Improving Denoising Diffusion Probabilistic Models via Exploiting Shared Representations

Delaram Pirhayatifard, Mohammad Taha Toghani, Guha Balakrishnan, César A. Uribe

2023 57th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-2574-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59524.2023.10476867

***Abstract*— In this work, we address the challenge of multi-task image generation with limited data for denoising diffusion probabilistic models (DDPM), a class of generative models that produce high-quality images by reversing a noisy diffusion process. We propose a novel method, SR-DDPM, that leverages representation-based techniques from few-shot learning to effectively learn from fewer samples across different tasks. Our method consists of a core meta architecture with shared parameters, i.e., task-specific layers with exclusive parameters. By exploiting the similarity between diverse data distributions, our method can scale to multiple tasks without compromising the image quality. We evaluate our method on standard image datasets and show that it outperforms both unconditional and conditional DDPM in terms of FID and SSIM metrics.**

I. Introduction

Diffusion models are a class of generative models that produce high-quality images by reversing a noisy diffusion process [1]. They have shown several advantages over previous state-of-the-art generative models such as GANs [2], such as their scalability and their ability to capture the underlying structure of the data, including the spatial relationships between different objects [3]. This enables them to generate images that are more realistic and diverse than those produced by other generative models [4], [5]. These advances have made diffusion models powerful and useful tools for generating images and other complex data for various applications, such as computer vision [6], [7], natural language processing [4], [8], [9], artistic image generation [10], medical image reconstruction [11], and music generation [12].

Diffusion models are based on non-equilibrium thermodynamics [1], [13], where diffusion increases the system's entropy. They generate samples by gradually introducing random noise to data and learning to reverse the process to obtain the desired data samples. However, diffusion models have some limitations, such as being time-consuming and computationally expensive to train [14], being difficult to troubleshoot and scale to large datasets [15], [16], and having high-dimensional latent variables similar to the original data.

Few-shot learning is a type of meta-learning [17] that enables models to learn from a limited amount of data [18], [19]. This is especially useful in scenarios where the data availability is scarce or the training costs and time are high [18]. In such cases, few-shot learning can be used to quickly learn from a small number of examples. Several optimization-based and hierarchical-based techniques have been proposed to enable meta-learning for different problems. These techniques allow researchers to make more accurate predictions and to better understand the underlying structure of data. Few-shot learning is also a powerful tool for image generation in limited data setups. It can be used to create a variety of images from a small dataset, such as images of a specific object in different poses or environments. This can be a useful tool for data augmentation, as well as for creating new images for different tasks.

In this work, we study image generation in multi-task setups with limited data per task. We propose to enhance the quality of image generation in diffusion models by leveraging the idea of shared and personalized representations. We introduce a novel hierarchical-based algorithm called *Shared-Representation Denoising Diffusion Probabilistic Model* (SR-DDPM), which exploits a combination of shared and exclusive features [20] to improve the sample fidelity under limited data regimes. We discuss how our method is capable of fast and light fine-tuning, as well as better scalability to unseen tasks, i.e., data from a new category. We evaluate the performance of SR-DDPM on four standard datasets: MNIST [21], Fashion-MNIST (FMNIST) [22], CIFAR-10 [23], and CIFAR-100 [23] under limited data samples.

The rest of this paper is structured as follows. Section II reviews the related works. Section III introduces

Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. This work is supported by the National Science Foundation under Grants #2211815 and #2213568. Email addresses: {dp43, mttoghani, guha, cauribe}@rice.edu.

the problem setup and our method, *SR-DDPM*, for improving the performance of DDPMs using a mixture of shared and exclusive layers. Section IV presents the numerical results and Section V concludes the paper.

II. Background

Recent advances in diffusion models have focused on improving the quality and efficiency of the generated images in various ways. For instance, [24], [25] introduced the concept of noise-conditioned score networks, which learn the corresponding noise for two consecutive images in the diffusion process. Rombach et al. [16] proposed a two-stage approach to distinguish the imperceptible details in high-quality photos via adversarial auto-encoders, which reduce the size of latent DDPM. Moreover, some works proposed non-Markovian and operator learning techniques for implicit fast sampling [14], [26], [27].

Ho et al. [28] found that cascaded diffusion models were capable of generating high-fidelity images without the assistance of auxiliary image classifiers. There have also been recent attempts to boost image quality by incorporating conditional approaches that use noise prediction [29], [30]. In recent studies, broader corruption processes such as blurring, pixelation, and desaturation have also been considered in training and sampling diffusion models [31], [32].

Furthermore, there has been a growing focus on score-based generative modeling using stochastic differential equations (SDE) [33]–[35], where the goal is to learn score functions, gradients of log probability density functions, on a wide range of noise-perturbed data distributions, and then sample with Langevin-type methods. Additionally, several exceptional efforts have been made for cases with multi-modal datasets and 3D image generation [4], [16], [34]. This has been achieved by incorporating additional information about the data, such as object labels or scene context, via using attention layers in the model [36].

III. Problem Setup & Algorithm

In this section, we first describe the underlying problem setup for few-shot image generation. Then after reviewing the notion and formulation of DDPM [24], we present our method, SR-DDPM.

Data Setup: We consider a set of n different tasks $\{\mathcal{T}_i\}_{i=1}^n$, where for each task $i \in [n] = \{1, 2, \dots, n\}$, there exist a set of m_i samples $\mathcal{S}_i = \{\mathbf{x}_i^j\}_{j=1}^{m_i}$, where each $\mathbf{x}_i^j \sim \mathcal{D}_i$ is an image in a d-dimensional space ($d = 32\times32\times3$ for CIFAR-10 [23]). In the conventional setup for diffusion models, the underlying mechanism is to aggregate all samples $\mathcal{S} = \cup_{i=1}^n \mathcal{S}_i$ irrespective of their task and every $\mathbf{x} \in \mathcal{S}$ is a realization of some generic (global) distribution $\mathbf{x} \sim \mathcal{D}$. *In this research, we unravel how to exploit the combination of diverse yet similar distributions $\mathcal{D}_i$ to improve the quality of image generation in diffusion models.*

We start by stating the problem setup for DDPM and then introduce our method for shared representations.

DDPM: Let $\mathbf{x}_0 \in \mathbb{R}^d$ be an image sampled from distribution $\mathcal{D}$. Moreover, let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ denote T latent variables where each $\mathbf{x}_t \in \mathbb{R}^d$, for all $t \in [T]$. The forward (diffusion) process q can be defined as follows:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2}$$

where $\beta_1, \beta_2, \dots, \beta_T$ is a variance schedule for the underlying gaussian noise with mean $\sqrt{1-\beta_t}\mathbf{x}_{t-1}$ and variance $\beta_t\mathbf{I}$ at each timestep t. According to [24], [37], the latent variable $\mathbf{x}_t$ can be directly derived based on the observed data $\mathbf{x}_0$ as

$$\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, \tag{3}$$

where $\overline{\alpha}_t \coloneqq \prod_{s=1}^{t} \alpha_s$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Moreover, the reverse (generative) process p_ψ, parameterized by a set of parameters ψ, can be summarized as follows:

$$p_\psi(\mathbf{x}_{0:T}) \coloneqq \prod_{t=1}^{T} p_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{4}$$

$$p_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\psi(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}\right), \tag{5}$$

$$\boldsymbol{\mu}_\psi(\mathbf{x}_t, t) \coloneqq \frac{1}{\sqrt{\alpha_t}}\left[\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}_\psi(\mathbf{x}_t, t)\right], \tag{6}$$

where $\sigma_t^2 = \beta_t$, and $\boldsymbol{\epsilon}_\psi : \mathbb{R}^d \times \mathbb{N} \to \mathbb{R}^d$ is a neural network with parameters ψ that takes $\mathbf{x}_t$ and timestep t as inputs and estimates the realization of $\boldsymbol{\epsilon}$ in (3). For example, UNet with attention is a proper candidate for $\boldsymbol{\epsilon}_\psi$. In [24], it is explained that $\sigma_t^2 = \tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$ provides similar experimental results to $\sigma_t^2 = \beta_t$. Note that the underlying assumption in this formulation is that $\boldsymbol{\epsilon}_\psi(\mathbf{x}_t, t)$ is a shared model across the Markov chain (from 0 to T) which is expressive enough to recover the noise value. Therefore, it is sufficient to optimize the network parameters with respect to some

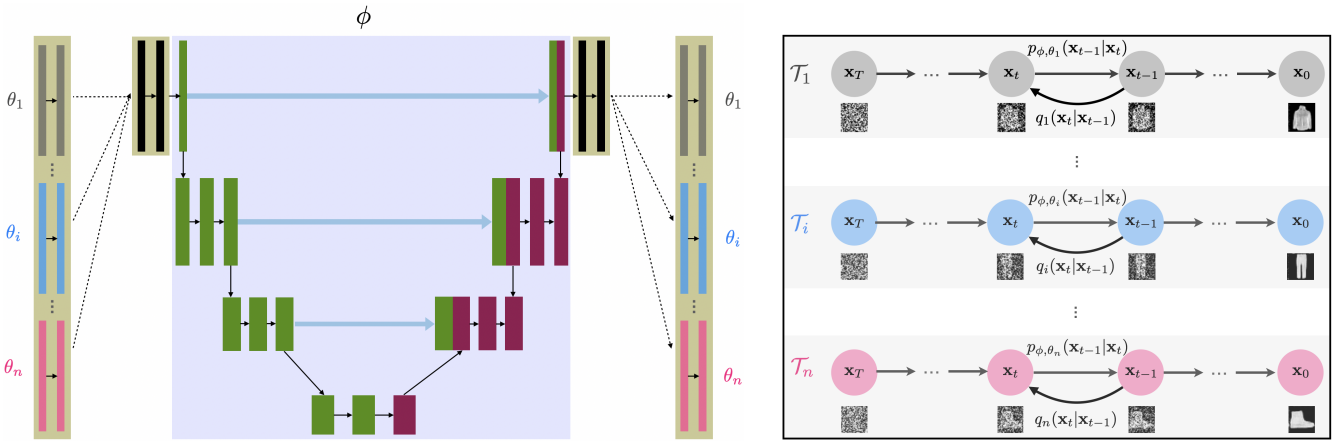


Fig. 1: **(Left)** UNet architecture contains a mixture of shared and exclusive layers. **(Right)** SR-DDPM for a mixture of n exclusive but similar distributions.

loss function $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, i.e., minimizing

$$\mathcal{L}\left(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\psi(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right), \tag{7}$$

which quantifies the distance between the original noise and prediction of the denoising model.

Next, we explain SR-DDPM for multi-task denoising diffusion models with shared and exclusive representations.

SR-DDPM: Our goal is to exploit the exclusiveness of each task $\{\mathcal{T}_i\}_{i=1}^n$ by splitting the denoising network architecture into shared and personal (exclusive) layers. Figure 1 depicts the UNet structure that uses a common set of parameters ϕ for all tasks $i \in [n]$, and a distinct set of parameters $\{\theta_i\}_{i=1}^n$ for each task. The set of parameters ψ in (4) is the combination of ϕ and θ_i, for any $i \in [n]$. This allows us to jointly capture both shared and unshared features. For example, Figure 1 shows that for different tasks involving various outfits, we train and sample from n parallel Markov chains with shared parameters ϕ and exclusive parameters $\{\theta_i\}_{i=1}^n$. In other words, we minimize the following:

$$\mathbb{E}_i\left[\mathcal{L}\left(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_{\phi,\theta_i}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0^i + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right)\right], \tag{8}$$

where $\mathbf{x}_0^i \sim \mathcal{D}_i$ and $i \sim \text{Uniform}([n])$. Algorithms 1 and 2 respectively describe the training and sampling processes of SR-DDPM. As shown in Algorithm 1, at the training phase, we randomly choose a task and an image from that task. Then, we use a first-order optimization method such as Adam [38] to optimize the stochastic gradient and minimize the cost in (8). Finally, we generate samples by feeding a noise signal to the network and applying the denoising process of Algorithm 2.

Algorithm 1 SR-DDPM: Training

1: **repeat**
2: $i \sim \text{Uniform}([n])$ {select a task $\mathcal{T}_i$ from n tasks}
3: $\mathbf{x}_0 \sim \mathcal{D}_i$ {sample data $\mathbf{x}_0$ from task $\mathcal{T}_i$}
4: $t \sim \text{Uniform}([T])$ {sample timestep t}
5: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
6: Compute the gradient and apply one step of optimizer:

$$\nabla_{\phi,\theta_i}\mathcal{L}\left(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_{\phi,\theta_i}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right)$$

7: **until** converged

IV. Experiments

In this section, we describe the experimental setup and the results of our proposed method. We compare our method with unconditional and conditional DDPM.

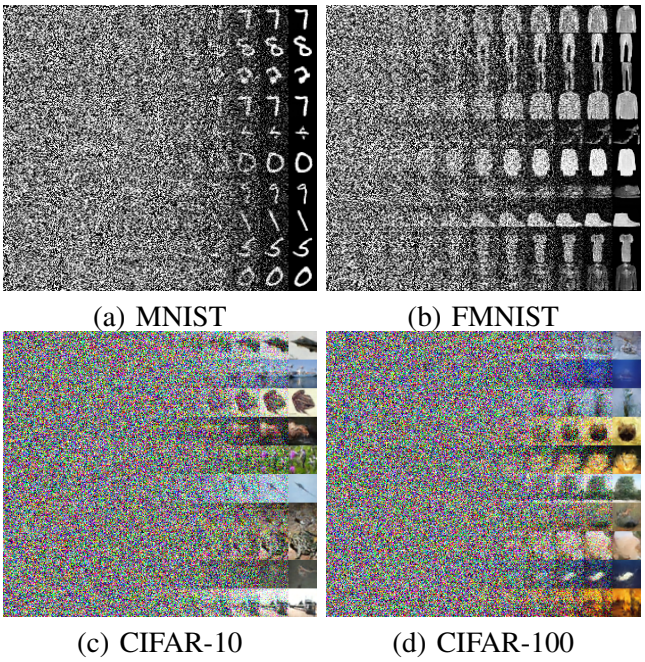
We consider four standard datasets: MNIST, FMNIST, CIFAR-10, and CIFAR-100. We implement a multi-task scheme with 500 samples per task. Following Section III, we adopt a UNet as the denoising network with four layers for all methods. The network

Algorithm 2 SR-DDPM: Sampling

1: $i \sim \text{Uniform}([n])$ {select a task $\mathcal{T}_i$ from n tasks}
2: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ {sample a noise signal $\mathbf{x}_T$}
3: **for** $t = T, \ldots, 1$ **do**
4: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
5: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\phi,\theta_i}(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}$
6: **end for**
7: **return** $\mathbf{x}_0$

TABLE I: Comparison of the performance of Denoising Diffusion Probabilistic Models (DDPM), Conditional DDPM (C-DDPM), and Shared-Representation DDPM (SR-DDPM) for T=500, 4-layer UNet with one exclusive layer, linear β schedule, and 600 training epochs.

Dataset	$\lvert\{\mathcal{T}_i\}_i\rvert$	Method	FID ↓	SSIM ↑
MNIST	10	DDPM	3.67	0.881
		C-DDPM	2.14	0.884
		SR-DDPM	**2.04**	**0.887**
FMNIST	10	DDPM	4.80	0.908
		C-DDPM	2.72	**0.915**
		SR-DDPM	**2.48**	0.909
CIFAR-10	10	DDPM	12.64	0.946
		C-DDPM	12.86	**0.949**
		SR-DDPM	**10.87**	**0.949**
CIFAR-100	20	DDPM	13.74	**0.944**
		C-DDPM	11.54	0.942
		SR-DDPM	**11.30**	**0.944**



(a) MNIST (b) FMNIST

(c) CIFAR-10 (d) CIFAR-100

Fig. 2: The process of diffusion image reconstruction using SR-DDPM. For each dataset, we generate 10 samples with $T = 500$ and visualize the reconstructed image at $t = 0, 50, 100, \ldots, 500$.

has a bottleneck in the middle to learn only the most important features of the data. We increase the number of channels by a factor of two and decrease the image size by the same factor per layer. We personalize one layer as the exclusive stage at the first and end of the network for all datasets. We train the model for each method within 600 epochs. We use the Adam optimizer with a learning rate of 5×10^{-4} for all experiments.

We quantitatively compare the performance of our model with DDPM and Conditional DDPM (C-DDPM) using the implementation of DDPM [24] on Hugging Face [39]. We measure the performance of SR-DDPM on the four different datasets using sample quality (FID@10k) and structural similarity (SSIM) on test data. Table I compares SR-DDPM with unconditional and conditional DDPM. Our method achieves better FID scores than the other two on all four datasets.

Figure 2 visualizes the reverse process for image generation for $T = 500$ on all datasets. We also display 20 samples from each task in [40]. The images generated from FMNIST show that the trained model can identify similarities between the tasks. Some of the images generated from one task overlap with the other task when using the corresponding exclusive layers. This implies that the method can capture the similarity between the tasks implicitly by using the shared and exclusive layers.

V. Conclusion

We presented a novel algorithm for training diffusion models with limited data. Our method outperforms unconditional and conditional DDPM on image generation in the same training time. We also found that the personal layer for each task can detect similarities among tasks automatically. This means that we can train a new personal layer for a new task without fine-tuning the whole network. Our method offers an interpretable way to generate images by using a combination of shared and unshared parameters to capture the differences among tasks.

References

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

[2] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[3] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[6] T. Amit, E. Nachmani, T. Shaharbany, and L. Wolf, "Segdiff: Image segmentation with diffusion probabilistic models," *arXiv preprint arXiv:2112.00390*, 2021.

[7] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[8] N. Savinov, J. Chung, M. Binkowski, E. Elsen, and A. v. d. Oord, "Step-unrolled denoising autoencoders for text generation," *arXiv preprint arXiv:2112.06749*, 2021.

[9] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *arXiv preprint arXiv:2205.14217*, 2022.

[10] R. Rombach, A. Blattmann, and B. Ommer, "Text-guided synthesis of artistic images with retrieval-augmented diffusion models," *arXiv preprint arXiv:2207.13038*, 2022.

[11] C. Peng, P. Guo, S. K. Zhou, V. M. Patel, and R. Chellappa, "Towards performant and reliable undersampled mr reconstruction via diffusion model sampling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 623–633.

[12] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," *arXiv preprint arXiv:2103.16091*, 2021.

[13] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[14] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[15] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," *arXiv preprint arXiv:2206.00941*, 2022.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[17] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[18] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial intelligence review*, vol. 18, no. 2, pp. 77–95, 2002.

[19] E. Robb, W.-S. Chu, A. Kumar, and J.-B. Huang, "Few-shot adaptation of generative adversarial networks," *arXiv preprint arXiv:2010.11943*, 2020.

[20] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2089–2099.

[21] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[22] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[23] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

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

[26] Q. Zhang, M. Tao, and Y. Chen, "gddim: Generalized denoising diffusion implicit models," *arXiv preprint arXiv:2206.05564*, 2022.

[27] H. Zheng, W. Nie, A. Vahdat, K. Azizzadenesheli, and A. Anandkumar, "Fast sampling of diffusion models via operator learning," *arXiv preprint arXiv:2211.13449*, 2022.

[28] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation." *J. Mach. Learn. Res.*, vol. 23, pp. 47–1, 2022.

[29] G. Giannone, D. Nielsen, and O. Winther, "Few-shot diffusion models," *arXiv preprint arXiv:2205.15463*, 2022.

[30] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[31] G. Daras, M. Delbracio, H. Talebi, A. G. Dimakis, and P. Milanfar, "Soft diffusion: Score matching for general corruptions," *arXiv preprint arXiv:2209.05442*, 2022.

[32] A. Bansal, E. Borgnia, H.-M. Chu, J. S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein, "Cold diffusion: Inverting arbitrary image transforms without noise," *arXiv preprint arXiv:2208.09392*, 2022.

[33] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 287–11 302, 2021.

[34] X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis, "Lion: Latent point diffusion models for 3d shape generation," *arXiv preprint arXiv:2210.06978*, 2022.

[35] T. Dockhorn, T. Cao, A. Vahdat, and K. Kreis, "Differentially private diffusion models," *arXiv preprint arXiv:2210.09929*, 2022.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] N. Rogge and K. Rasul, "The annotated diffusion model," *Hugging Face Blog*, 2022, https://huggingface.co/blog/annotated-diffusion.

[40] D. Pirhayatifard, M. T. Toghani, G. Balakrishnan, and C. A. Uribe, "Improving denoising diffusion probabilistic models via exploiting shared representations," 2023.