# Taming the Tail in Class-Conditional GANs: Knowledge Sharing via Unconditional Training at Lower Resolutions

Saeed Khorram[1], Mingqi Jiang[1], Mohamad Shahbazi [2], Mohamad H. Danesh[3], Li Fuxin [1]
[1]Oregon State University [2] ETH Zürich [3] McGill University
{khorrams, jiangmi, lif}@oregonstate.edu
mshahbazi@vision.ee.ethz.ch, mohamad.danesh@mail.mcgill.ca

## Abstract

*Despite extensive research on training generative adversarial networks (GANs) with limited training data, learning to generate images from long-tailed training distributions remains fairly unexplored. In the presence of imbalanced multi-class training data, GANs tend to favor classes with more samples, leading to the generation of low quality and less diverse samples in tail classes. In this study, we aim to improve the training of class-conditional GANs with long-tailed data. We propose a straightforward yet effective method for knowledge sharing, allowing tail classes to borrow from the rich information from classes with more abundant training data. More concretely, we propose modifications to existing class-conditional GAN architectures to ensure that the lower-resolution layers of the generator are trained entirely unconditionally while reserving class-conditional generation for the higher-resolution layers. Experiments on several long-tail benchmarks and GAN architectures demonstrate a significant improvement over existing methods in both the diversity and fidelity of the generated images. The code is available at https://github.com/khorrams/utlo.*

## 1. Introduction

In the past few years, research on Generative Adversarial Networks (GANs) [7] has led to remarkable advances in generating realistic images [2, 17, 41]. Conditional GANs [29] (cGANs) have garnered particular attention due to their ability to accept user inputs which can additionally guide the generation process. They enable a wide range of applications such as class-conditional image generation [42], image manipulation [43], image-to-image translation [12], super-resolution [22] and text-to-image synthesis [52].

Despite the advances, past research on cGANs has focused primarily on learning from balanced data. In real-world scenarios, however, data often follows a power-law distribution, with a few classes dominating most of the training data. This is referred to as the "*long-tail (LT) problem*" [55], where most of the training data come from a few classes (referred to as the "*head*" of the distribution). In contrast, data from a large number of classes rarely occur (referred to as the "*tail*" of the distribution). This imbalance in data poses a challenge to the effective learning of tail classes, as standard machine learning algorithms tend

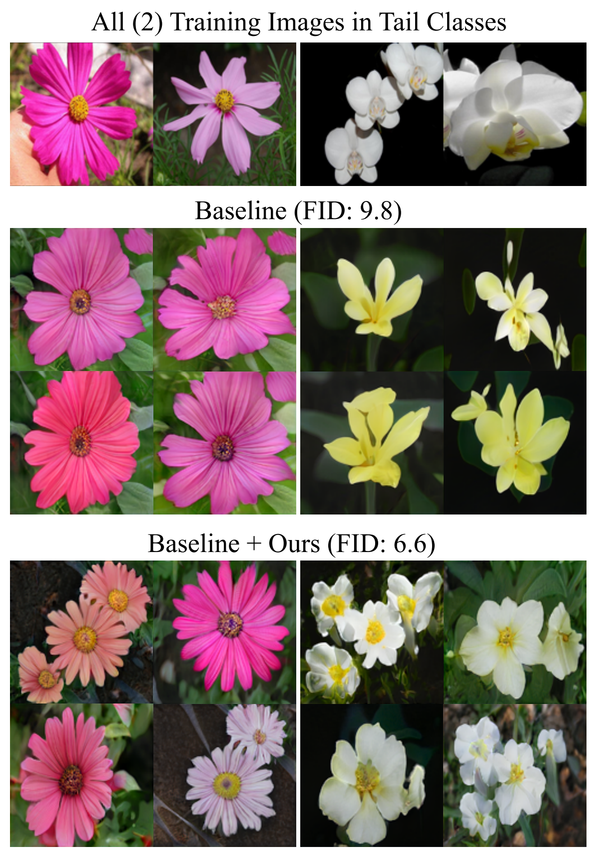


Figure 1. Generating images from rare tail classes in the Flowers-LT with only **two** training images. Our proposed approach allows for a more diverse set of features such as backgrounds, colors, poses, and object layouts to be infused into the tail classes.

to favor densely sampled regions of the input distribution.

Although there has been a surge of interest in the long-tail problem in recent years, the focus has been primarily on recognition tasks [49]. Generative learning on the long-tail data, particularly cGANs which is the focus of this work, remains underexplored. Recently, [36, 37] identified mode collapse on tail classes as a result of the spectral norm explosion in class-conditional BatchNorms and the collapse of the latents in the StyleGANs $\mathcal{W}$ space. While techniques such as regularization, normalization, and/or class-balancing techniques [49] might improve the performance of long-tail generative models to some extent, their effectiveness is limited. The primary challenge in the long-tail problem is the under-representation of tail classes in the training data, e.g., in Fig. 1 each tail class contains merely 2 training images, resulting in insufficient observation of the tail classes by the learning algorithm.

To address this challenge, we propose to leverage the abundant information in the head classes, with the aim to learn and infuse knowledge from the head classes into the tail ones, thereby enriching the training distribution for the latter. However, it is very difficult to explicitly disentangle the class-specific information pertaining to head classes from class-independent information shared between head and tail classes. We make a key observation that the similarity between head and tail instances is often *higher at lower resolutions*: discriminative features such as shape or specific texture are usually unveiled at higher resolutions. Conversely, information from the lower resolutions tends to be more class-independent such as background, configuration, or direction of objects (Fig. 6), thus can be shared between the head and tail classes.

Building upon this observation, we propose Unconditional Training at LOwer resolution (UTLO), a novel approach for training cGANs in the long-tail setup. In addition to the standard conditional GAN objective, UTLO trains the intermediate low-resolution output of the generator with an *unconditional* GAN objective. The unconditional GAN objective encourages the learning of low-resolution features common to both the head and tail classes, which are infused into the subsequent layers of the generator, particularly benefiting the under-represented tail classes.

Through extensive experiments and analysis on several long-tailed datasets, we demonstrate that using UTLO to combine high-resolution conditional and low-resolution unconditional training effectively facilitates knowledge sharing between head and tail classes, thereby improving the overall generative modeling of cGANs in the long-tail setup. Due to the strong class imbalance in the long-tailed data, naive usage of existing GAN metrics can be misleading. To mitigate this issue, we propose a few practices to adapt commonly-used GAN metrics to the long-tail setup.

Below, we highlight our main contributions:

- We propose UTLO, a novel knowledge-sharing framework tailored for training cGANs in the long-tailed setup. UTLO allows infusing information from the head classes to the tail classes via an additional unconditional objective applied to the low-resolution part of the cGAN generator. To the best of our knowledge, this work is the first to demonstrate that not all layers in a cGAN need to receive class-conditional information, i.e., a cGAN featuring a partially unconditional generator.
- We present a set of practices and metrics designed to adapt the commonly-used GAN evaluation metrics for long-tail setups, enabling a more precise evaluation of the image generation quality.
- Through extensive experiments across multiple benchmarks and architectures, we validate the effectiveness of our proposed method in improving the training of cGANs in the long-tail setup, achieving state-of-the-art results across several long-tail datasets.

## 2. Related Work

**Conditional Generative Adversarial Networks** The initial work on conditional image generation using GANs [7] was presented by [29], in which the class conditions were concatenated to the inputs of the generator and discriminator networks. The AC-GAN [33] proposed the use of auxiliary classification in the discriminator. [2] introduced BigGAN and set a milestone in large-scale and high-resolution conditional image generation. The authors of StyleGAN [17, 18] first map the input noise and class condition to a latent style space, which is then passed to the multiple layers of the generator for image synthesis. More recently, an extension of the StyleGAN called StyleGAN-XL [41] has become the state-of-the-art in conditional image generation on several datasets and outperforms more complex and time-consuming approaches such as diffusion models [6].

**GAN Regularization under Limited Data** Training GANs under limited data is challenging as the discriminator can memorize the training samples, resulting in the collapse of the training or quality degradation of the generated images. Recently, data augmentation and regularization techniques have been incorporated into GAN training as a means to mitigate this problem [15, 17, 46, 53, 57, 58]. Additionally, [47] introduced a regularization term to the GAN objective, which tracks the predictions of the discriminator for real and generated images using separate moving averages. [21] employed off-the-shelf vision models in an ensemble of discriminators, demonstrating improved performance in both limited-data and large-scale GAN training.

In the context of few-shot image generation, FastGAN [23] proposed a lightweight architecture with a self-supervised discriminator. [34] introduced cross-domain distance consistency in order to transfer diversity from a source domain to a target domain. Subsequently, [19] proposed a latent-mixup strategy to smooth the latent space through controlled latent interpolation.

Recently, [42, 57] have observed that class-conditioning can cause mode collapse in the limited data regime. Their proposed work learns from limited but balanced/unlabeled data, which is different from our setup: heavily imbalanced long-tail data. [42] proposed Transitional-cGAN, a training strategy that starts with unconditional GAN objective and injects class-conditional information during a transition period. This approach was found to be effective in mitigating mode-collapse in the limited data regime. Our main idea of separating pathways between low-resolution and high-resolution is novel w.r.t. [42]: to promote knowledge sharing, we modify the cGAN architecture so that the lower-resolution part of the generator is entirely unconditional while only the higher-resolution part is conditional and received class-conditional information.

**Long-tail Recognition** To address the long-tail recognition problem [49, 55], previous research has primarily focused on techniques such as class re-balancing [3, 28, 38], learning algorithm and model design [13, 14, 59], and information augmentation [24, 45, 48]. More related to our work, re-sampling [8, 56] has been widely used to handle class imbalance problems. This involves balancing the number of samples used during training through over-sampling and/or under-sampling. Oversampling increases the frequency of tail classes [9, 31] while under-sampling reduces the imbalance by reducing the frequency of head class instances [25, 50]. Note that our research diverges from prior studies on long-tail recognition by addressing the more challenging task of generative modeling of the long-tail data, as opposed to the conventional long-tail classification task.

**Training GANs on Long-tail Data** Recent years have seen a growing interest in learning from long-tail data. However, most focus has been directed toward recognition tasks, with a limited number of studies addressing the development of generative models for long-tail data. GAMO [31] uses adversarial training to over-sample from minority classes. [35] introduced class-balancing regularization to the GAN objective, which leverages the predictions of a pretrained classifier to improve the generation of underrepresented classes in an unconditional setting. However, this method trains an unconditional GAN and is restricted to having access to a pre-trained classifier on the long-tail data to guide the training. The most closely related work to ours is the Group Spectral Regularization (GSR) [36] and NoisyTwins [37] regularizations. The authors identified mode collapse on tail classes as a result of the spectral norm explosion in class-conditional BatchNorms and the collapse of the latents in the StyleGANs $\mathcal{W}$ space, respectively. In contrast, we present a novel approach for sharing knowledge between head and tail classes to enhance long-tail learning, which is orthogonal to regularization techniques. In addition, while [37] is restricted to StyleGANs, our framework can be extended to different GAN architectures.

# 3. Taming the Tail in cGANs

## 3.1. Background: Conditional Generative Adversarial Networks

Conditional Generative Adversarial Networks (cGANs) are suitable for many applications due to providing user control over the generated samples at inference time. This is particularly helpful in the long-tail setup, where it is desirable to explicitly generate samples from rare (tail) classes. In contrast, unconditional GANs are likely to generate samples that track the training distribution, resulting in most samples coming from common (head) classes. To train a cGAN over a dataset containing $n$ instances $\mathbb{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and their corresponding labels $\mathbb{Y} = \{y_1, \ldots, y_n\}$ , we can formulate the adversarial training objective that alternatively optimizes the following loss terms,

$$\begin{aligned}\mathcal{L}_c^D &= \mathbb{E}_{\boldsymbol{x},y}[f_D\left(-D(\boldsymbol{x}|y)\right)] + \mathbb{E}_{\boldsymbol{z},y}[f_D\left(D\left(G\left(\boldsymbol{z},y\right)\right)\right)] \\ \mathcal{L}_c^G &= \mathbb{E}_{\boldsymbol{z},y}[f_G\left(-D\left(G\left(\boldsymbol{z},y\right)\right)\right)]\end{aligned} \quad (1)$$

where $G$ and $D$ are the generator and discriminator networks with their corresponding loss functions $f_G$ and $f_D$ [27]; $\boldsymbol{x}|y \sim p_{\text{data}}(\boldsymbol{x}|y)$ is real data drawn from class $y$. The generator $G$ conditions on both $\boldsymbol{z} \sim p_z(\boldsymbol{z})$, a random noise vector from a prior distribution over latent space, and $y$, the class conditioning vector. This adversarial training objective encourages the cGAN to reach an equilibrium between the generator and the discriminator, resulting in the generator producing realistic samples indistinguishable from real data by the discriminator.

## 3.2. Caveat of Training cGANs on Long-tail Data

Training GANs on long-tail data is challenging due to the inherent difficulty in modeling the rare examples that comprise the tail of the data distribution. This skewness in the data distribution hinders balanced learning across classes. In particular, the discriminator does not see enough examples from the tail of the distribution during training, which can lead to poor discriminative signals for the generator. As a result, the generator learns to generate a small subset of the possible outputs that fools the inadequately learned discriminator – exhibiting a classic "mode collapse" scenario.

Although regularizations introduced in [36, 37] can help mitigate the collapse of learning in tail classes that contain a sufficiently large number of samples, we still observe mode collapse occurring when a very limited number of samples are in tail classes, which is inherently challenging. This can be seen in Fig. 2 for tail classes of CIFAR100-LT with as few as 5 training instances. Although early-stopping can be adopted to obtain reasonable performance, the key in boost-

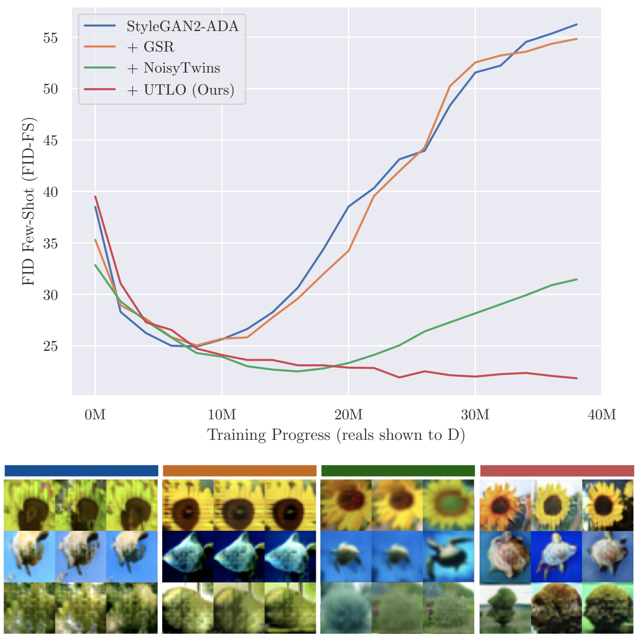


Figure 2. Convergence of different methods on CIFAR100-LT ($\rho = 100$), where tail classes have as few as 5 training examples. Incorporating our framework into the baseline alleviates overfitting as a result of knowledge sharing from head to rare tail classes.

ing the performance would be an approach to infuse additional diversities into the tail classes rather than only relying on techniques such as regularization.

### 3.3. Knowledge Sharing via Unconditional Training at Lower Resolutions (UTLO)

We devise a novel generative adversarial network (GAN) training scheme for long-tail data that aims to share knowledge from predominant *head* classes to rare *tail* classes, assuming some of the tail and head classes are at least *coarsely* similar. We propose a framework that uses unconditional learning in low-resolution images / features at an intermediate layer of the generator using the **unconditional** GAN objective. This helps with learning universal features from both head and tail classes. The subsequent layers, responsible for introducing finer details at higher resolutions, are trained using a **class-conditional** GAN objective.

In our proposed framework, we perform conditional training on the high-resolution image output from the final layer of the generator network. This is built upon the features learned at lower layers through the use of an unconditional objective, inheriting features primarily from head classes. The conditioning on the class labels gives control at the inference time to explicitly generate images from tail classes, meeting our design desideratum. The combination of unconditional and conditional objectives using our proposed method enables knowledge sharing, which, in turn, improves the quality of GAN training on long-tail data.

Our framework is general and applicable to many GAN architectures. In this section, we illustrate the application of our framework on StyleGAN2 with adaptive data augmentation (ADA) [17], a state-of-the-art and solid baseline for training GANs, particularly in the limited data regime. In the following, we demonstrate the necessary modifications to the architecture of the generator and discriminator networks. Although we show the necessary modifications for StyleGAN2-ADA, our method can be easily extended to other GAN architectures such as FastGAN [23], which is also used in our experiments (see Section 4).

#### 3.3.1 Modifying the Generator

In the generator design of the StyleGAN2-ADA [17, 18], the class-conditional information, combined with the latent vector, is first embedded in the style space $w$ using the style-mapping network $G_{\text{map}}$. The style vectors are then broadcasted across the layers of the synthesis network $G_{\text{syns}}$ in order to generate diverse images. To meet our design desideratum, we dissect the synthesis network into two sub-networks $G_{\text{syns}} = G_h \circ G_l$ where $G_l$ represents the earlier part of the network that produces intermediate features and/or images at a low resolution $L \times L$. $G_h$ on the other hand, represents the latter part of $G_{\text{syns}}$ that generates the output image at the high resolution $H \times H$ where $H > L$.

To block the flow of class-conditional information to the lower layers of $G_{\text{syns}}$, we generate separate $w$ vectors, one containing the class-conditional embeddings, referred to as $w_{z,y}$, while for the other one, referred to as $w_z$, class information was set to zero. Note that both vectors share the same latent variable $z$ and mapping network $G_{\text{map}}$ parameters, and the only difference is the presence of class-conditional information. The lower layers of the synthesis network $G_l$ are conditioned on $w_z$ while the subsequent layers at higher resolutions receive $w_{z,y}$ (see Fig. 3).

Generators typically follow a network design that gradually increases the resolution of intermediate features and/or images as the network progresses. This allows us to select a desired low-resolution, such as $8 \times 8$ or $16 \times 16$, during training. Recent generator designs often incorporate skip connections [16] or residual connections [30] to improve gradient flow. The generator of StyleGAN2-ADA uses skip connections, which explicitly generates intermediate RGB images at each layer. We take this low-resolution image as the input to the discriminator. Note that in the case of a generator design that uses residual connections, a $1 \times 1$ convolutional layer can be incorporated to convert the residual channels to RGB channels.

In a forward pass through the generator, images at low and high resolutions can be generated simultaneously without additional overhead. By inputting a latent code $z \sim p_z(\boldsymbol{z})$ and a target label $y$, we can obtain an unconditional synthetic image $\hat{\boldsymbol{x}}_l \in \mathbb{R}^{3 \times L \times L}$ along with a conditional

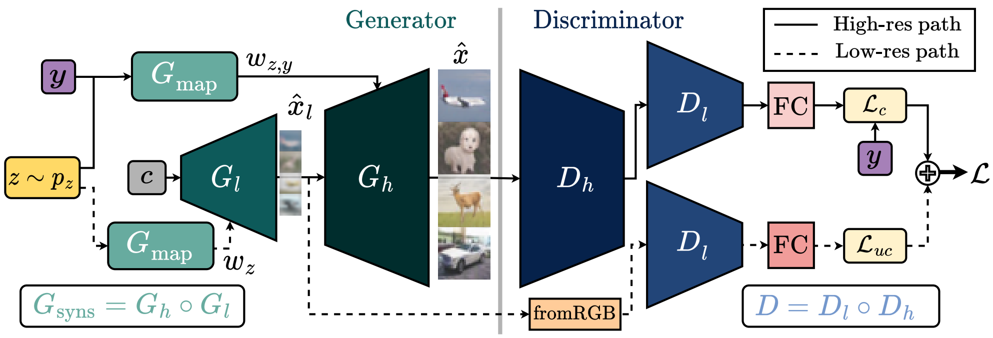


Figure 3. The proposed framework, UTLO, illustrated for the StyleGAN2-ADA architecture. Low and high resolution image pathways are used for unconditional and class-conditional objectives, respectively. $z$ is the input latent code, $y$ indicates the class embeddings, and $c$ is a constant input. Separate style vectors, $w_z$ (class-independent) and $w_{z,y}$ (class-conditional), are generated using the same z and a shared style-mapping network $G_{\text{map}}$ which are then passed to $G_l$ and $G_h$, respectively. The high-resolution generated image $\hat{x} \in \mathbb{R}^{3\times H\times H}$ is passed through the discriminator $D = D_l \circ D_h$ to calculate the conditional objective $\mathcal{L}_c$ while the low-resolution image $\hat{x}_l \in \mathbb{R}^{3\times L\times L}$ is passed only through $D_l$ to calculate the unconditional objective $\mathcal{L}_{uc}$. The final objective $\mathcal{L}$ is the combination of the two. While $D_l$ is shared, two *separate* prediction heads (FC layers) are used for unconditional and conditional objectives. The fromRGB is designed to increase the dimensionality of RGB channels to match the input channels of the $D_l$.

high-resolution synthetic image $\hat{\boldsymbol{x}} \in \mathbb{R}^{3\times H\times H}$ at the output. While $\hat{\boldsymbol{x}}$ is trained to be from class $y$, $\hat{\boldsymbol{x}}_l$ has no class-specific constraints. At inference time, the intermediate images can be discarded.

#### 3.3.2 Modifying the Discriminator

Opposite to the generator design, discriminators gradually reduce the resolution of the intermediate features as the network progresses. In order to pass images to an intermediate layer of the discriminator, we dissect it into two sub-networks. Formally, we define the discriminator as $D = D_l \circ D_h$ where $D_h$ represents the earlier part of the original network that takes the images at training resolution $\mathbb{R}^{3\times H\times H}$. On the other hand, $D_l$ represents the latter part of the discriminator taking a lower resolution input $\mathbb{R}^{C\times L\times L}$ with $C$ channels.

The default discriminator design in StyleGAN2-ADA uses residual connections. This does not allow direct passing of RGB inputs to the intermediate layers. To overcome this, we add a $1\times1$ convolutional layer, referred to as "fromRGB", that increases the channels of the RGB images to match $C$, the number of input channels of $D_l$.

More concretely, $D_l$ is a shared network in both the unconditional low-resolution and class-conditional high-resolution image pathways (see Fig. 3). To obtain separate unconditional and conditional predictions, we use two different fully-connected (FC) layers at the final layer of $D_l$. The output of the $D_h$ and the low-res image $\hat{\boldsymbol{x}}_l$ are passed through the $D_l$ which are then used for calculating the class-conditional loss $\mathcal{L}_c$ and unconditional loss $\mathcal{L}_{uc}$, respectively.

Note, when passing real images to $D_l$ for the unconditional loss, they are first downsized to the low-resolution $L \times L$ using bi-linear interpolation to be comparable with the ones generated from $G_l$. They are then passed through the fromRGB layer and consequently $D_l$. For the class-conditional loss, the original images are passed to $D$ without any modification.

#### 3.3.3 Final Training Objective

The final objective of our framework is simply a combination of the conditional and unconditional GAN losses, using a weighting parameter $\lambda$. To put it formally we have,

$$\mathcal{L}^D = \mathcal{L}_c^D + \lambda \cdot \mathcal{L}_{uc}^D, \ \mathcal{L}^G = \mathcal{L}_c^G + \lambda \cdot \mathcal{L}_{uc}^G \tag{2}$$

where $\mathcal{L}_c^D$ and $\mathcal{L}_c^G$ represent the class-conditional losses for the discriminator and generator, respectively (as outlined in Eq. 3.1), and $\mathcal{L}_{uc}^D$ and $\mathcal{L}_{uc}^G$ represent their unconditional counterparts, where $\boldsymbol{x}$ is sampled from $p_{\text{data}}(\boldsymbol{x})$ instead of $p_{\text{data}}(\boldsymbol{x}|y)$. The overall framework of our proposed method is depicted in Fig. 3.

## 4. Experiments

### 4.1. Setup

**Datasets.** We use 6 different long-tailed datasets in our experiments: CIFAR10 and CIFAR100[20], LSUN [51], Flowers [32], iNaturalist2019 [11], and AnimalFaces [44]. This selection is intended to encompass a wide spectrum of image domains, dataset sizes, resolutions, and imbalance ratios (denoted as $\rho$). $\rho$ represents the ratio of the number of training samples between the classes with the most and the least number of examples. To ensure the tail classes stay *few-shot*, we keep the number of images in the smallest tail classes under 50 [49]. A detailed description of these datasets is provided in Supp. A.

**Baselines.** In order to demonstrate the effectiveness of our proposed method, we use various cGAN architectures covering different generator and discriminator designs and data augmentation pipelines:

- *StyleGAN2 with Adaptive Data Augmentation (ADA)* [17]. We also integrate *Transitional* training [42], *Group Spectral Regularization (GSR)* [36], and *NoisyTwins* [37] to this baseline.
- *Projected GAN (PGAN)* [40]: we use projected discriminator with *Differentiable Augmentation (DA)* [57]. We also add GSR [36] to this baseline.

For comprehensive details on the baselines, hyperparameters, and implementation please refer to Supp. F.

### 4.2. Evaluation Metrics For Long-Tail Datasets

Evaluating the images generated by GANs on a long-tailed setup poses challenges, primarily due to the imbalanced data and access to a limited number of samples for the tail classes. For our experiments, we employ widely-used GAN metrics: Fréchet Inception Distance (FID) [10], and Kernel Inception Distance (KID) [1]. Following [10, 17], we report all metrics by computing statistics between $50k$ generated images and all available training images. In the following, we present a set of practices for adapting the aforementioned metrics to the long-tail setup.

First, when a larger and (more) balanced dataset following the long-tail training data distribution is available, it is used for metric calculation, e.g., full CIFAR10 before artificial imbalance. For naturally imbalanced datasets such as iNaturalist2019, we use the unmodified training set to calculate metrics. Note that when generating images for evaluation, we sample from the same distribution as the available real dataset, which may be imbalanced. While FID and KID indicate how well the real data distribution matches the generated images across all classes, we suggest additional metrics to mitigate the disproportionate influence of head classes in data statistic calculations.

In response to the varied number of training instances across classes in a long-tail setup, we propose evaluating metrics specifically on few-shot (FS) categories [49]. We calculate the FID and KID for the few-shot subset and refer to them as *"FID-FS"* and *"KID-FS"*, respectively. This is tailored to evaluate the quality of generated samples in the tail classes. Contrasting with the standard FID/KID, we maintain an equal number of real images across all classes during our FID-FS/KID-FS calculation. This emphasizes learning quality on tail classes, irrespective of any imbalances that might be present in the few-shot subset. Reporting both standard and few-shot metrics provides a more comprehensive evaluation of long-tail learning, considering the performance of both head and tail classes.

### 4.3. Results

Below, we report the results obtained across different benchmarks and cGAN architectures. We pick the *best* model for reporting results as the one with the lowest FID-FS from two independent runs. Note that our method exhibits significantly less reliance on early stopping compared to the baselines, as illustrated in Fig. 2.

Table 1. Our proposed method, UTLO, outperforms the baselines in terms of quantitative image quality metrics on the AnimalFaces-LT dataset ($64 \times 64$ resolution).

| Methods | FID ↓ | FID-FS ↓ | KID ↓ ×1000 | KID-FS ↓ ×1000 |
|---|---|---|---|---|
| StyleGAN2-ADA UnCond. [17] | 39.4 | 104.1 | 17.3 | 27.6 |
| StyleGAN2-ADA [17] | 51.4 | 87.1 | 24.7 | 35.9 |
| + Transitional [42] | 62.1 | 99.0 | 38.5 | 48.9 |
| + GSR[36] | 39.2 | 67.2 | 21.2 | 32.7 |
| + NoisyTwins [37] | 29.4 | 50.2 | 16.7 | 21.2 |
| + UTLO (Ours) | **26.2** | **48.4** | **12.6** | **19.6** |

**AnimalFaces-LT** This benchmark has a naturally occurring imbalance across its classes. We compare our method against the StyleGAN2-ADA benchmarks in terms of quantitative image quality metrics in Table. 1. Our method outperforms the baselines across all metrics, as demonstrated in the results. For a visual comparison against the baselines, please refer to Supp. H.

**Ablation on the Choice of Low-resolution for Unconditional Training ($res_{uc}$).** In our proposed method, one of the hyperparameter choices is to select an intermediate low-resolution $res_{uc}$ for unconditional training. All layers with equal or lower resolution than $res_{uc}$ do not receive class-conditional information. An unconditional GAN objective is added over the images and/or features at $res_{uc}$. To study the impact of $res_{uc}$, we conducted an ablation study on the AnimalFaces-LT dataset, containing images at $64 \times 64$. Table. 2 presents the results for selecting $res_{uc}$ from resolutions lower than $64 \times 64$, namely $8 \times 8$, $16 \times 16$, and $32 \times 32$. Studying the obtained results, we observe that resolutions of $8 \times 8$ and 16 achieve relatively close performance. Conversely, the performance diminishes when layers up to $32 \times 32$ are trained unconditionally. This is anticipated as the AnimalFaces-LT is at $64 \times 64$, leaving only one up-sampling layer to learn the class-conditional information which is shown to be insufficient. In all our experiments, we use $res_{uc}$ of $8 \times 8$ unless stated otherwise. Additional visual analysis on the role of $res_{uc}$ as well as an ablation on the choice of unconditional training objective weight $\lambda$ (see Eq. 2), need for unconditional layers in the discriminator, and distinction from "coarse-to-fine" training strategies are presented in the Supp. C.

Table 2. Ablation on choice of low resolution for unconditional training ($res_{uc}$).

| $res_{uc}$ | FID ↓ | FID-FS ↓ | KID ↓ ×1000 | KID-FS ↓ ×1000 |
|---|---|---|---|---|
| $8 \times 8$ | **26.2** | **48.4** | **12.6** | **19.6** |
| $16 \times 16$ | 27.5 | 50.3 | 13.7 | 20.8 |
| $32 \times 32$ | 38.0 | 64.9 | 23.3 | 34.3 |

**Can any low-resolution be the starting point for high-res generations?** In general, our observations indicate that high-res images are guided, yet not entirely restricted by the low-res images, as shown in Fig. 4. Each row in this figure presents a set of different high-resolution images that are generated from the same unconditional low-resolution image. This illustration reveals that high-resolution images can substantially differ from low-resolution ones, both in terms of color and texture. This suggests that subsequent conditional blocks are capable of significant modification in the background or foreground.

**Importance of FS metrics: comparison against unconditional training.** We also include results from training an unconditional StyleGAN2-ADA model [17] in Table 1, referred to as StyleGAN2-ADA UnCond. The unconditional model generates samples that follow the training distribution, which is mainly dominated by head classes. This bias favors unconditional learning in terms of the FID/KID metrics, which do not consider the skewness in the data distribution. Nevertheless, it can be observed the UnCond. baselines significantly perform worse in terms of few-shot metrics FID-FS/KID-FS. This suggests solely relying on FID/KID can be misleading. We strongly recommend including few-shot metrics, i.e., FID-FS and KID-FS, when evaluating cGANs on long-tailed datasets.

**CIFAR10-LT** In this benchmark, our method is compared with various baselines with imbalanced ratios $\rho$ of 50 and 100. This indicates the least frequent tail class includes only 100 and 50 training instances, respectively. Table. 3 quantitatively assesses the quality of the samples generated by our method and the baselines. All the baselines use the StyleGAN2-ADA [17] architecture. Our proposed method consistently outperforms the baselines across all metrics. This demonstrates that beyond improving generative learning across all classes (FID & KID), UTLO can also notably improve the learning from tail classes (FID-FS & KID-FS).

**Effect of Imbalance Ratio $\rho$.** Not surprisingly, all methods demonstrate improved performance with lower values of $\rho$, as seen in Table. 3. Further, we notice the benefits of knowledge sharing via UTLO become more pronounced as the imbalance in the dataset increases, i.e., learning tail classes becomes more challenging. It is also worth mentioning that the FID-FS and KID-FS metrics are critical to represent the quality of the few-shot classes as those are under-represented in the dataset-wide metrics that assume balanced sampling. For instance, at $\rho = 50$, the FID and KID metrics obtained from UTLO show little changes, while FID-FS and KID-FS exhibit improvements close to 50% in comparison to GSR.

Figure 4. Different class-conditional images generated given the same unconditional low-resolution images (left-most column).

Table 3. Quantitative comparison of the generated images from UTLO against baselines on CIFAR10-LT dataset with different imbalance ratios $\rho$. UTLO shows substantial improvements on few-shot (FS) metrics compared to StyleGAN2-ADA [17] while outperforming existing methods including NoisyTwins [37].

| $\rho$ | Methods | FID ↓ | FID-FS ↓ | KID ↓ ×1000 | KID-FS ↓ ×1000 |
|---|---|---|---|---|---|
| 50 | StyleGAN2-ADA [17] | 6.5 | 21.4 | 2.4 | 9.0 |
| | + Transitional [42] | 9.4 | 17.7 | 4.7 | 8.6 |
| | + GSR [36] | 6.4 | 21.3 | **2.3** | 8.1 |
| | + NoisyTwins [37] | 6.2 | 12.2 | 2.4 | 5.0 |
| | + UTLO (ours) | **6.1** | **11.8** | 2.4 | **4.8** |
| 100 | StyleGAN2-ADA [17] | 9.0 | 24.2 | 4.0 | 9.7 |
| | + Transitional [42] | 11.3 | 20.6 | 5.4 | 9.2 |
| | + GSR [36] | 8.4 | 24.3 | 3.9 | 11.8 |
| | + NoisyTwins [37] | 7.1 | 14.1 | 2.9 | 5.9 |
| | + UTLO (Ours) | **6.8** | **13.4** | **2.8** | **5.4** |

**Knowledge-sharing at Low Resolutions**. As explained in Sec. 3, our proposed generator builds on top of low-resolution (e.g., $8\times8$) unconditionally trained images which are subsequently used to generate conditional images at higher resolutions (e.g., $32 \times 32$). Several low-resolution images are shown in Fig. 6 along with their high-resolution conditional images generated for the head and tail classes of the dataset. It can be seen that conditional images generated from the tail classes evidently share certain features from the head classes. This demonstrates that knowledge sharing at low resolutions is an effective approach for infusing information from head classes to tail ones. A quantitative analysis of knowledge-sharing is presented in Supp. D. For additional qualitative examples see Supp. H.

**CIFAR100-LT** In addition to CIFAR10-LT, we evaluate our method and various baselines on the more challenging CIFAR100-LT dataset, wherein the tail classes include as few as five training instances. The quantitative results, shown in Table. 4, reveal that our method outperforms the baselines in a benchmark where a large number of classes with high diversity are present. In Fig. 2, we demonstrate the FID-FS curve during the course of the training along with the generated samples from the tail classes of the CIFAR100-LT dataset. This shows the effectiveness of our proposed method in addressing mode collapse while regularization methods diverge and require early stopping. Further visual comparisons of our proposed method against the baselines for CIFAR100-LT can be found in Supp. H.

**LSUN5-LT** We include LSUN5-LT as a higher resolution ($128 \times 128$) and highly-imbalanced ($\rho = 1000$) benchmark. In addition, we use Projected GAN with FastGAN[26] generator and DA [40]) as the baseline model. This is a different generator and discriminator design compared to StyleGAN2 along with a different aug-


Figure 5. Generated images from LSUN5-LT dataset. Despite only 50 training instances for the tail class `kitchen`, the proposed UTLO framework produces diverse, high-fidelity images.

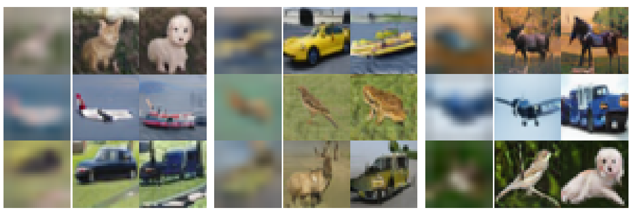

Figure 6. Knowledge sharing from head to tail classes in CIFAR10-LT dataset ($\rho = 100$) using UTLO. The conditional images generated from the head (middle columns) and tail (right columns) classes share and are built on top of the same low-resolution (unconditional) images (left columns). Low-resolution images ($8 \times 8$) are upsampled to that of CIFAR10-LT ($32 \times 32$) for improved visualization.

Table 4. Comparison of our proposed UTLO method with StyleGAN2-ADA baselines on the CIFAR100-LT dataset.

| Methods | FID ↓ | FID-FS ↓ | KID ↓ ×1000 | KID-FS ↓ ×1000 |
|---|---|---|---|---|
| StyleGAN2-ADA [17] | 10.8 | 24.9 | 5.1 | 9.3 |
| + Transitional [42] | 10.6 | 23.7 | **4.2** | 8.5 |
| + GSR[36] | 11.1 | 25.0 | 5.0 | 8.2 |
| + NoisyTwins [37] | 10.1 | 22.5 | 5.0 | 7.9 |
| + UTLO (Ours) | **9.9** | **21.8** | 4.6 | **7.5** |

mentation method. In conditional FastGAN implementation, the class-specific information is injected using class-conditional batch normalization, differing from the style mapping network utilized in StyleGAN2. UTLO can be readily adapted to this design by using standard batch normalization layers at low resolutions and utilizing class-conditional batch normalization only at higher resolutions. However, it is noteworthy that methods such as [37] are restricted to StyleGANs and cannot be applied in this context.

Table. 5 compares the results obtained from our method against the baseline and with the addition of GSR [36]. Our method consistently surpasses the baselines by around 30% across all metrics. As previously shown for the CIFAR10-LT (Table 3), UTLO not only improved on the tail classes but also significantly improved the dataset-wide FID/KID due to the high imbalance ratio ($\rho = 1000$).

Qualitative results in Fig. 5 contrast the generated images from our method against the baselines across all classes. For the `kitchen` class which only includes 50 training examples, the baselines struggle to effectively learn from the tail class whereas UTLO succeeds in generating diverse and high-quality images. More visual examples from our method and baselines are presented in Supp. H.

Table 5. Evaluating the quality of generated images from the proposed method and comparing against baselines on LSUN5-LT.

| Methods | FID ↓ | FID-FS ↓ | KID ↓ ×1000 | KID-FS ↓ ×1000 |
|---|---|---|---|---|
| PGAN (FastGAN)+DA [40] | 15.0 | 60.2 | 4.6 | 52.7 |
| + GSR [36] | 15.7 | 63.7 | 5.7 | 58.0 |
| + UTLO (Ours) | **10.9** | **43.6** | **3.5** | **35.3** |

Due to space constraints, additional results and analysis are provided in the supplementary material. e.g., Analysis of the role of weighted sampling (Supp. B.), The distinction of training class-conditional GANs in the Long-tailed setup vs. Limited-data setup (Supp. E.), and additional evaluation on Flowers-LT and iNaturalist2019 datasets (Supp. G).

## 5. Conclusion

In this paper, we proposed UTLO, a novel framework designed to improve the training of cGANs on long-tailed data. Inspired by the observation that head and tail classes often have more similarities at lower resolutions, our method facilitates knowledge sharing from head to tail classes using unconditional training at lower resolutions. The proposed method enriches the limited training distribution of the tail classes and effectively addresses mode collapse, leading to significant improvement in image generation for tail classes. We have also introduced the FID-FS/KID-FS metric, an adaptation of widely-used GAN metrics, specifically tailored for tail classes. We hope our findings are useful for future work in long-tail learning.

## Acknowledgements

This work is partially supported by the National Science Foundation (NSF) under grant 1751412 and 1927564.

## References


[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 6

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 1, 2

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[4] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *ieee transactions on pattern analysis and machine intelligence*, 2021. 12

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9268–9277. Computer Vision Foundation / IEEE, 2019. 12

[6] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, A aron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 1, 2

[8] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15089–15098, 2021. 3

[9] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 3

[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. *Advances in neural information processing systems*, 30, 2017. 6

[11] Grant Van Horn and Oisin Mac Aodha. The inaturalist 2019 competition dataset. `https://www.kaggle.com/c/inaturalist-2019-fgvc6`. 5, 12

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[13] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 3

[14] Bingyi Kang, Yu Li, Sai Nan Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021. 3

[15] Tejan Karmali, Rishubh Parihar, Susmit Agrawal, Harsh Rangwani, Varun Jampani, Maneesh Kumar Singh, and R. Venkatesh Babu. Hierarchical semantic regularization of latent spaces in stylegans. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, pages 443–459. Springer, 2022. 2

[16] Animesh Karnewar and Oliver Wang. MSG-GAN: multi-scale gradients for generative adversarial networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7796–7805. Computer Vision Foundation / IEEE, 2020. 4

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 1, 2, 4, 6, 7, 8, 12, 13, 16, 17, 18

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. Computer Vision Foundation / IEEE, 2020. 2, 4, 18

[19] Chaerin Kong, Jeesoo Kim, Donghoon Han, and Nojun Kwak. Few-shot image generation with mixup-based distance learning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, pages 563–580. Springer, 2022. 2

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *preprint*, 2009. 5, 12

[21] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10651–10662, 2022. 2

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[23] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and A. Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. *iclr*, 2021. 2, 4, 18

[24] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data:

A learnable embedding augmentation perspective. *Computer Vision And Pattern Recognition*, 2020. 3

[25] Shudong Liu and Ke Zhang. Under-sampling and feature selection algorithms for s2smlp. *IEEE Access*, 8:191803–191814, 2020. 3

[26] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 7

[27] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018. 3

[28] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. 3

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv: Arxiv-1411.1784*, 2014. 1, 2

[30] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *ICLR*, 2018. 4

[31] S. S. Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. *Ieee International Conference On Computer Vision*, 2019. 3

[32] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 5, 12

[33] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2642–2651. PMLR, 2017. 2

[34] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10743–10752, 2021. 2

[35] Harsh Rangwani, Konda Reddy Mopuri, and R. Venkatesh Babu. Class balancing gan with a classifier in the loop. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1618–1627. PMLR, 2021. 3

[36] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R. Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. *European Conference On Computer Vision*, 2022. 2, 3, 6, 7, 8, 12, 13, 16, 18, 19

[37] Harsh Rangwani, Lavish Bansal, Kartik Sharma, Tejan Karmali, Varun Jampani, and R Venkatesh Babu. Noisytwins: Class-consistent and diverse image generation through stylegans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5987–5996, 2023. 2, 3, 6, 7, 8, 18, 19

[38] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, pages 4175–4186. Curran Associates, Inc., 2020. 3

[39] Shibani Santurkar, Ludwig Schmidt, and A. Madry. A classification-based study of covariate shift in gan distributions. *International Conference On Machine Learning*, 2017. 12

[40] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 6, 7, 8, 18

[41] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 49:1–49:10. ACM, 2022. 1, 2, 15, 16

[42] Mohamad Shahbazi, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Collapse by conditioning: Training class-conditional GANs with limited data. In *International Conference on Learning Representations*, 2022. 1, 3, 6, 7, 8, 16, 18

[43] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 1

[44] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1354–1367, 2012. 5, 12

[45] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, pages 1513–1524. Curran Associates, Inc., 2020. 3

[46] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2021. 2

[47] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021. 2

[48] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[49] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130:1837–1872, 2022. 2, 3, 5, 6, 12

[50] Baofeng Yao and Lei Wang. An improved under-sampling imbalanced classification algorithm. In *2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 775–779. IEEE, 2021. 3

[51] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset

using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5, 12

[52] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 1

[53] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2019. 2, 12

[54] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Ieee/cvf Conference On Computer Vision And Pattern Recognition*, 2018. 16

[55] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv: Arxiv-2110.04596*, 2021. 1, 3

[56] Zizhao Zhang and Tomas Pfister. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 725–734, 2021. 3

[57] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020. 2, 3, 6, 12, 16, 18

[58] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv: 2006.02595*, 2020. 2

[59] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16489–16498. Computer Vision Foundation / IEEE, 2021. 3