Exploring Synthetic Data Generation using Diffusion Models

Arunava Basu [1] **Rahul Karanam** [1] **Aditi Ramadwar** [1]

Abstract

Although large datasets of labeled images are necessary for deep learning's greatest triumphs in computer vision, obtaining and annotating such a large amount of data is frequently prohibitively expensive or impractical in real-world applications. One interesting approach is to deploy models in real-world scenarios after training them on synthetic data for which we already know the correct labels. Unfortunately, supervised learning methods struggle when the distributions of the training and test data diverge. Performance is considerably decreased by the minute variations between real and synthetic data. To explore how synthetic data can be used to enhance and augment existing unbalanced datasets, our work looks towards two predominant avenues - the effect of synthetically augmented datasets on downstream tasks such as image classification, measuring the fidelity of said generated synthetic data

1. Introduction

Training modern training models for computer vision (CV) continues to require a lot of data. On the back of ImageNet, a massive dataset with one million photos, and 1000 representatives from each of 1000 categories, a remarkable advance in object recognition was achieved. Creating datasets of this magnitude needs both acquisition (of relevant photos) and annotation (usually by humans). When data cannot be passively acquired, acquisition alone may be a significant investment. Annotation is generally expensive. It may be expensive or impossible for a small research team or business to produce its own enormous datasets. Data availability is frequently a significant barrier to using machine learning algorithms, even for large businesses.

Parallel to this, high-fidelity photorealistic images may now be created via generative models that try to simulate real-data distributions. Recent text-to-image generation models have made enormous strides, in particular, in the creation of excellent images from text descriptions. Generative models work by trying to estimate the underlying data distribution of datasets, the ability to properly do enables us to sample the estimated distribution and generate novel samples. GANs have previously been used to generate synthetic data for image recognition tasks. None have explored the recent revolutionary class of generative modeling, diffusion models, which hold more promise to benefit recognition tasks.

In this work, we explore using the recent class of Denoising Diffusion Probabilistic Models to augment datasets with synthetic images and measure their effects on downstream image recognition tasks.

2. Related Work

GAN or the rendering of 2D or 3D models using graphics engines has been the main method over the years for creating synthetic datasets. The disparity between generated data and data from the real world and the need for extensive resources to produce these datasets are two major problems for graphics models .(Lea, 2018) have used Blender to generate photo-realistic images which are later used for downstream tasks such as Image recognition.

GAN-generated synthetic datasets are much more realistic than the images produced by graphic engines. GANs produce more realistic images and can be conditioned based on the downstream task, which helps them sample from low-density regions and generate high-fidelity images. (Mariani et al., 2018), which generates images from the low-density regions to balance the dataset distribution, is one of the few works that have used GANs to generate synthetic datasets. For tasks further down the line, such as object segmentation, (Zhang et al., 2021) have used synthetic labeled data. (Jahanian et al., 2021) employed contrastive learning to produce multi-view images using GAN. As we can see, using GAN has produced better synthetic images. However, there is still room to produce a wider variety of images that are quicker to produce and more photorealistic.

The popularity of generative models has grown as a result of recent developments in text-image generation, which creates realistic images from a text prompt using diffusion models(Sohl-Dickstein et al., 2015). Given their advantages over GAN models, researchers frequently use these models for conditional and unconditional generation in image synthesis. These models have the tendency to capture the underlying data distribution of the given dataset and produce

diverse samples with high sample quality and good mode coverage.

The diffusion sampling process is used by (Sehwag et al., 2022)to produce high-fidelity images by using the underlying representation of low-density regions. (Nichol and Dhariwal, 2021)directs the sampling procedure towards a specific class for generation using class-conditional classifier guidance. Despite being state-of-the-art generative models with amazing results, the creation of synthetic datasets using these models is still largely unexplored. We will experiment in this paper with creating synthetic datasets using cutting-edge diffusion models.

3. Methodology

We aim to create datasets that make use of synthetic samples as well as real-life samples. To achieve this, we carry out experiments using conditional and unconditional sampling on the diffusion models. We evaluate a CNN model trained on the long-tailed dataset with class-specific synthetic generated samples. We also observe the fidelity of synthetic images by replacing the real images with them in a balanced dataset.

3.1. Long Tailed CIFAR 10

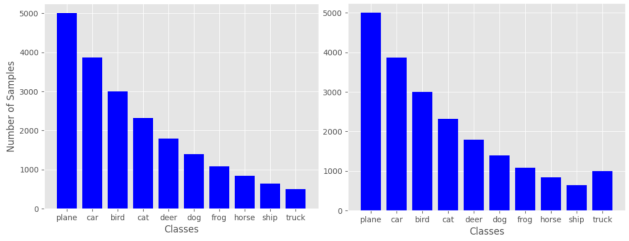

Figure 1. CIFAR-LT(Left), Samples distribution of CIFAR-LT after augmenting 500 synthetic samples on the truck class(Right)

$$LT = N_i * \lambda^{\left(\frac{i}{10-1}\right)} \tag{1}$$

To achieve a balanced dataset using our synthetically generated images, it was required to acquire an imbalanced dataset, to begin with. We carried out our experiments using the CIFAR-10 dataset (Canadian Institute For Advanced Research) with 10 classes. This dataset is inherently balanced with 6000 images per class, which is split into 5000 images per class for the train set and 1000 images per class for the test set.

An imbalanced dataset in the form of a long tail is created using the CIFAR-10 train set. This is carried out by decreasing the number of training examples per class with an exponential function. In order to showcase our approach, a synthetic long-tailed CIFAR-10 dataset (CIFAR-LT) is created by decreasing the number of training examples per class with an exponential function $LT = Ni^{*}\mu$i where LT is the modified number of samples, i is the class-index, Ni is the original number of training images for the ith class and $\mu \in (0, 1)$, $\lambda = 0.1$. Hence, Equation 1 is used to generate the CIFAR-LT dataset shown in the left of Figure 1.

Table 1. Training setup for experiments augmenting the long-tailed CIFAR-10 dataset

PARAMETER	VALUE
EPOCHS	30
BATCH SIZE	256
LEARNING RATE	0.01
OPTIMIZER	SGD

Table 2. Training setup for experiments replacing real images in the balanced CIFAR-10 dataset

PARAMETER	VALUE
EPOCHS	50
BATCH SIZE	256
LEARNING RATE	0.01
OPTIMIZER	SGD

3.2. Experimental Setup

For all our experiments we use the Denoising Diffusion Probabilistic Model as proposed in (Nichol and Dhariwal, 2021) for generating synthetic data. For the downstream task, we have chosen image classification using a standard ResNet18 model on the test set or the CIFAR-10 dataset.

For all the experiments where we augment the long-tailed CIFAR-10 dataset, we use the setup given in Table 1. to train the ResNet18. Class-conditioned versions of similar DDPMs were also trained where the input is conditioned with an embedding of the class ID. This enables us to guide the diffusion process to generate images from the required class. Conditional generation has been used in some of the experiments which are explained in more detail in the following sections.

For all the experiments where we replace real images of the balanced CIFAR-10 dataset with synthetic images, we use the setup given in Table 2. to train the ResNet18.

For making use of the synthetic data generated by unconditional sampling, we pseudo-label the images. For pseudo-labeling, we use a pre-trained Convolutional Neural Network that exhibits around 97% accuracy on the CIFAR10 dataset.

3.3. Learned Data Augmentation

3.3.1. Using Class Conditioned Model trained on ImageNet

Due to fine labeling in ImageNet and coarse labeling in CIFAR, we identified and extracted the fine labels in ImageNet corresponding to the course labels in CIFAR as shown in Table 3

CIFAR-10 Label	ImageNet Label
Truck	Fire engine, fire truck, garbage truck, dustcart, pickup, pickup truck, tow truck, tow car, wrecker, trailer truck, tractor-trailer, trucking rig, rig, articulated lorry, semi
Ship	Container ship, container vessel, pirate ship

Table 3. Mapping between coarse CIFAR labels and ImageNet fine labels

500 synthetic samples were generated using the fine labels and the class-conditioned model trained on ImageNet for the two classes with the least number of samples, Truck, and Ship. These synthetic samples were augmented on their respective classes in the CIFAR-LT dataset.

This enabled us to generate different augmented versions of the CIFAR-LT dataset such as:

- Dataset added with 500 truck images
- Dataset added with 500 ship images
- Dataset added with 500 truck and ship images

For each of these augmented datasets, we trained a ResNet18 model using the experimental setup mentioned in Table 1. The most interesting results are shown in Fig. 2 and Fig. 3.

In Fig. 2, we can see improved performance in the ship and truck classes which are the original classes that we augmented. In Fig. 3, we compare per-class accuracy for different combinations of augmentations as shown in the legend. From the graph, we can see that the classes that have been augmented in each combination clearly show an increase in performance. This result is not unexpected as the absolute number of samples has increased.

3.3.2. Using Class Conditioned & Unconditional Diffusion Model trained on CIFAR-10

In the previous section, the experiments clearly show an increase in the performance of the augmented classes since the number of training samples inadvertently has increased.

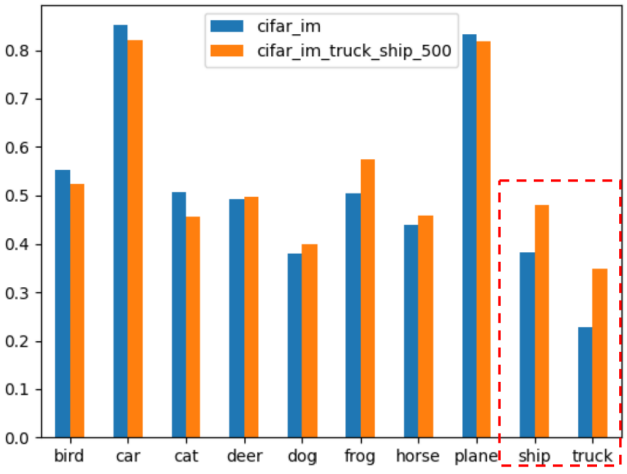

Figure 2. Comparison of evaluation results between the original CIFAR-LT(Blue), CIFAR-LT after augmenting 500 synthetic samples on the truck and ship class(Orange)

Additionally, we have also assumed that we have access to ImageNet which is a larger and more diverse dataset.

To address these concerns we run similar experiments, keeping the number of training samples the same across comparison datasets and generating synthetic samples from a diffusion model that is trained on CIFAR-10.

In this case, to keep the number of samples equal, we generate the control dataset by adding pseudo-labeled synthetic images generated from an unconditional diffusion model.

The targeted class augmentations remain the same as by using the class-conditioned diffusion model trained on CIFAR-10.

For example, the data distribution shown on the left of Figure 4 is generated by augmenting 500 synthetic samples for three classes in the CIFAR-LT dataset, dog, ship, and truck. 1500 Synthetic samples were randomly generated by unconditional diffusion model trained on CIFAR-10 and augmented to their respective classes on CIFAR-LT dataset. The unlabeled unconditional samples were given pseudo labels using the pseudo-labeler model. The distribution of this new dataset can be observed in the right of Figure 4.

This keeps the number of training samples the same when training the ResNet18 models on these augmented datasets. The ResNet18 models are trained with the setup mentioned in Table 1.

The performance of the downstream tasks trained on these datasets is shown in Figure 5. The confusion matrices for these synthetic augmentations are shown in Figure 6.

Fig 5 shows the performance of the classification task on

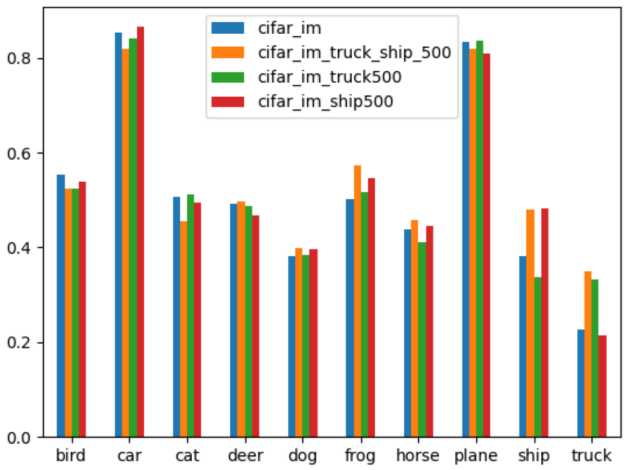

Figure 3. Comparison of evaluation results between the original CIFAR-LT(Blue), CIFAR-LT after augmenting 500 synthetic samples on the truck class(Green), CIFAR-LT after augmenting 500 synthetic samples on the ship class(Red), CIFAR-LT after augmenting 500 synthetic samples each on the truck as well as ship class(Orange)

the augmented datasets. There is no clear increase in performance for the augmented classes only as in the previous section since we have added unconditional samples to the control set as well. This shows us the emergence of shared representations between classes. Such results can be confirmed from the confusion matrix as well.

4. Fidelity of Synthetic Images

In this approach, we have used an unconditional CIFAR-10 diffusion model to generate synthetic images, and then used a pseudo-labeling model(CNN) to label the generated images. This allows you to evaluate the fidelity of the synthetic images by comparing their performance on downstream tasks, such as classification, to the performance of real images.

To do this, we have replaced a portion of the real images in the dataset with the synthetic images as shown in Figure 8 and then evaluate the performance of the classification model on the modified dataset. By comparing the performance of the model on the modified dataset to its performance on the original dataset, you can assess the fidelity of the synthetic images and determine how closely they resemble real-world images.

We have used the balanced CIFAR-10 as our original dataset and using the unconditional Diffusion model trained on CIFAR-10, we have generated 600 samples for each class. We performed an ablation study to evaluate the impact of replacing a certain percentage of the real images from the

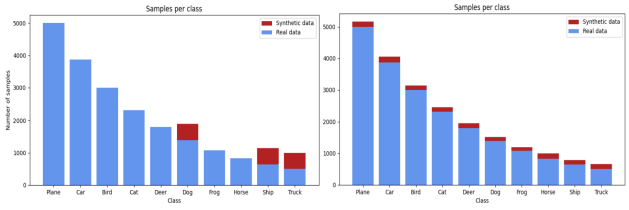

Figure 4. 500 Synthetic samples(Red) generated by conditional diffusion model trained on CIFAR-10 augmented to their respective classes on CIFAR-LT real images(Blue) (Left). Randomly generated 1500 Synthetic samples(Red) by unconditional diffusion model trained on CIFAR-10 augmented to their respective classes on CIFAR-LT real images(Blue) (Right)

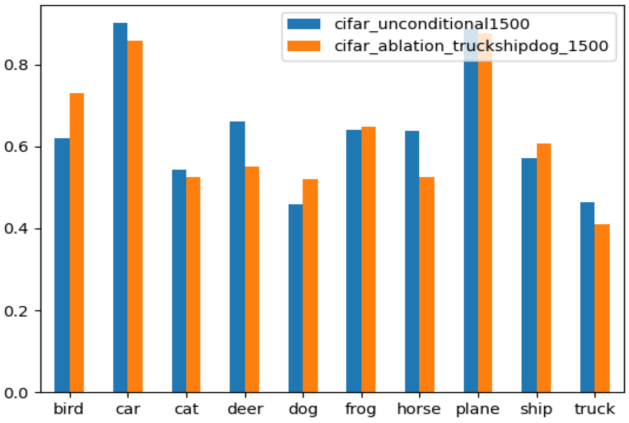

Figure 5. Per class accuracy on datasets generated by conditionally generated synthetic samples(Orange) and unconditionally generated synthetic samples(Blue)

CIFAR-10 dataset with synthetic images. We evaluated the performance of a classification model(2 on the modified dataset for different percentages of replacement, and recorded the results of the study in a table(As **??**).

4.1. Results and Analysis

Based upon the experiments in table **??**, it appears that replacing a small percentage of real images in the CIFAR-10 dataset with synthetic images generated by your CIFAR-10 diffusion model can slightly improve the performance of a classification model on the dataset. In particular, replacing 2% of the real images with synthetic images resulted in a 0.74% increase in test accuracy. However, replacing a larger percentage of the real images with synthetic images seems to have a negative impact on the model's performance, with test accuracy decreasing as the percentage of replacement increases.

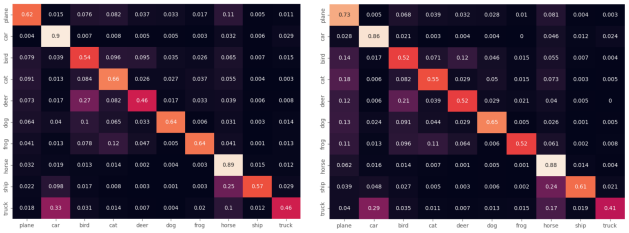

Figure 6. Confusion matrix for evaluation on the dataset generated by unconditional sampling (Left) and the same for dataset generated by conditional sampling(Right)

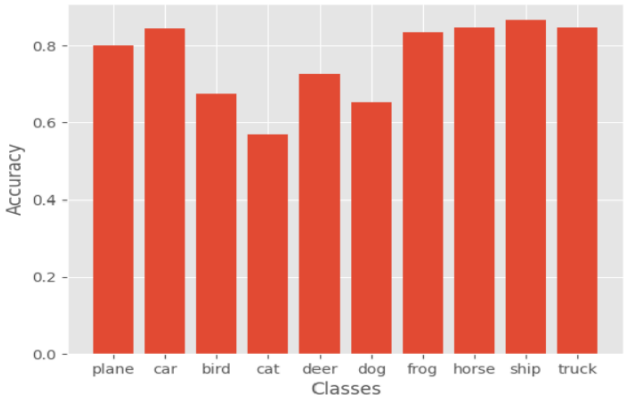

Figure 7. Accuracy Per class of ResNet model trained on Balanced CIFAR-10 dataset

We ran another experiment where we replaced the real samples in the original dataset with the generated samples and evaluated the per-class accuracy of these modified datasets. It has a mixed impact on the performance of the classification model on the dataset. For some classes, such as planes, birds, deer, and frogs, the accuracy of the model decreased slightly compared to the original dataset. For other classes, such as cars, cats, and dogs, the accuracy of the model increased slightly. For the remaining classes, horse, ship, and truck, the accuracy of the model remained relatively unchanged.

On the other hand, we can see from the confusion matrix for these two datasets that some classes have shared representation because of that we have seen some increase or decrease in the per-class accuracy in the previous experiments.

These results suggest that the synthetic images generated by your CIFAR-10 diffusion model have relatively high fidelity, as they were able to slightly improve the performance of a classification model when used in conjunction with real images. However, the performance of the model decreases when a larger percentage of the real images are replaced

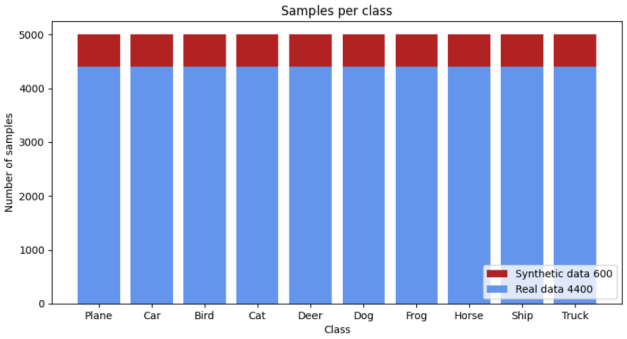

Figure 8. Replacment of 600 real images(Blue) per class with synthetic image(Red)

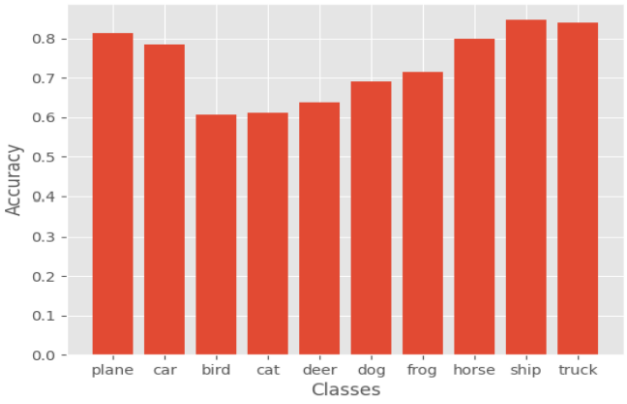

Figure 9. Accuracy Per class of ResNet model trained on the modified CIFAR-10 dataset with 600 synthetic samples per class

with synthetic images, indicating that the synthetic images may not be as high quality as the real images.

It is important to note that the performance of the classification model on the modified dataset may be lower than its performance on the original dataset, even if the synthetic images have high fidelity. This is because replacing real images with synthetic images can introduce additional variability and noise into the dataset, which can make it more difficult for the model to learn.

However, if the performance of the model decreases significantly when synthetic images are used, it may indicate that the synthetic images are not of sufficient quality to be used as a replacement for real images.

4.2. Findings

While training and sampling images from the CIFAR-10 Conditional and unconditional model, we have observed that images generated from the unconditional model have

Figure 10. Confusion matrix for evaluation on the balanced CIFAR-10 dataset (Left) and the same for dataset generated presented in Figure 8 (Right)

Training Setup	% of Replacement	Test Accuracies
Real + Synthetic	2%	**77.37**%
Real + Synthetic	4%	76.65%
Real + Synthetic	6%	75.54%
Real + Synthetic	8%	76.20%
Real + Synthetic	10%	76.20%
Real + Synthetic	12%	73.49%

Table 4. Ablation Study of CIFAR-10 (Fidelity of Synthetic Samples)

dark backgrounds as compared to class-conditional images where the backgrounds are diverse and less dark.

This has been observed after verifying with more than 9000 samples. Having a mostly dark background in the synthetic images could affect the performance of a classification model on the modified dataset, as it may be more difficult for the model to discern the details of the objects depicted in the images. This could result in a decrease in performance, particularly for classes that are difficult to distinguish based on shape or color alone. This might be one reason why our accuracy has dropped when compared with the real dataset. Please refer figure for more details.

5. Conclusion

5.1. Results

When working with Conditional sampling using ImageNet and appending synthetic data on CIFAR-LT, a jump in performance was observed for the classes that provided synthetic data. This shows that the synthetic data can be used to balance the dataset.

The fidelity of synthetic images was evaluated by replacing real images with synthetic images and analyzing the resulting performance of a classification model. Results showed that replacing a small percentage of real images

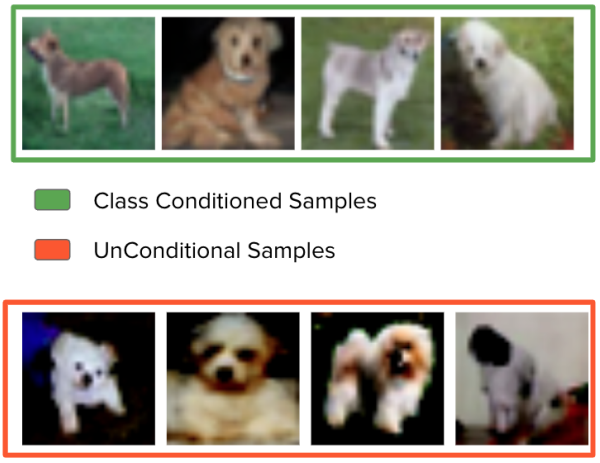

Figure 11. Sample generated using conditional vs unconditional

with synthetic images improved performance slightly, but replacing a larger percentage decreased performance.

We also observe the role of shared representation in Figure 6 and Figure 10. It can be observed that due to the augmentation of synthetic data, the model seems to have a slight bias over a few classes as compared to the other. It can be hypothesized that the diffusion model generates its synthetic data with more prominent features from classes like birds and deer.

Synthetic image characteristics, including background, also impacted performance. Further optimization may be necessary to improve synthetic image performance as a replacement for real images.

5.2. Future Work

Further investigation into the use of synthetic images as a replacement for real images in downstream tasks could include identifying the underlying cause of unexpected changes in accuracy for some classes, determining the optimal percentage of image replacement that maximizes performance, and examining the impact of synthetic image replacement on datasets to identify any newly introduced biases.

References

Learning image classifiers from (limited) real and (abundant) synthetic data. 2018.

Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan, 2018. URL https://arxiv.org/abs/1803.09655.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10140–10150, 2021. doi: 10.1109/CVPR46437.2021.01001.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning, 2021. URL https://arxiv.org/abs/2106.05258.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton Ferrer. Generating high fidelity data from low-density regions using diffusion models, 2022. URL https://arxiv.org/abs/2203.17260.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.