SUD^2: Supervision by Denoising Diffusion Models for Image Reconstruction

Matthew Chan
Department of Computer Science
University of Maryland, College Park
College Park, MD, 20742, USA
mattchan@cs.umd.edu

Sean I. Young
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, 02139, USA
siyoung@mit.edu

Christopher A. Metzler
Department of Computer Science
University of Maryland, College Park
College Park, MD, 20742, USA
metzler@umd.edu

Abstract

Many imaging inverse problems—such as image-dependent in-painting and dehazing—are challenging because their forward models are unknown or depend on unknown latent parameters. While one can solve such problems by training a neural network with vast quantities of paired training data, such paired training data is often unavailable. In this paper, we propose a generalized framework for training image reconstruction networks when paired training data is scarce. In particular, we demonstrate the ability of image denoising algorithms and, by extension, denoising diffusion models, to supervise network training in the absence of paired training data. (The unabridged version of this manuscript is available at https://arxiv.org/abs/2303.09642)

1 Introduction

Imaging inverse problems can generally be described in terms of a forward operator $\mathcal{F}(\cdot)$ that maps a scene x to a measurement y according to $y = \mathcal{F}(x)$. Historically, computational imaging research has focused on solving inverse problems with known forward models. For instance, computed tomography's forward model can be represented as a Radon transform and magnetic resonance imaging's forward model can be represented as 2D Fourier Transform. Knowledge of these forward models allows one to reconstruct scenes x from measurements y using any number of classical or learning-based algorithms [1].

Since the onset of the deep learning era, significant progress has been made in solving inverse problems which lack explicit forward models. Using vast amounts of training pairs $\{x_i, y_i\}_{i=0}^N$, neural networks learn to directly map samples from a source distribution, y_i, to images from a target distribution, x_i. In doing so, the network implicitly learns the inverse operator $\mathcal{F}^{-1}$ without any explicit knowledge of the forward model $\mathcal{F}$. The main drawback of such methods is that their performance is directly related to the size and quality of the training dataset. As a result, they often struggle whenever little to no paired training data is available.

Our goal in this work is to train a network $f_\theta(\cdot)$ to reconstruct images/scenes x from measurements y using three sets of training data:

- A small set P of paired examples (x_p, y_p) drawn from the joint distribution $p_{x,y}$.
- A large set U_y of unpaired measurements y_u drawn from the marginal distribution p_y.
- A large set U_x of unpaired images x_u drawn from the marginal distribution p_x.

Such mixed datasets naturally occur in applications where gathering unpaired data is easy, but gathering paired data is a challenge. For instance, it is straightforward to capture images with fog and images without fog, but capturing two paired images of the same scene with and without fog (with lighting conditions and all other nuisance variations fixed) is very challenging[2, 3]. Often times, the latter paired dataset is restricted to only a few images captured in a lab.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

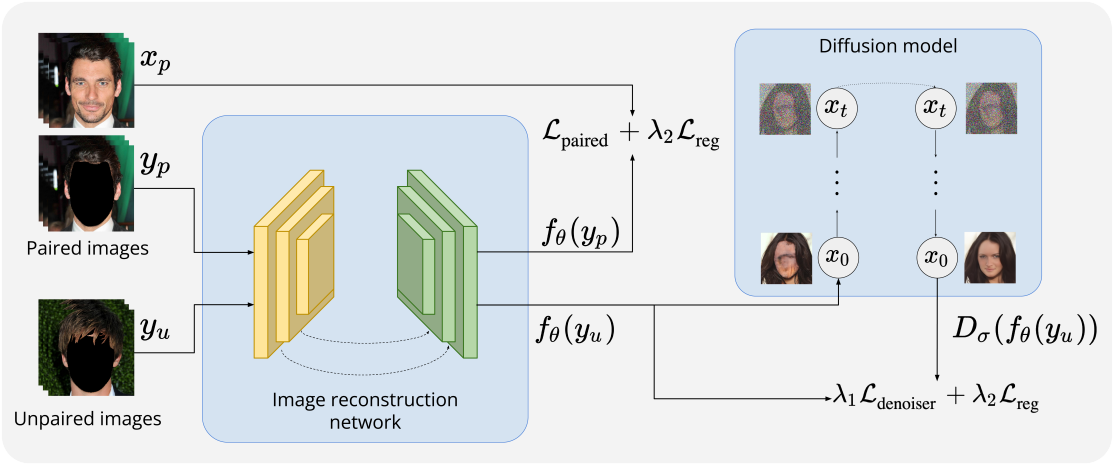

Figure 1: **Overview of the training pipeline.** A pre-trained diffusion model supervises training by pushing outputs of the image reconstruction network towards the desired target image distribution.

A paired training set, P, allows one to optimize $f_\theta(\cdot)$ by minimizing the empirical risk

$$\mathcal{L}_{\text{paired}} = \frac{1}{|P|} \sum_{(x_p, y_p) \in P} \|x_p - f_\theta(y_p)\|^2, \tag{1}$$

where $|P|$ denotes the cardinality of P. However, as the size of P decreases $\mathcal{L}_{\text{paired}}$ becomes a poor approximation of the true risk and $f_\theta(\cdot)$ overfits to the training set. As an alternative, we seek to leverage unpaired datasets U_x and U_y to improve the quality of our reconstructions.

2 Supervision-by-denoising

Young et al. [4] recently introduced the supervision-by-denoising (SUD) framework which extends the ideas behind regularization-by-denoising [5] to semi-supervised learning. The key intuition behind SUD and similar works [6–8] is that learned denoisers, $D_\sigma(u)$, (which are trained using the set U_x of unpaired images) encode strong priors on the distribution p_x. SUD enforces that the network's reconstructions $f_\theta(y_u)$ on the unpaired training data are consistent with the priors encoded in the denoiser.

When used in combination with an ℓ_2 loss and without temporal-ensembling [9] or mean-teacher [10], SUD effectively minimizes $\mathcal{L}_{\text{paired}} + \lambda_1 \mathcal{L}_{\text{denoiser}}$, where λ_1 is a scalar weight and

$$\mathcal{L}_{\text{denoiser}} = \frac{1}{|U_y|} \sum_{y_u \in U_y} \|f_\theta(y_u) - D_\sigma(f_\theta(y_u))\|^2. \tag{2}$$

When updating the network weights θ to minimize (2), SUD treats $D_\sigma(f_\theta(y_u))$ as a fixed pseudo-label and does not propagate gradients through the denoiser. That is, SUD defines the gradient of $\mathcal{L}_{\text{denoiser}}$ with respect to a single reconstruction $f_\theta(y_u)$ as

$$\nabla_{f_\theta(y_u)} \mathcal{L}_{\text{denoiser}} = \frac{2[f_\theta(y_u) - D_\sigma(f_\theta(y_u))]}{|U_y|}. \tag{3}$$

As demonstrated in [4], SUD is a powerful and effective semi-supervised learning technique in the context of medical segmentation, where the goal is to map an image to a discrete-valued segmentation map. Using only a handful paired images and segmentation maps, Young et al. were able to train a denoiser to segment brains, kidneys, and tumors.

3 Improving SUD

Unfortunately, we find that without modification, SUD, with or without temporal ensembling, is far less effective at general image restoration tasks. In particular, minimizing the SUD loss for general inverse tasks almost always led to mode collapse. To better understand and overcome these weaknesses, we provide the following analyses of SUD.

Figure 2: **CelebAHQ-Mask in-painting results.** Supervised results are trained using 5 paired faces while semi-supervised methods are trained on an additional 1000+12,500 unpaired faces. The forward model (the mask) is unknown, which greatly complicates the reconstruction task.

Theorem 3.1 *When D_σ is a minimum-mean-squared error (MMSE) denoiser, minimizing $\mathcal{L}_{denoiser}$ minimizes the cross entropy between the distribution of $f_\theta(y_u)$ and the smoothed version of p_x.*

Let ν follow an independent zero-mean white Gaussian distribution with variance σ^2. We will use ν to smooth the distributions p_x (recall $p_{x+\nu} = p_x * p_\nu$, where $*$ denotes convolution). The cross entropy $H(\cdot,\cdot)$ between $p_{f_\theta(y)}$ and $p_{x+\nu}$ is, by definition, $-\mathbb{E}_{f_\theta(y)}[\ln p_{x+\nu}(f_\theta(y))]$. We can form a Monte-Carlo approximate of this expectation by averaging over U_y:

$$H(p_{f_\theta(y)}, p_{x+\nu}) \approx -\frac{1}{|U_y|} \sum_{y_u \in U_y} \ln p_{x+\nu}(f_\theta(y_u)). \tag{4}$$

Then, we can express the gradient of this loss with respect to a reconstruction $f_\theta(y_u)$ as

$$\nabla_{f_\theta(y_u)} H(p_{f_\theta(y)}, p_{x+\nu}) \approx -\frac{\nabla_{f_\theta(y_u)} \ln p_{x+\nu}(f_\theta(y_u))}{|U_y|}. \tag{5}$$

To efficiently evaluate (5) we turn to Tweedie's Formula [11], which states that for a signal corrupted with zero-mean additive white Gaussian noise, $r = x + \nu$ where $\nu \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, the output of a MMSE denoiser $D_\sigma(\cdot)$ (and by extension a neural network trained to act as a MMSE denoiser) can be expressed as

$$D_\sigma(r) = r + \sigma^2 \nabla_r \ln p_{x+\nu}(r). \tag{6}$$

In other words, denoisers perform gradient ascent on the log-likelihood of $p_{x+\nu}$ where the step size corresponds to the noise variance. By applying Tweedie's formula to (5) we arrive at

$$\nabla_{f_\theta(y_u)} H(p_{f_\theta(y)}, p_{x+\nu}) \approx \frac{[f_\theta(y_u) - D_\sigma(f_\theta(y_u))]}{\sigma^2 |U_y|}. \tag{7}$$

Up to constants, this is the same expression as the gradients in (3). Therefore, $\mathcal{L}_{\text{denoiser}}$ minimizes the cross entropy between $p_{f_\theta(y)}$ and $p_{x+\nu}$.

Corollary 3.2 *Minimizing $\mathcal{L}_{denoiser}$ encourages mode collapse.*

Minimizing $\mathcal{L}_{\text{denoiser}}$ minimizes the cross entropy between $p_{f_\theta(y_u)}$ and $p_{x+\nu}$. The cross entropy of $H(p, q)$ of two distributions p and q is minimized with respect to p when p is a Dirac distribution with a non-zero support where distribution q is largest, i.e., a mode.

To fight mode collapse, we introduce an additional penalty $\mathcal{L}_{\text{reg}} = \frac{\text{cov}(A,B)}{\sigma_A \sigma_B}$ which computes the normalized covariance between intermediate latent vectors outputted by the encoder block of our U-net. By minimizing $\mathcal{L}_{\text{reg}}$ over all latent

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
Supervised	18.44	0.71	0.29	0.48
CycleGAN	8.77	0.23	0.66	**0.17**
CycleGAN-SSL	11.38	0.60	0.39	1.14
SUD	11.28	0.29	0.69	3.07
SUD^2 (Ours)	**18.71**	**0.71**	**0.28**	0.31

Figure 3: **In-painting metrics.** Average test scores attained by each method are listed above.

vectors in a mini-batch, we encourage network to produce outputs which are uncorrelated in latent space. Therefore, our updated loss function becomes

$$\mathcal{L}_{\text{paired}} + \lambda_1 \mathcal{L}_{\text{denoiser}} + \lambda_2 \mathcal{L}_{\text{reg}}. \tag{8}$$

Corollary 3.3 *Minimizing $\mathcal{L}_{denoiser}$ can encourage blurry reconstructions.*

Minimizing $\mathcal{L}_{\text{denoiser}}$ minimizes the cross entropy between $p_{f_\theta(y_u)}$ and $p_{x+\nu}$ and will result in solutions $f_\theta(y_u)$ that maximize $p_{x+\nu}(f_\theta(y_u))$. For sufficiently large σ, $p_{x+\nu}$ is maximized not where p_x is large (along the manifold of natural images) but rather at some point in between high-probability points.

To alleviate this problem, we inject noise onto the reconstructions $f_\theta(y_u)$ before passing them through the denoiser. That is, we redefine $\nabla_{f_\theta(y_u)} \mathcal{L}_{\text{denoiser}}$ as

$$\nabla_{f_\theta(y_u)} \mathcal{L}_{\text{denoiser}} = \frac{2[f_\theta(y_u) - D_\sigma(f_\theta(y_u) + \nu_2)]}{|U_y|}, \tag{9}$$

where $\nu_2 \sim N(0, \sigma_2^2 \mathbf{I})$. This simple modification allows us to compare the smoothed distribution $p_{f_\theta(y)+\nu_2}$ with the smoothed distribution $p_{x+\nu}$.

3.1 Diffusion models

An alternative interpretation of denoising algorithms is that they project the reconstructions onto a manifold $\mathcal{M}$ of allowable reconstructions, e.g., faces or "natural images". Traditional denoising algorithms perform this projection in a single step. However, existing theory [12] suggests that one should navigate image manifolds gradually, in a smooth-to-rough/coarse-to-fine manner.

Loosely inspired by this observation, we propose replacing our single-step MMSE denoising algorithm with a multi-step denoising diffusion probabilistic model (DDPM) [13]. That is, we replace our denoiser $D_\sigma(\cdot)$ from (9) with an iterative forward "noising" operator $F(\cdot)$ and an iterative reverse "denoising" operator $R(\cdot)$ such that $D_\sigma(f_\theta(y_u) + \nu) = R(F(f_\theta(y_u)))$. Conceptually, the DDPM serves to first project r onto the smooth manifold of noisy images and then gradually project r onto correspondingly less smooth manifolds of less noisy images.

4 Experiments

Image in-painting is a generative process for reconstructing missing regions of an image such that restored image fits a desired—often natural—image distribution. We test our method on the CelebAMask-HQ dataset [14], which contains 30,000 images and their corresponding segmentation maps. In this experiment, we mask out the subject's face from each image and train a few shot, semi-supervised in-painting network on 5 paired images and 1000+12,500 unpaired images. A U-net is trained using 1,000 unpaired images, while a denoiser and diffusion model are pre-trained on 12,500 unpaired images. As an additional baseline, we train an image reconstruction network using a pre-trained CycleGAN [15] as a pseudo-label generator, which we refer to as CycleGAN-SSL.

Quantitatively, SUD^2 achieves the most consistent results across our test set of 768 images, with a 48% higher average PSNR over CycleGAN-SSL and a 43% lower average FID score compared to the supervised baseline. Although CycleGAN achieves the best FID score overall, it tends to produce qualitatively poor faces with disproportionately sized features, which is reflected in its poor PSNR, SSIM, and LPIPS scores. Likewise, while the supervised baseline achieves PSNR, SSIM, and LPIPS scores comparable to SUD^2, it often generates faces with missing features (i.e. eyes, nose, mouth), which is indicated by its high FID score. Notably, as described in Corollary (3.2), the SUD baseline collapses to a mode during training with high probability, yielding highly correlated reconstructions.

5 Conclusion

We introduce SUD^2, a generalized deep learning framework for solving few-shot, semi-supervised image reconstruction problems. Inspired by the recent success of denoising diffusion models on image generation tasks, we leverage diffusion models to regularize network training, encouraging solutions that lie close to the desired image distribution.

Acknowledgments and Disclosure of Funding

M.C. and C.M. were supported in part by the AFOSR Young Investigator Program Award FA9550-22-1-0208 and a Northrop Grumman seed grant.

References

[1] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

[2] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. I-haze: a dehazing benchmark with real hazy and haze-free indoor images. In *arXiv:1804.05091v1*, 2018.

[3] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *IEEE Conference on Computer Vision and Pattern Recognition, NTIRE Workshop*, NTIRE CVPR'18, 2018.

[4] Sean I Young, Adrian V Dalca, Enzo Ferrante, Polina Golland, Bruce Fischl, and Juan Eugenio Iglesias. Sud: Supervision by denoising for medical image segmentation. *arXiv preprint arXiv:2202.02952*, 2022.

[5] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

[6] Karen Egiazarian, Alessandro Foi, and Vladimir Katkovnik. Compressed sensing image reconstruction via recursive spatially adaptive filtering. In *2007 IEEE International Conference on Image Processing*, volume 1, pages I–549. IEEE, 2007.

[7] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013.

[8] Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016.

[9] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[10] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1195–1204, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[11] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106:1602–1614, 12 2011. doi: 10.1198/jasa.2011.tm11181.

[12] Michael B Wakin, David L Donoho, Hyeokho Choi, and Richard G Baraniuk. High-resolution navigation on non-differentiable image manifolds. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–1073. IEEE, 2005.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[14] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.