampling of $320/R$ frequencies out of the original 320. More formally,

$$\mathbf{H} = \mathbf{I}\mathbf{\Sigma}\mathbf{F}, \tag{11}$$

where $\mathbf{F}$ is the discrete Fourier transform matrix, and $\mathbf{\Sigma}$ is a square diagonal matrix containing ones for frequency indices that are kept by $\mathbf{H}$, and zeroes elsewhere. Incidentally, Equation 11 is a valid SVD of $\mathbf{H}$, and can be efficiently simulated using the fast Fourier transform algorithm.

This operator matches our assumptions: (i) We assume each $\mathbf{H}$ and the additive white Gaussian noise standard deviation σ_0 to be known; (ii) All matrices $\mathbf{H}$ share the same left-singular vectors defined by $\mathbf{F}$ (and not depending on the randomness); and (iii) The central $120/R$ horizontal frequencies are always sampled, and each of the remaining frequencies are equally likely to be sampled, with probability $200/(320R - 120)$. Thus $\mathbb{E}[\mathbf{P}] = \mathbb{E}[\mathbf{\Sigma}^\dagger\mathbf{\Sigma}]$ is a diagonal matrix with nonzero diagonal values, making it positive definite.

D. Implementation Details

Our experiments were conducted using DDPM (Ho et al., 2020) U-Net architecture with base channel width 128. All networks were trained using the Adam optimizer, dropout with probability 0.1, EMA with decay factor of 0.9999. The diffusion process considered in training for all experiments has 1000 steps, with a linear β schedule ranging from $\beta_1 = \sigma_0^2$ (σ_0^2 is the variance of the AWGN in the data) to $\beta_{1000} = 0.2$. All experiments were conducted on 8 NVIDIA A40 GPUs.

In the human faces experiment we ignore the weighting matrix $\mathbf{W}$ during training because the probability for each pixel to be masked is uniform. For the knee MRI experiment the weighting matrix $\mathbf{W}$ was set to 1 for the centeral lines that were not masked by $\mathbf{H}$, and $\sqrt{5.8}$ for all other lines matching their inverse square root masking probability (for $R = 4$).

All models, including the oracle ones, were trained with the hyperparameters listed in Table 2. The "mean type" hyperparameter refers to whether the neural network predicts the image $\mathbf{x}$ or the added noise $\boldsymbol{\epsilon} = (\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}) \,/\, \left(\sqrt{1-\bar{\alpha}_t}\right)$.

Table 2. Architecture and training hyperparameters for CelebA and fastMRI experiments.

	CelebA	**FastMRI**
Iterations	180,000	31,000
Batch Size	128	32
Learning Rate	$5e-5$	$1e-5$
Mean Type	`predict_x`	`predict_epsilon`
Channel Multipliers	$[1,2,2,2,4]$	$[1,1,2,2,4,4]$
Attention Resolutions	[16]	[20]
γ_t	1	$\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}$
λ_t	0.0001	$0.0001 \cdot \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}$

For the MRI model, we apply an inverse Fourier transform and a Fourier transform to the network's input and output respectively, to utilize the convolutional architecture's advantage on image data (rather than frequencies). Due to the orthogonality and linearity of the Fourier transform and its inverse, the additive white Gaussian noise remains so, and maintains the same variance.

We provide our code, configuration files, and trained model checkpoints at https://github.com/bahjat-kawar/gsure-diffusion/.

E. Pragmatic Loss Function Considerations

E.1. Divergence Term Estimation

The GSURE-Diffusion training loss in Equation 7 contains a divergence term, which is highly expensive to accurately obtain, in both memory consumption and computation time. Similar to other SURE-based methods (Metzler et al., 2018; Soltanayev & Chun, 2018; Aggarwal et al., 2022), we use an unbiased Monte Carlo approximation (Ramani et al., 2008) of the divergence. Considering the divergence as the trace of the Jacobian matrix $\mathbf{J}$ of the term being differentiated ($\mathbf{P}\mathbf{W}^2 f_\theta^{(t)}(\bar{\mathbf{x}}_t)$) , Monte Carlo SURE (Ramani et al., 2008) uses Hutchinson's trace estimator (Hutchinson, 1989). We compute the estimate by sampling a random Gaussian vector $\mathbf{v} \sim \mathcal{N}(0, \mathbf{I})$, and calculating $\mathbf{v}^\top \mathbf{J} \mathbf{v}$ using automatic differentiation tools. Notably, this differs from previous methods (Metzler et al., 2018; Soltanayev & Chun, 2018; Aggarwal et al., 2022) that used numerical estimates for differentiation, which may suffer from numerical inaccuracies.

E.2. MSE Term Variance

The GSURE-Diffusion loss function in Equation 7 contains the following squared error term

$$\left\| \mathbf{W}\mathbf{P} \left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}} \bar{\mathbf{x}}_t \right) \right\|_2^2 .$$

Because of the possibly strong noise-to-signal ratio present in $\bar{\mathbf{x}}_t$, this term may suffer from high variance, effectively impeding the training process. To alleviate this, we propose replacing $\frac{1}{\sqrt{\bar{\alpha}_t}} \bar{\mathbf{x}}_t$ with the less noisy $\bar{\mathbf{y}}$, resulting in the loss function

$$\sum_{t=1}^{T} \gamma_t \mathbb{E} \left[\left\| \mathbf{W}\mathbf{P} \left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}} \right) \right\|_2^2 + 2\lambda_t \left(\nabla_{\bar{\mathbf{x}}_t} \cdot \mathbf{P}\mathbf{W}^2 f_\theta^{(t)}(\bar{\mathbf{x}}_t) \right) + c \right] . \tag{12}$$

Note that the difference between the expectations in Equation 7 and Equation 12 is negligible, while Equation 12 has significantly less variance (as $\bar{\mathbf{y}}$ is less noisy than $\frac{1}{\sqrt{\bar{\alpha}_t}} \bar{\mathbf{x}}_t$). From Equation 4 we get

$$\frac{1}{\sqrt{\bar{\alpha}_t}} \bar{\mathbf{x}}_t = \bar{\mathbf{y}} + \frac{1}{\sqrt{\bar{\alpha}_t}} \left((1 - \bar{\alpha}_t) \mathbf{I} - \bar{\alpha}_t \sigma_0^2 \mathbf{\Sigma}^\dagger \mathbf{\Sigma}^{\dagger\top} \right)^{\frac{1}{2}} \boldsymbol{\epsilon}_t .$$

We denote $\bar{\boldsymbol{\epsilon}} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left((1 - \bar{\alpha}_t) \mathbf{I} - \bar{\alpha}_t \sigma_0^2 \mathbf{\Sigma}^\dagger \mathbf{\Sigma}^{\dagger\top} \right)^{\frac{1}{2}} \boldsymbol{\epsilon}_t$, and show that

$$\begin{aligned}
&\left\| \mathbf{W}\mathbf{P} \left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}} \bar{\mathbf{x}}_t \right) \right\|_2^2 \\
&= \left\| \mathbf{W}\mathbf{P} \left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}} - \bar{\boldsymbol{\epsilon}} \right) \right\|_2^2 \\
&= \left\| \mathbf{W}\mathbf{P} \left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}} \right) \right\|_2^2 + \left\| \mathbf{W}\mathbf{P}\bar{\boldsymbol{\epsilon}} \right\|_2^2 - 2\bar{\boldsymbol{\epsilon}}^\top \mathbf{P}\mathbf{W}\mathbf{W}\mathbf{P} \left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}} \right) \\
&= \left\| \mathbf{W}\mathbf{P} \left(f_\theta^{(t)}(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}} \right) \right\|_2^2 - 2\bar{\boldsymbol{\epsilon}}^\top \mathbf{P}\mathbf{W}^2\mathbf{P} f_\theta^{(t)}(\bar{\mathbf{x}}_t) + 2\bar{\boldsymbol{\epsilon}}^\top \mathbf{P}\mathbf{W}^2\mathbf{P}\bar{\mathbf{y}} + \left\| \mathbf{W}\mathbf{P}\bar{\boldsymbol{\epsilon}} \right\|_2^2 .
\end{aligned}$$

The final two terms are constants w.r.t. θ. Effectively, this means that the difference between the squared error terms in Equation 7 and Equation 12 is $2\bar{\boldsymbol{\epsilon}}^\top \mathbf{P}\mathbf{W}^2\mathbf{P} f_\theta^{(t)}(\bar{\mathbf{x}}_t)$. Under the manifold hypothesis, if $f_\theta^{(t)}(\bar{\mathbf{x}}_t)$ outputs valid images residing on the manifold, and because $\bar{\boldsymbol{\epsilon}}$ is a random Gaussian vector, $f_\theta^{(t)}(\bar{\mathbf{x}}_t)$ and $\bar{\boldsymbol{\epsilon}}$ are perpendicular. Therefore, the expected difference between the squared error terms, $\mathbb{E}\left[2\bar{\boldsymbol{\epsilon}}^\top \mathbf{P}\mathbf{W}^2\mathbf{P} f_\theta^{(t)}(\bar{\mathbf{x}}_t) \right]$, is zero. This motivates us to replace Equation 7 with Equation 12, resulting in significantly lower variance in the loss at little to no cost in terms of bias.

F. Additional Results

We show more training data examples and generated images for all tested degradation levels in CelebA experiments in Figure 4 and Figure 5.

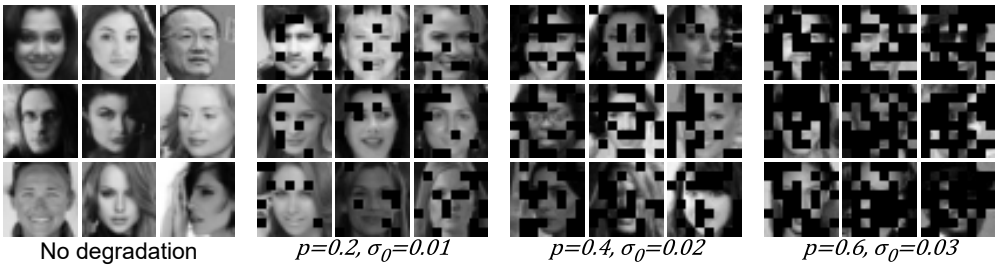

Figure 4. Training set samples of the different degradation settings in CelebA experiments.

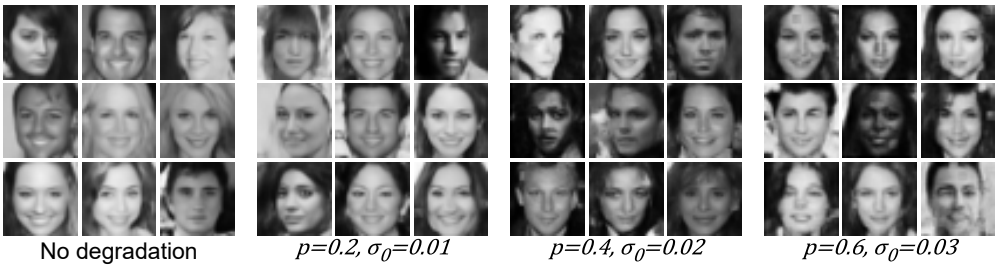

Figure 5. Generated samples (with 50 DDIM steps) from models trained on different degradation settings in CelebA experiments.

In all MRI experiments, we show the "zero-filled" result (the MR image with zeroes in the missing frequencies) as a baseline. We observe in Figure 6 that GSURE-Diffusion achieves similar performance to the oracle for $R = 4$. The model can also generalize for higher acceleration factors, as we show in Figure 7, as well as subsampling masks of different characteristics, as we show in Figure 3. Furthermore, as a generative model, GSURE-Diffusion can provide uncertainty estimates for its outputs. We follow (Chung & Ye, 2022) and quantify the uncertainty using the standard deviation of 8 stochastic outputs made by the model. We add synthetic Gaussian noise to MR images with $\sigma_0 = 0.4$, and show uncertainty quantification results using GSURE-Diffusion and the oracle model in Figure 8. We jointly normalize standard deviations to ensure a fair visual comparison. We attach a color bar to accurately illustrate the standard deviation intensities. This uncertainty quantification technique can potentially aid medical practitioners, providing clues towards anomalous regions in MRI scans.

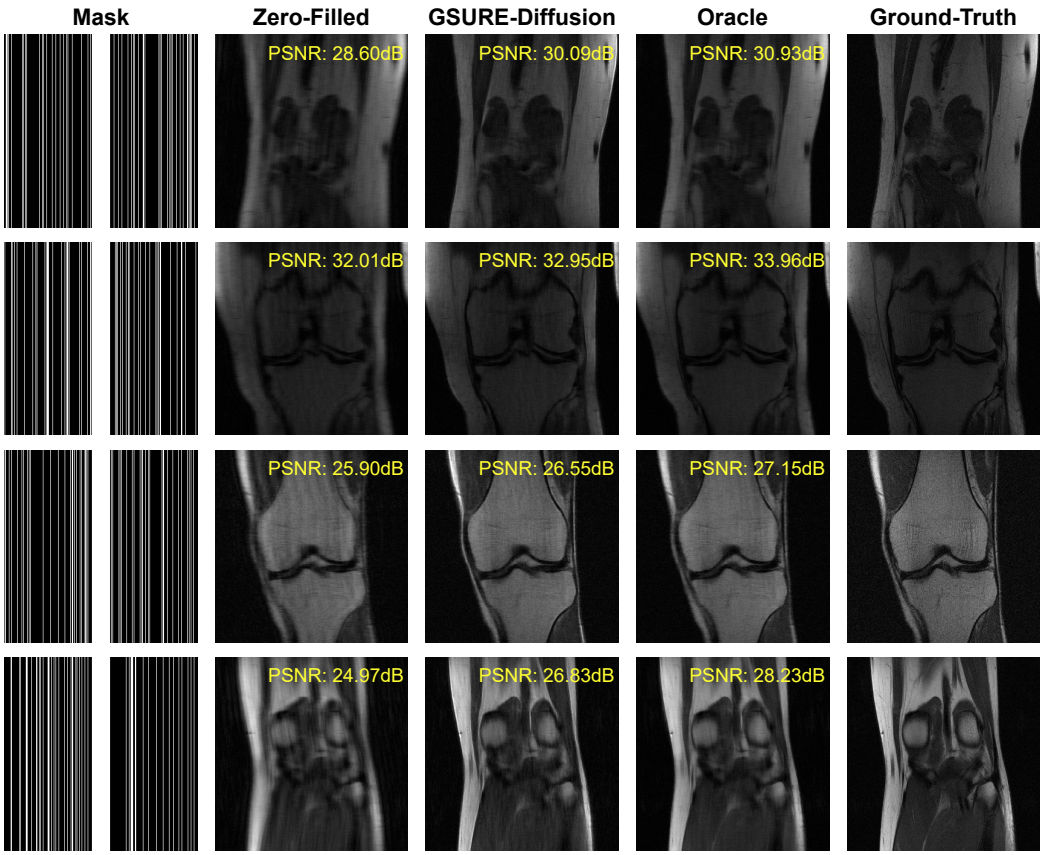

Figure 6. Accelerated MRI reconstruction results for $R = 4$ and $\sigma_0 = 0.01$.

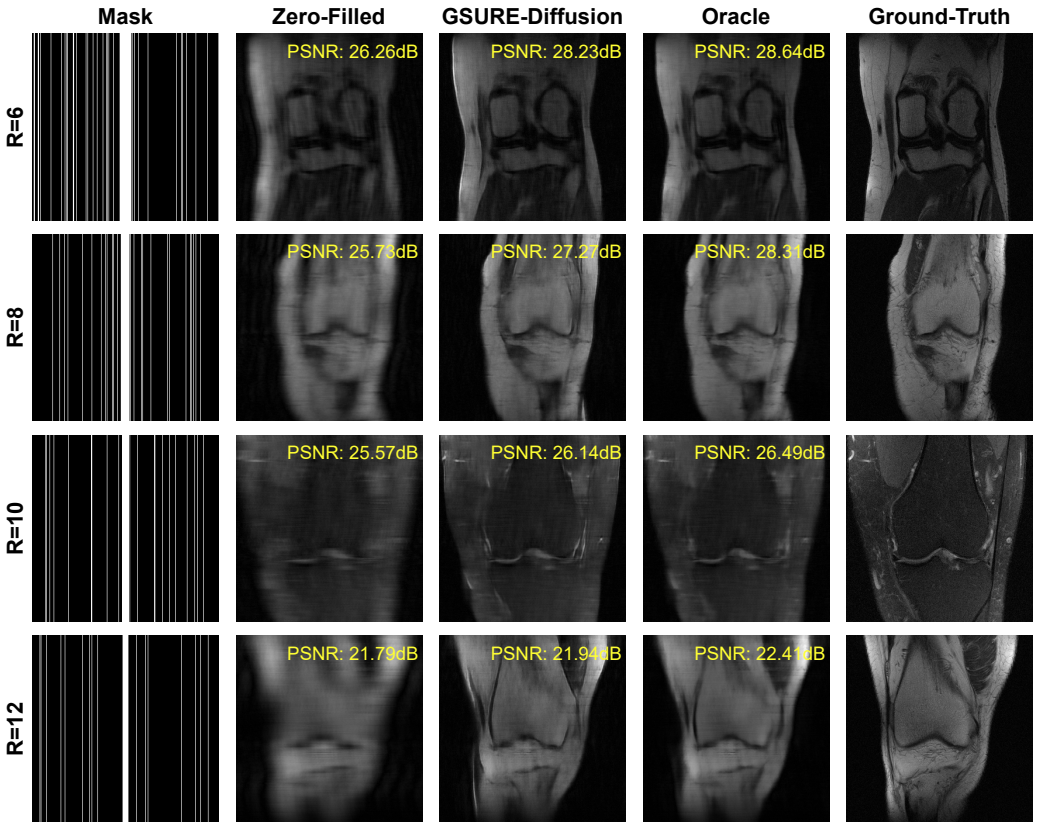

Figure 7. Accelerated MRI reconstruction results for $R \in \{6, 8, 10, 12\}$ and $\sigma_0 = 0.01$. GSURE-Diffusion can generalize well across different acceleration factors.

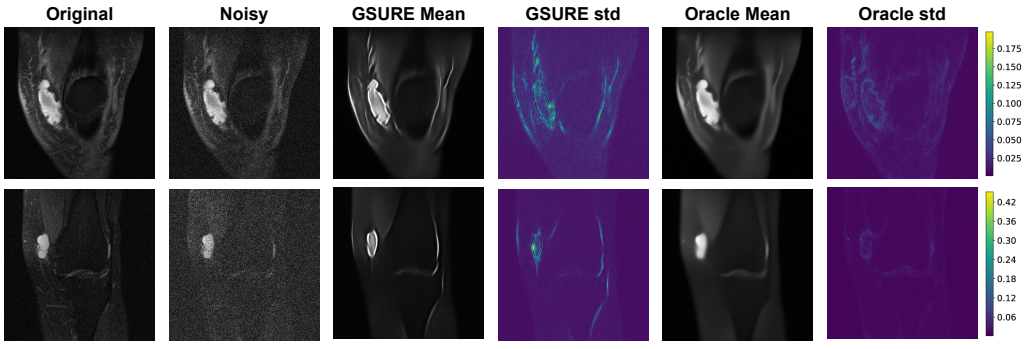

Figure 8. Uncertainty quantification for MR image denoising with GSURE-Diffusion and oracle models. Means and standard deviations are calculated for 8 stochastic diffusion reconstructions.