# Class-Balancing Diffusion Models

Yiming Qin[1]     Huangjie Zheng[2]     Jiangchao Yao[1,3]✉     Mingyuan Zhou[2]     Ya Zhang[1,3]

[1]Cooperative Medianet Innovation Center, Shanghai Jiao-Tong University
[2]University of Texas, Austin
[3]Shanghai AI Laboratory

## Abstract

*Diffusion-based models have shown the merits of generating high-quality visual data while preserving better diversity in recent studies. However, such observation is only justified with curated data distribution, where the data samples are nicely pre-processed to be uniformly distributed in terms of their labels. In practice, a long-tailed data distribution appears more common and how diffusion models perform on such class-imbalanced data remains unknown. In this work, we first investigate this problem and observe significant degradation in both diversity and fidelity when the diffusion model is trained on datasets with class-imbalanced distributions. Especially in tail classes, the generations largely lose diversity and we observe severe mode-collapse issues. To tackle this problem, we set from the hypothesis that the data distribution is not class-balanced, and propose Class-Balancing Diffusion Models (CBDM) that are trained with a distribution adjustment regularizer as a solution. Experiments show that images generated by CBDM exhibit higher diversity and quality in both quantitative and qualitative ways. Our method benchmarked the generation results on CIFAR100/CIFAR100LT dataset and shows outstanding performance on the downstream recognition task.*

## 1. Introduction

In recent years, log-likelihood-based diffusion models have evolved rapidly and established new benchmarks on a range of generation tasks [1,7]. Based on them, researchers have been able to further control the model generation process and the generation quality. This improves the applications of generative models in numerous domains including text-image generation [31], image editing [28,38], speech synthesis [17], medical imaging [26,44], video generation [13] and adversarial learning [18,33], etc.

Although diffusion models are known for the power of high fidelity and diversity in generation, most of the existing diffusion models are trained with the hypothesis that the data are uniformly distributed *w.r.t.* their labels. However,
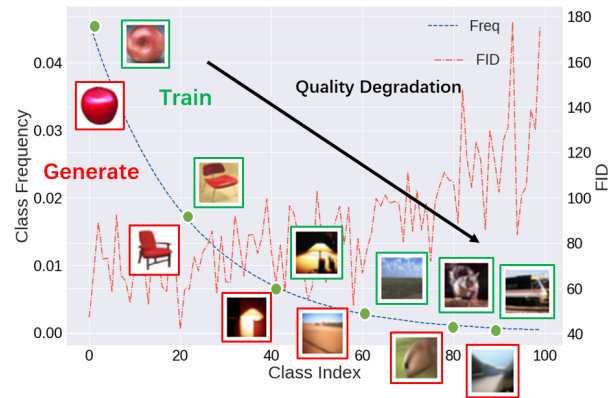


Figure 1. Generation degrades along with class frequency. Semantics of generated images become less recognizable when class frequency decreases, while the FID score increases significantly.

in the real world, the distribution is often very skewed. Especially for many domain-specific generation tasks such as medical images [14], fine-grained dataset for taxology [15] and data grabbed from the web [24], it is difficult to collect large amounts of data for each class equally, and the size of the training set for head and tail categories can differ by a factor of hundred or more. For such datasets, unconditional diffusion models tend to produce a significant portion of low-quality images. Conditional models, as shown in Figure 1, generate head class images with satisfying performance, while conversely the generated images on tail classes are very likely to show unrecognizable semantics. Concerning training generative models with limited data, there already exist several methods [21, 37, 50] based on GAN models [2]. However, quite few studies examine the impact of imbalanced class distribution [34] especially on diffusion models, which is practical yet under-explored.

Our work first introduces diffusion models to imbalance generation tasks on several long-tailed datasets [19], and then build some straightforward baselines according to the common methods used in long-tailed recognition [27, 29].

To overcome the potential degeneration induced by the skewed distribution, we propose a novel Class-Balancing Diffusion Model (CBDM). Theoretically, CBDM resorts to adjusting the conditional transfer probability during sampling in order to implicitly force generated images to have a balanced prior distribution during every sampling step. Technically, the adjusted transfer probability of CBDM results in an additional MSE-form loss for a conditional diffusion model, which functions as a regularizer. Intuitively, this loss augments the similarity of generated images conditioned on different classes, and turns out to be an effective approach to transfer common information from head classes to tail classes without hurting the model's expressiveness on head classes. CBDM can be implemented within several flines of codes, and its lighter version admits fine-tuning an existing conditional model. We conducted extensive experiments on CIFAR10/CIFAR100 and their corresponding long-tailed dataset to show the promise of CBDM over existing state-of-the-art methods. In a nutshell, the contributions of this work can be summarized as follows:

- We identify the severe degeneration problem of diffusion models in long-tailed generation tasks and benchmark some straightforward baselines in this direction.

- We propose a new perspective to handle the generation quality collapse on tail classes, and derive a novel Class-Balancing Diffusion Model, which is effective and lightweight as a regularizer to existing methods.

- We validate that CBDM is capable of generating more diverse images with convincing fidelity, especially for datasets with large number of categories. In addition, CBDM is robust with accelerating algorithms such as DDIM [43], and can be transplanted to different conditional diffusion-based backbones easily.

## 2. Related Works

**Diffusion models** Diffusion models are recently proposed generative models [42] based on non-equilibrium thermodynamics. Conditional diffusion models [7] encode label information into the generation process and improve largely the generation performance. The guidance structure proposed in [7] makes it possible to control the generation process through an external module. Based on a similar intuition, researchers arrive to realize diverse functions, such as guided adversarial purification [33], few-shot generation [9] and so on [36, 41]. The drawback of classifier guidance (noted as CG) [7] lies in its requirement of training another auxiliary classifier. To address the issue, classifier-free guidance (noted as CFG) [12] proposed a mechanism that uses the generator itself to express the class guidance information. CFG is proved to be a resource efficient method

and achieves outstanding performance on large models [31]. Moreover, CFG only requires to add one line in training, which can be easily transplanted on different models.

**Long-tail recognition** The problem of long-tailed distribution is a common dilemma in machine learning and have been widely explored in the area of discriminative models, *i.e.,* long-tailed recognition. There mainly exist three paradigms in this domain, namely Class Re-balancing [27,29], Information Augmentation [25,47,49] and Module Improvement [16, 32]. Among them, Class Re-balancing provides the best explainability, and its most common practice is re-sampling [27]. Thereafter, stemming from modifying the objective function from the global error rate to the class average one, [29] propose logit adjustment which has shown an impressive performance. Another effective method is based on Information Augmentation, which uses head class feature information to augment tail classes [5, 25, 47, 49]. However, discriminative models map data from higher to lower dimensions, while generative models map images from lower to higher dimensions. Thus, the mechanism of class rebalancing between them may be completely different and how to design the balancing method remains under-explored.

**Generative models based on limited data** GAN [2, 6] is the most dominant model in the field of image synthesis in recent years. Given the requirement of large-scale data to train the generative models, a part of researchers focus on improving the performance of GAN models under the small datasets. To address the overfitting issue of the discriminator, a number of regularization methods [22, 37, 45] has been proposed. An alternative solution is Data Augmentation. As augmentation information is very prone to leak to the generator, researchers proposed improved augmentation strategies such as Differentiable Augmentation (DiffAug) [50] and Adaptive Augmentation (ADA) [21]. These augmentation ways can be transferred to diffusion models as baselines to help the generation of tail-class samples.

## 3. Method

In this section, we first give the basic notations following the classical DDPM model, and then introduce the class-imbalanced generation setting. In the third part, we introduce our CBDM algorithm and present its training details including implementation and hyper-parameter settings.

### 3.1. Preliminary

Diffusion models leverage a pre-defined forward process in training, where a clean image distribution $q(\boldsymbol{x}_0)$ can be corrupted to a noisy distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ at a specified timestep $t$. Given a pre-defined variance schedule $\{\beta_t\}_{1:T}$,

(a) Principle overview of CBDM

(b) DDPM(top)/CBDM(bottom) comparison when denoising a same noised image in class 70/86. We use both generators to recover a noisy image of two classes and observe that CBDM is able to produce more diversity based on the same starting point.
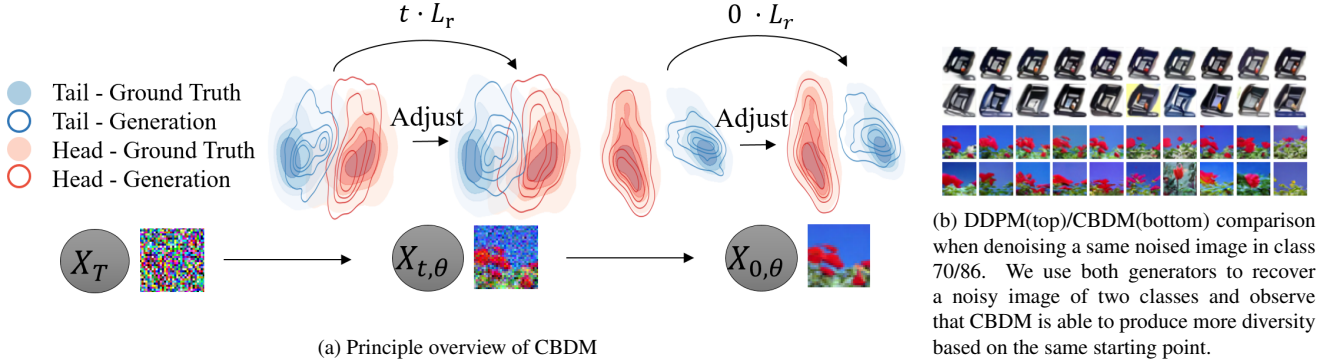
Figure 2. Algorithm (*left*) and generation (*right*) visualization. In the *left* figure, we show that an extra regularization loss $gL_r$ proportional to the diffusion step $t$ is added during training. This loss function pushes the sampling distribution (curves on the surface) to wider region while preventing it to be excessively distorted compared to the ground truth distribution (gradient color on the background).

the noisy distribution at any intermediate timestep is

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1 - \bar{\alpha}_t)\boldsymbol{I}); \quad \bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_i).$$

To reverse such forward process, a generative model $\theta$ learns to estimate the analytical true posterior in order to recover $\boldsymbol{x}_{t-1}$ from $\boldsymbol{x}_t$ as follows:

$$\min_{\theta} D_{\mathrm{KL}}[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_{\theta}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)]; \quad \forall t \in \{1, ..., T\},$$

and such an objective can be reduced to a simple denoising estimation loss [11]:

$$\mathcal{L}_{\mathrm{DDPM}} = \mathbb{E}_{t, \boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta}(\sqrt{\bar{\alpha}}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2\right]$$
(1)

For the case where label information is available, the model is trained to estimate the noise as above in both conditional cases $\boldsymbol{\epsilon}_{\theta}(\boldsymbol{x}_t, y, t)$ with data-label pairs $(\boldsymbol{x}_0, y)$ and unconditional case $\boldsymbol{\epsilon}_{\theta}(\boldsymbol{x}_t, t)$. In the sampling, the label-guided model estimates the noise with a linear interpolation $\hat{\boldsymbol{\epsilon}} = (1 + \omega)\boldsymbol{\epsilon}_{\theta}(\boldsymbol{x}_t, y, t) - \omega\boldsymbol{\epsilon}_{\theta}(\boldsymbol{x}_t, t)$ to recover $\boldsymbol{x}_{t-1}$, which is often referred as Classifier-Free Guidance (CFG) [12].

### 3.2. Class-Balancing Diffusion Models

Current diffusion models assume the data distribution to be uniform in class, and thus equally treat samples in the training stage. However, based on our observation, such training strategy leads to degradation in generation quality. Below, we provide an analysis that motivates our Class-Balancing Diffusion Models (CBDM).

Suppose $q(\boldsymbol{x}, y)$ is the data distribution that we need to match with the joint distribution $p_{\theta}(\boldsymbol{x}, y)$ predicted by a generative model. We analyze their difference from the density ratio $r = \frac{q(\boldsymbol{x}, y)}{p_{\theta}(\boldsymbol{x}, y)} = \frac{q(\boldsymbol{x} \mid y)}{p_{\theta}(\boldsymbol{x} \mid y)} \cdot \frac{q(y)}{p_{\theta}(y)}$. When the true label distribution $q(y)$ is the same as the prior $p_{\theta}(y)$,

which is usually assumed to be uniform, the density ratio $r$ is reduced the conditional term to learn a conditional model $p_{\theta}(\boldsymbol{x} \mid y)$. However, when such a hypothesis is violated, for head classes, $\frac{q(y)}{p_{\theta}(y)}$ would result in a larger weight that makes the model biased and hurt tail classes, and vice versa. Empirically, we observe that the generation degrades more on tail classes, as illustrated in Figure 1. Moreover, as shown in Figure 2a, compared to head classes, DDPM cannot well capture the tail-class data distribution and the mode is poorly covered during the sampling process. As a result, generations of tail classes often have poor quality and diversity, shown in Figure 2b.

To tackle this issue, the most intuitive approach lies in adjusting the prior label distribution through a class balanced re-sampling. However, such abrupt adjustment easily leads to negative improvement in experiments. The step-by-step sampling nature of diffusion models provides another aspect to adjust this distribution more softly. In this spirit, we propose to calibrate the learning process through the conditional transfer probability $p_{\theta}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)$ when there exists a gap between the class distribution and the prior.

Let $p_{\theta}^{\star}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)$ be the optimum trained in the case that $\frac{q(y)}{p_{\theta}(y)}$ is correctly estimated, and $p_{\theta}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)$ be the one trained in a class-imbalanced case. The relation between such two generative distributions can be described as the following proposition.

**Proposition 1.** *When training a diffusion model parameterized with $\theta$ on a class-imbalanced dataset, its conditional reverse distribution $p_{\theta}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)$ can be corrected with an adjustment schema:*

$$p_{\theta}^{\star}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y) = p_{\theta}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)\frac{p_{\theta}(\boldsymbol{x}_{t-1})}{p_{\theta}^{\star}(\boldsymbol{x}_{t-1})}\frac{q^{\star}(\boldsymbol{x}_t)}{q(\boldsymbol{x}_t)} \quad (2)$$

The proposition above shows that, when trained on a

class-imbalanced dataset, a diffusion model can still approach the true data distribution by applying a distribution adjustment schema $\frac{p_\theta(\boldsymbol{x}_{t-1})}{p_\theta^\star(\boldsymbol{x}_{t-1})}\frac{q^\star(\boldsymbol{x}_t)}{q(\boldsymbol{x}_t)}$ at every reverse step $t$. However, approximating this schema is not feasible at every sampling step, so CBDM incorporates it into the training loss function to achieve an equivalent objective, and thus gets rid of the model-free part $\frac{q^\star(\boldsymbol{x}_t)}{q(\boldsymbol{x}_t)}$. By further decomposing $p_\theta(\boldsymbol{x}_{t-1})$ and $p_\theta^\star(\boldsymbol{x}_{t-1})$ to the expectation of the conditional probability $p_\theta^\star(\boldsymbol{x}_{t-1}|\boldsymbol{x}_{t:T}, y)$, we present a upper bound to approximate this probability in Proposition 2.

**Proposition 2.** *For the adjusted loss $\mathcal{L}_{DM}^\star = \sum_{t=1}^T \mathcal{L}_{t-1}^\star$, an upper-bound of the target training objective to calibrate at timestep $t$ (i.e. $\mathcal{L}_{t-1}^\star$) can be derived as:*

$$\sum_{t\geq 1}\mathcal{L}_{t-1}^\star = \sum_{t\geq 1} D_{\mathrm{KL}}[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta^\star(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)]$$

$$\leq \sum_{t\geq 1}[\underbrace{D_{\mathrm{KL}}[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)]}_{\textit{Diffusion model loss } \mathcal{L}_{DM}}$$

$$+ \underbrace{t\mathbb{E}_{y'\sim q_{\mathcal{Y}}^\star}[D_{\mathrm{KL}}[p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y')]]]}_{\textit{Distribution adjustment loss } \mathcal{L}_r},$$

The upper bound in the above proposition can be considered as two parts. The first term $\mathcal{L}_{\mathrm{DM}}$ corresponds to an ordinary DDPM loss [11] *e.g.,* Eqn. (1), and the second loss $\mathcal{L}_{\mathrm{r}}$ is used to adjust the distribution as a regularization term. Roughly speaking, $\mathcal{L}_{\mathrm{r}}$ increases the similarity between the model's output and a random target class. Thus, it reduces the risk of overfitting on the head classes, and enlarges the generation diversity for tail class through knowledge obtained from other classes. When $q_{\mathcal{Y}}^\star$ is less longtailed than the dataset, this loss also increases the probability for underrepresented tail samples to be chosen during training.

### 3.3. Training algorithm

The detailed training algorithm of CBDM is presented in Alg. (1). In the algorithm, we reduce the distribution adjustment loss $\mathcal{L}_r$ as a square error loss with Monte-Carlo samples as indicated by Eqn.(3), where $\mathcal{Y}$ is a set of samples that drawn following the distribution $q_{\mathcal{Y}}^\star$ and $y$ denotes the image label. Note for CFG [12], there is a fixed probability (usually 10%) to drop the condition, *i.e.,* $y = $ None.

$$\mathcal{L}_{\mathrm{r}}(\boldsymbol{x}_t, y, t) = \frac{1}{|\mathcal{Y}|}\sum_{y'\in\mathcal{Y}}[t||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, y) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, y')||^2], \quad (3)$$

For the implementation in practice, CBDM can be plugged into any existing conditional diffusion models by adopting their model architecture and adjusting the training loss $\mathcal{L}_{\mathrm{DM}}$ following lines $6-11$. Specifically, the choice

---

**Algorithm 1** Training algorithm of CBDM.

1: **for** Every batch of size N **do**
2:     **for** $(\boldsymbol{x}_0^{(i)}, y^{(i)})$ in this batch **do**
3:         Sample $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $t \sim \mathcal{U}(\{0, 1, ..., T\})$
4:         $\boldsymbol{x}_t^{(i)} = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}^{(i)}$
5:         Calculate $\mathcal{L}_{\mathrm{DM}} = ||\boldsymbol{\epsilon}^{(i)} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^{(i)}, y^{(i)})||^2$
6:         Sample $y'^{(i)}$ from $q_{\mathcal{Y}}^\star$
7:         Calculate the first regularization term
8:         $\mathcal{L}_r = t\tau||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^{(i)}, y^{(i)}) - \mathrm{sg}(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^{(i)}, y'^{(i)}))||^2$
9:         Calculate the regularization commitment term
10:        $\mathcal{L}_{rc} = t\tau||\mathrm{sg}(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^{(i)}, y^{(i)})) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^{(i)}, y'^{(i)})||^2$
11:        Update with $\mathcal{L}_{\mathrm{CBDM}} = \mathcal{L}_{\mathrm{DM}} + \mathcal{L}_r + \gamma\mathcal{L}_{rc}$
12:     **end for**
13: **end for**

---

of the regularization weight $\tau$ affects the sharpness of the density ratio $\frac{p_\theta(\boldsymbol{x}_t)}{p_\theta^\star(\boldsymbol{x}_t)}$. For the theoretical analysis, please refer to our proof of Prop. (2) in the Appendix. In addition, the choice of the sampling set $\mathcal{Y}$ is another important perspective of CBDM, which depends on the target distribution we wish to adjust. Without loss of generality, we discuss two cases here. On the one hand, we can adjust the label distribution to a class-balanced label distribution, where we sample labels to construct $\mathcal{Y}^{bal}$. On the other hand, if the data distribution is heavily long-tailed, we can also target the adjusted distribution to a relatively less class-imbalanced distribution for stabilized training. In our experiments, we show CBDM can work well for both cases in different mechanisms.

Moreover, we observe that naively optimizing with this loss could make the model collapse to some trivial solutions, where the model outputs the same result regardless of the condition $y$ and thus degenerates the conditional generation performance. Therefore, we follow previous works to apply a stop gradient operation [4, 46] to prevent this issue. The final loss of CBDM is

$$\mathcal{L}_{\mathrm{CBDM}}^{\tau, \gamma, \mathcal{Y}}(\boldsymbol{x}_t, y, t, \boldsymbol{\epsilon}) = ||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, y) - \boldsymbol{\epsilon}||^2$$

$$+ \frac{\tau t}{|\mathcal{Y}|}\sum_{y'\in\mathcal{Y}}(||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, y) - \mathrm{sg}(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, y'))||^2$$

$$+ \gamma||\mathrm{sg}(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, y)) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, y')||^2), \quad (4)$$

where "sg$(\cdot)$" denotes the stop gradient operation; $\tau$, $\gamma$ are weights for the regularization and the commitment term respectively with $\gamma$ set to $\frac{1}{4}$ in a default setting, and $|\mathcal{Y}|$ denotes the number of elements in the label set.

(a) DDPM / Tail Class 94          (b) CBDM / Tail Class 94



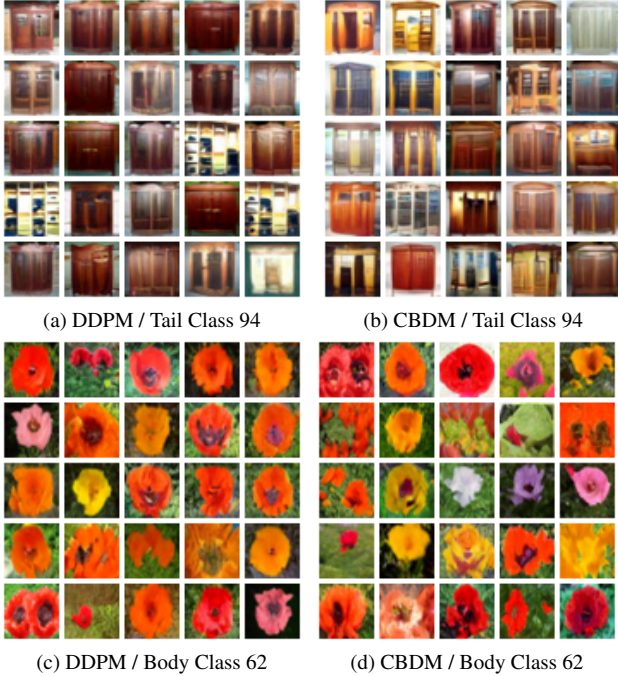(c) DDPM / Body Class 62          (d) CBDM / Body Class 62

Figure 3. Comparison of image generation on heavily tail-distributed (94) and mild tail-distributed (62) classes between DDPM and CBDM.

## 4. Experimental results

### 4.1. Experimental setup

**Datasets** We first chose two common datasets that are widely used in the domain of image synthesis, CIFAR10/CIFAR100, and their corresponding long-tailed versions CIFAR10LT and CIFAR100LT. The construction of CIFAR10LT and CIFAR100LT follows [3], where the size decreases exponentially with its category index according to the imbalance factor $imb = 0.01$. We also conduct experiments on 3 higher-resolution datasets, whose details and results could be found in the Appendix.

**Implementation details** We strictly follow the training configurations of the baseline models. For DDPM, we set diffusion schedule $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$ with $T = 1,000$, and optimize the network with an Adam optimizer whose learning rate is $0.0002$ after $5,000$ epochs of warmup. Considering that the size and semantic complexity of the datasets vary greatly, we choose appropriate epochs for each dataset. In the ablation studies, we also include EDM [20] and Diffusion-ViT [48] as our backbones to show the compatibility with different architectures and score-matching methods. Note that, when a model does not use conditional input, we follow [7] to slightly modify the backbone [8] by adding an extra embedding layer.

**Baselines** We first adopt following classic methods to build baselines: re-sampling (RS), soft re-sampling methods (RS-SQRT) [27], and augmentation-based methods including Differentiable Augmentation (DiffAug) [50] and Adaptive Augmentation (ADA) [21]. Besides diffusion-based baselines, previous state-of-the-art generative models that study the long-tailed distribution, *e.g.,* CBGAN [35], and group spectral regularization for GANs [34], are also included in the comparison. Precisely, RS follows the uniform class distribution and RS-SQRT [27] samples with a probability determined by the square root of class frequency. And DiffAug [50] is directly applied to all training images following their default threshold. Following [20], ADA [21] is not only applied to images, the augmentation pipeline is also encoded as a condition via an additional embedding layer in the U-Net.

**Metrics** CBDM and corresponding baselines are evaluated in terms of both generation diversity and fidelity via Frechet Inception Distance (FID) [10], Inception Score (IS) [40], Recall [23] and $F_\beta$ [39]. We measure Recall and $F_\beta$ using Inception-V3 features, and take respectively $K = 5$ for Recall, 1/8 and 8 for the threshold in $F_\beta$, and 20 times of class number as the clustering number of $F_\beta$ to capture the inner class variance. To this end, Recall and $F_8$ can be regarded as diversity metrics and IS and $F_{1/8}$ are more inclined to measure fidelity. In evaluation, we take the class-balanced version corresponding to those used in training, and all metrics are measured with 50k generated images (10k for ablation experiments). The classifier-free guidance is adapted in sampling and we tune the guidance strength $\omega$ for both baselines and our method to ensure the best performance. Specifically, we take $\omega = 1.6, 0.8, 1.0, 0.8$ respectively for the four CIFAR datasets mentioned above.

### 4.2. Main results

In this part, we consider DDPM as the most direct baseline. As shown in Table 1, CBDM overall outperforms DDPM on all datasets except the IS is slightly lower on CIFAR10. For datasets with more classes, the improvements are more significant in terms of both diversity and fidelity, as on CIFAR100LT and CIFAR100. We further investigate the performance of diffusion models when combined with augmentation-based and re-sampling methods. Surprisingly, except for ADA, we consistently observe a performance degradation when these methods are compared with the vanilla DDPM. Moreover, adding the ADA augmentation with CBDM training leads to a further improvement in terms of both diversity and fidelity.

As a qualitative justification, Figure 3 provides the visualization comparison between DDPM and CBDM on a relatively mild tail-distributed class (62) and a tail-distributed class (94). We remark that CBDM generates more diverse images. For example, on the tail class 94, the cabinets

| Dataset | Model | FID↓ | $F_8$ ↑ | Recall↑ | IS↑ | $F_{1/8}$ ↑ |
|---|---|---|---|---|---|---|
| CIFAR100LT | DDPM [11] | 7.38 | 0.85 | 0.52 | 13.11 | 0.88 |
| | +ADA [21] | 6.16 | 0.91 | 0.57 | 12.71 | 0.90 |
| | +DiffAug [50] | 9.19 | 0.88 | 0.47 | 11.56 | 0.86 |
| | + RS [27] | 10.50 | 0.65 | 0.49 | 12.60 | 0.83 |
| | +SQRT- RS [27] | 9.72 | 0.66 | 0.47 | **13.47** | 0.83 |
| | CBDM (ours) | 6.26 $_{(-1.12)}$ | 0.91 $_{(+0.06)}$ | 0.57 $_{(+0.05)}$ | 13.24 $_{(+0.13)}$ | 0.89 $_{(+0.01)}$ |
| | +ADA [21] | **5.81** $_{(-1.57)}$ | **0.91** $_{(+0.06)}$ | **0.57** $_{(+0.05)}$ | 13.34 $_{(+0.23)}$ | **0.90** $_{(+0.02)}$ |
| CIFAR100 | DDPM | 3.11 | 0.97 | 0.65 | 13.65 | 0.96 |
| | CBDM (ours) | **2.72** $_{(-0.39)}$ | 0.97 $_{(\pm 0)}$ | **0.67** $_{(+0.02)}$ | **14.03** $_{(+0.38)}$ | 0.96 $_{(\pm 0)}$ |
| CIFAR10LT | DDPM | 5.76 | 0.97 | 0.57 | 9.17 | 0.95 |
| | CBDM (ours) | **5.46** $_{(-0.30)}$ | 0.97 $_{(\pm 0)}$ | **0.59** $_{(+0.02)}$ | **9.28** $_{(+0.11)}$ | 0.95 $_{(\pm 0)}$ |
| CIFAR10 | DDPM | 3.16 | 0.99 | 0.64 | **9.80** | 0.98 |
| | CBDM (ours) | **3.03** $_{(-0.13)}$ | 0.99 $_{(\pm 0)}$ | **0.65** $_{(+0.01)}$ | 9.63 $_{(-0.17)}$ | 0.98 $_{(\pm 0)}$ |

Table 1. CBDM performance on different datasets. In the table, the first three columns are diversity-related, and we mark the best results in bold. As the comparison between CBDM and DDPM is more straightforward, we mark the performance gain in parentheses next to the CBDM results, using blue and red to indicate improvements and degradation respectively.

| PT | FT | $\mathcal{Y}$ | FID | $F_8$ | Recall | IS |
|---|---|---|---|---|---|---|
| CIFAR 100LT | - | $\mathcal{Y}^{train}$ | 6.26 | 0.91 | 0.57 | **13.24** |
| | | $\mathcal{Y}^{bal}$ | 6.10 | **0.94** | **0.64** | 12.29 |
| | | $\mathcal{Y}^{sqrt}$ | **6.00** | 0.93 | 0.59 | 13.00 |
| CIFAR 100LT | CIFAR 100LT | - | 7.38 | 0.85 | 0.52 | 13.11 |
| | | $\mathcal{Y}^{train}$ | 6.20 | 0.88 | 0.54 | **13.36** |
| | | $\mathcal{Y}^{bal}$ | **5.85** | **0.92** | **0.60** | 13.29 |
| CIFAR 100 | CIFAR 10LT | - | 5.39 | 0.66 | 0.58 | 9.43 |
| | | $\mathcal{Y}^{train}$ | **5.05** | **0.67** | 0.60 | 9.48 |
| | | $\mathcal{Y}^{bal}$ | 5.90 | 0.66 | 0.60 | **9.49** |

Table 2. CBDM performance for 3 mechanisms under different regularization sampling set $\mathcal{Y}$. Column *PT* indicates the dataset used in (pre)training, column *FT* indicates the dataset used for fine-tuning. The rows where $\mathcal{Y}$ marked as "-" represent the results with DDPM, and rows where *FT* marked as "-" represent the results of CBDM trained from scratch. Three diversity-preferred metrics and the best results are in bold for emphasis.



Figure 4. FID improvement per class compared to DDPM. The curve is smoothened by a moving average of 5 unit.

shown in Figure 3b have more color and texture, which justifies the improvements of diversity-related metrics in Table 1. On the contrary, DDPM only generates images that are highly similar to the training data.

**Case-by-case study.** To better understand the generation conditioned on each class, we compare the FID case-by-case between DDPM and CBDM on each class. The results are shown in Figure 4. Compared to DDPM, CBDM shows more consistent improvements for the class index greater than 40. In the tail classes, CBDM outperforms than DDPM the FID by a more significant margin.

**The choice of label set $\mathcal{Y}$** We investigate the effects of label distribution for CBDM under three mechanisms. The first mechanism is the vanilla CBDM given in Algorithm 1 which trains a diffusion model with regularization from scratch. Besides, we also study CBDM with two fine-tuning configurations. The first setting concerns using CBDM to adjust an existing model pre-trained on a long-tailed dataset to improve its performance. The second one refers to fine-tuning a pre-trained model to adapt to a smaller dataset with an imbalanced class distribution using CBDM. For those mechanisms, we consider three settings for sampling the label set: (1) a label distribution similar to the training set (denoted $\mathcal{Y}^{train}$); (2) a totally uniform label distribution (denoted $\mathcal{Y}^{bal}$); (3) a less long-tailed distribution compared to $\mathcal{Y}^{train}$, whose class frequency is the square root of the original one (denoted $\mathcal{Y}^{sqrt}$).

The results are shown in Table 2, where we found that CBDM performs well on all these configurations. As we

| Dataset | Model | FID | IS |
|---|---|---|---|
| CIFAR 100 | BigGAN+DA+$R_{LC}$ [45] (2021) | 2.99 | - |
| | DDPM | 3.11 | 13.65 |
| | CBDM (ours) | **2.72** | **14.03** |
| CIFAR 100LT | CBGAN [35] (2021) | 28.17 | - |
| | DDPM | 12.31 | 12.69 |
| | CBDM (ours)+ADA [21] | **10.50** | **12.81** |
| CIFAR 10LT | CBGAN [35] (2021) | 32.93 | - |
| | SNGAN [30]+gSR [34] (2022) | 18.58 | 7.80 |
| | DDPM | 9.68 | 9.00 |
| | CBDM (ours) | **9.38** | **9.12** |

Table 3. Comparision with long-tailed SoTAs on CIFAR. Following their setting, all methods are evaluated with 10k generated images and ground truth images are from their validation dataset. The publish year of every baselines are marked in next to the citation.

| Training data | Precison | Recall |
|---|---|---|
| CIFAR100 | 0.70 | 0.67 |
| CIFAR100LT | 0.45 | 0.39 |
| + DDPM gens (50k) | 0.48 (+0.03) | 0.44 (+0.05) |
| + CBDM gens (50k) | **0.49** (+0.04) | **0.47** (+0.08) |

Table 4. Recognition results of different training data. All configurations are evaluated on the testing set of normal CIFAR100. Gains based on CIFAR100LT dataset is noted in blue.

| Backbone | FID | $F_8$ | Recall | IS | $F_{1/8}$ |
|---|---|---|---|---|---|
| EDM [20] | 8.64 | 0.82 | 0.48 | 12.16 | 0.86 |
| +CBDM (ours) | **7.97** | **0.88** | **0.52** | **12.01** | **0.86** |
| Diffuse-ViT [48] | 37.6 | 0.76 | 0.45 | 7.66 | 0.62 |
| +CBDM (ours) | **30.0** | **0.81** | **0.51** | **8.00** | **0.65** |

Table 5. CBDM performance using different backbones on CIFAR100LT dataset. Three diversity-biased metrics.

| Dataset | FID↓ | $F_8 \uparrow$ | Recall↑ | IS↑ | $F_{1/8} \uparrow$ |
|---|---|---|---|---|---|
| CIFAR 100 | 3.44 (+0.72) | 0.97 (±0) | 0.69 (+0.02) | 13.11 (−0.92) | 0.96 (±0) |
| CIFAR 100LT | 6.27 (+0.01) | 0.93 (+0.02) | 0.61 (+0.04) | 12.56 (−0.68) | 0.89 (±0) |

Table 6. Sampling with 100 DDIM steps on CIFAR100 and CIFAR100LT. The difference compared with those using DDPM steps are reported below each result.

analyzed in section 3.3, when the empirical distribution of $\mathcal{Y}$ has a significant difference from the label distribution of the dataset, adjusting with $\mathcal{Y}^{bal}$ could hurt training stability and result in performance degradation. We observe that $\mathcal{Y}^{bal}$ only shows better performance regarding generation diversity, while the generation fidelity has an obvious gap than the other settings. Using a relatively mild set $\mathcal{Y}^{sqrt}$ has better performance in terms of IS and preserves the generation diversity, which leads to the best FID among all these settings. On the contrary, when a pre-trained model is available, fine-tuning on top of it ensures better stability, and we can observe in such case using $\mathcal{Y}^{bal}$ has significant improvement than the vanilla DDPM or fine-tuning using $\mathcal{Y}^{train}$. When adapting to a different dataset, we also observe using $\mathcal{Y}^{bal}$ causes stability issue and produce undesired results.

**Enhancement of training classifiers on long-tailed data**
As training models on long-tailed data usually leads to undesired classification results, we investigate whether the generated data could help improve the classifier trained on long-tailed data as a complementary evaluation metrics. Here, we train ResNet-32 models, respectively with CIFAR100, CIFAR100LT and CIFAR100LT augmented with DDPM and CBDM generations (50k samples). Table 4 shows that the training on long-tailed data results in se-

vere performance degradation. When augmenting the long-tailed data with the generated data, we observe fairly improvements in the trained classifier. Moreover, comparing the results between DDPM and CBDM, we observe that the improvements by using CBDM are more significant, which validates the effectiveness of our approach. Especially, the gain on recall is more remarkable than the gain on precision, which means that the generated images are more diverse than DDPM and justifies the results in Table 1.

**Comparison with other benchmarks** We include the representative state-of-the-art long-tailed generative modeling approaches in the comparison, most of which are based on GANs since they suffer more severe problems when the training data is skewed. For a fair comparison, we strictly follow their evaluation setting and the results are reported in Table 3. From the results, DDPM surpasses the performance of the baselines on all the datasets, and CBDM exhibits even stronger performance.

### 4.3. Ablations

**Compatibility with different backbones** We first investigate the performance of our method with different diffusion models. Without loss of generality, we adopt EDM [20] and Diffusion-ViT [48], which use a different score-matching method and different denoising network backbone. As shown in Table 5, compared with the corresponding baselines, CBDM improves the generation result on CIFAR100LT dataset for both backbones, demonstrating the compatibility with a variety of backbones and the effectiveness of handling long-tailed data.

**CBDM with DDIM sampling** Apart from DDPM reverse sampling and neural SDE methods, recent deterministic ODE methods such as DDIM [43] are widely used in the
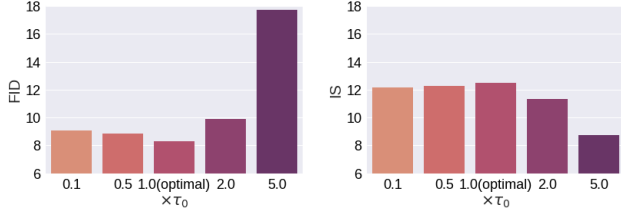
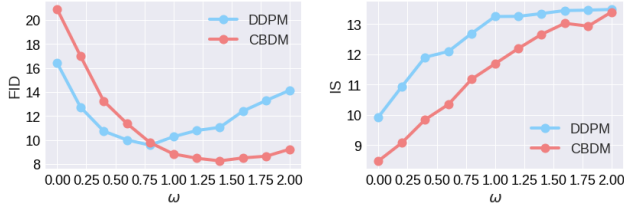Figure 5. FID/IS score under different regularization weight $\tau$



Figure 6. FID/IS score under different guidance strength $\omega$



Figure 7. Image fidelity and diversity controlled by guidance strength $\omega$ and regularization weight $\tau$.

generation process of diffusion models. We also conduct experiments to study whether CBDM is compatible with DDIM steps. Table 6 shows the comparison of using DDPM and DDIM reverse steps when the generation steps are compressed to $1/10$. We note on normal CIFAR100, FID and IS become slightly worse than using 1000 DDPM steps. On CIFAR100LT, except for IS, all metrics remain almost the same as using DDPM steps.

**Effects of hyperparameters**  We tested the effects of regularization weight $\tau$ in CBDM, which is set in default $\tau = \tau_0 = 0.001$ so that the weight does not surpass 1 at all steps. As shown in Figure 5, we search weights across different scales and found that the optimal weight is $\tau = \tau_0$. The results also point out that $\tau$ should not be too large or too small. We also tested the difference between CBDM models trained using commitment loss $\mathcal{L}_{rc}$ and those trained without it. Our experiments show that the FID score of the models using the commitment loss (8.30) are significantly lower than those of the models not using it (8.84).

**Guidance strength $\omega$**  As we adapt the Classifier-Free Guidance (CFG) during sampling, we also show the effects of guidance strength $\omega$. In Figure 6, we searched $\omega$ from 0 to 2 with an interval of 0.2 and compare FID and IS score of DDPM and CBDM models. We observe that the FID of CBDM remains decreasing at larger guidance strengths compared to DDPM; and although the IS of CBDM is consistently weaker than that of DDPM, they reach the same level when the guidance strength reaches 2.0, where the FID of CBDM (6.75) is still significantly lower than the best FID of DDPM (7.38).

**Fidelity-diversity control**  [12] describes the phenomenon where excessive guidance strength $\omega$ tends to lead to better fidelity at the price of overfitting and diversity degradation, as a fidelity-diversity tradeoff. CBDM
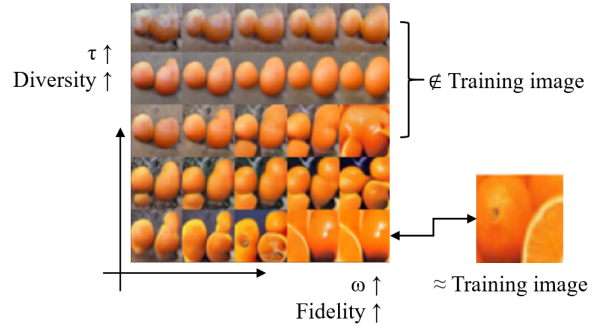
additionally provides another hyperparameter: the regularization weight $\tau$ that promotes model diversity through cross-class information interaction. Figure 7 shows the body class (53) generation results under different guidance strengths and different regularization weight. We observe that a DDPM model with high guidance strength is able to generate very realistic results, but its images overlap almost exactly with a sample in the training set. In contrast, the guidance term of CBDM does not drastically change the image content, but rather refines the image content to be closer to the selected class than the unguided result. This means that CBDM retains more of the diversity of the unguided case and effectively enhances the category information with guide terms.

## 5. Conclusion

In this paper, we focus on the problem when the training data is imbalanced and the diffusion model drops in generation quality on tail classes. We first establish a low baseline for this task by examining the most common paradigms in long-tail scenarios and in training generative models with limited data. Thereafter, we propose the Class-Balancing Diffusion Model (CBDM) through theoretical analysis. This approach can be implemented very cleanly in the training of any conditional diffusion model and thus has the potential to be widely used in other fields. Our experiments show that the CBDM approach significantly improves model generation diversity with high fidelity, on both class-balanced and class-imbalanced datasets.

## 6. Acknowledgement

# References

[1] Jan Kautz Arash Vahdat, Karsten Kreis. Score-based generative modeling in latent space. In *NeurIPS*, 2021. 1

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 2

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32, 2019. 5

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 4

[5] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 694–710, Cham, 2020. Springer International Publishing. 2

[6] Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced Wasserstein distance. In *CVPR*, pages 3483–3491, 2018. 2

[7] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, 2021. 1, 2, 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[9] Didrik Nielsen Giorgio Giannone and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2021. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4, 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 4, 8

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1

[14] Greg Holste, Song Wang, Ziyu Jiang, Thomas C. Shen, George L. Shih, Ronald M. Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. *Data augmentation, labelling, and imperfections : second MICCAI workshop, DALI 2022, held in conjunction with MICCAI 2022, Singapore, September 22, 2022, proceedings. DALI (Workshop)*, 13567:22–32, 2022. 1

[15] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 1

[16] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, pages 5375–5384, 2016. 2

[17] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-TTS: A Denoising Diffusion Model for Text-to-Speech. In *Proc. Interspeech 2021*, pages 3605–3609, 2021. 1

[18] Alexia Jolicoeur-Martineau, Remi Piche-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*. 1

[19] Haeyong Kang, Thang Vu, and Chang D. Yoo. Learning imbalanced datasets with maximum margin loss. In *ICIP*, pages 1269–1273, 2021. 1

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 5, 7

[21] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, Red Hook, NY, USA, 2020. Curran Associates Inc. 1, 2, 5, 6, 7

[22] Minsu Ko, Eunju Cha, Sungjoo Suh, Huijin Lee, Jae-Joon Han, Jinwoo Shin, and Bohyung Han. Self-supervised dense consistency regularization for image-to-image translation. In *CVPR*, pages 18280–18289, 2022. 2

[23] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 32, 2019. 5

[24] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 1

[25] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, pages 2967–2976, 2020. 2

[26] Guanxiong Luo, Martin Heide, and Martin Uecker. MRI reconstruction via data driven markov chain with joint uncertainty estimation. *CoRR*, abs/2202.01479, 2022. 1

[27] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, pages 185–201, Cham, 2018. Springer International Publishing. 1, 2, 5, 6

[28] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1

[29] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*. OpenReview.net, 2021. 1, 2

[30] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 7

[31] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022. 1, 2

[32] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *CVPR*, pages 864–873, 2016. 2

[33] Yuntian Gu Quanlin Wu, Hang Ye. Guided diffusion model for adversarial purification from random noise. *arXiv preprint arXiv:2206.10875*, 2022. 1, 2

[34] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R. Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, pages 426–442, Cham, 2022. Springer Nature Switzerland. 1, 5, 7

[35] Harsh Rangwani, Konda Reddy Mopuri, and R. Venkatesh Babu. Class balancing gan with a classifier in the loop. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1618–1627. PMLR, 27–30 Jul 2021. 5, 7

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[37] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NeurIPS*, page 2015–2025, Red Hook, NY, USA, 2017. Curran Associates Inc. 1, 2

[38] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 1

[39] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *NeurIPS*, page 5234–5243, Red Hook, NY, USA, 2018. Curran Associates Inc. 5

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *NeurIPS*, volume 29. Curran Associates, Inc., 2016. 5

[41] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *CVPR*, pages 11492–11501, 2022. 2

[42] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015. 2

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2021. 2, 7

[44] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2022. 1

[45] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *CVPR*, pages 7917–7927, 2021. 2, 7

[46] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, NIPS'17, page 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. 4

[47] Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *CVPR*, pages 3783–3792, 2021. 2

[48] Xiulong Yang, Sheng-Min Shih, Yinlin Fu, Xiaoting Zhao, and Shihao Ji. Your vit is secretly a hybrid discriminative-generative diffusion model. *ArXiv*, abs/2208.07791, 2022. 5, 7

[49] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, pages 5697–5706, 2019. 2

[50] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, Red Hook, NY, USA, 2020. Curran Associates Inc. 1, 2, 5, 6