Label Augmentation as Inter-class Data Augmentation for Conditional Image Synthesis with Imbalanced Data

Kai Katsumata Duc Minh Vo Hideki Nakayama
The University of Tokyo, Japan
{katsumata,vmduc,nakayama}@nlab.ci.i.u-tokyo.ac.jp

Abstract

*Conditional image synthesis performs admirably when trained on well-constructed and balanced datasets. However, in practice, training datasets frequently contain minorities (i.e., a class with a few samples), known as **imbalanced data**, which causes difficulties in learning generative models. To address conditional image synthesis with imbalanced data, we analyze a diversity issue of label-preserving data augmentation and an affinity issue of non-label-preserving data augmentation. From this observation, we present **label augmentation**, which works as inter-class data augmentation that effectively augments data by predicting a new label for a given image using the prediction of a pretrained image classification model (i.e., probabilities for each class). We incorporate our label augmentation into the discriminator of a seminal conditional generative adversarial network (GAN) model, proposing **Softlabel-GAN**. Using class probabilities extracts class-invariant and shared features between similar classes, achieving data augmentation with high affinity and diversity. Our experiments on imbalanced datasets show that Softlabel-GAN produces images with high quality and diversity while being hardly affected by the number of samples in each class. Code: https://github.com/raven38/softlabel-gan.*

1. Introduction

The impressive success of conditional deep generative models [12, 24, 37, 45, 48, 49, 62] has been largely aided by a large amount of well-collected, balanced, and diverse data, typically consisting of not only a large number of images in total but also a certain number of images in each class. Despite the enormous number of images available online, collecting special or rare objects is not always feasible owing to annotation costs, data constraints (*e.g.*, paintings of a specific artist), unauthorized data, and privacy concerns. As a result, imbalanced data [20] is inevitable in real-world scenarios, leading to the failure of generative models and amplifying biases. The potential risk of the latter has been discussed in terms of AI ethics [2, 35, 50]. Training generative models in the imbalanced data regime is thus considered necessary, potentially broadening generative models' real-world applications and providing safety to them.

Figure 1. Generated examples on AnimeFace [1]. DiffAug-GAN (DiffuAug) generates images with low diversity (red rectangles). Smooth-GAN (Smooth) generates images that differ from the given class (dashed black rectangles). In contrast, Softlabel-GAN (Ours) can avoid overlapped images and images without respect to the given class. Each class has five samples.

Data augmentation is a straightforward solution for image generation with limited data. According to [13], two important aspects of data augmentation are affinity and diversity, with affinity indicating how small the distribution shifts between the augmented and training data distributions and diversity indicating the complexity of the augmented data. Low-affinity data causes the model to learn incorrect features, whereas low-diversity data causes the model to easily memorize training data, degrading the quality and diversity of generated images. Achieving augmentation with high affinity and diversity is difficult for imbalanced data because of two reasons: (i) the number of training samples is insufficient in minor classes and (ii) the minor class samples are not diverse. As a result, data augmentation with high affinity and diversity is vital for successfully conditional image synthesis with imbalanced data.

Data augmentation consists of two parts: label-preserving data augmentation [11, 25, 56, 64, 65] and non-label-preserving data augmentation [54, 61]. The former part directly modifies only image inputs, whereas the lat-

ter one does the labels of inputs. The label-preserving data augmentation for images mostly employs geometric transformations, image filters, and color intensity transformations, and their operations are limited to those that maintain input labels. On the other hand, non-label-preserving data augmentation allows arbitrary operations. Label-preserving data augmentation is widely used in generative adversarial networks (GANs) [25, 56, 64, 65], with notable results on limited and balanced data. However, in our case, where the data is not only limited but also imbalanced (*i.e.*, the appearance of minor classes), the current data augmentation is insufficient to expand the data distribution. The main reason is that such data augmentation creates additional data by solely reusing the samples within each class, leading to augmented data with high affinity yet low diversity. For instance, simply applying geometric, color, corruption, and/or filtering transformations to the training data (*e.g.*, DiffAugment [64]) leads to a rapid imbalance between a generator and a discriminator, yielding images with low diversity (*i.e.*, repeat almost the same images) (Fig. 1).

Our main idea is to predict a new label for a given image by assigning probabilities to all classes to which the image belongs. Therefore, our novel augmentation is a type of non-label-preserving data augmentation, which we call label augmentation. We approach our problem by incorporating our simple yet effective label augmentation into the discriminator of a seminal cGAN model. More precisely, we first prepare an image classification model pretrained on the imbalanced data. Then, we use the output of the softmax function (*i.e.*, probabilities for each class) obtained through the pretrained classifier as the class condition in the discriminator. Using class probabilities enables our model to take into account semantic similarities between classes with respect to the perception of pretrained classifiers. Naive label augmentation methods [54,61] blindly distribute class probabilities to the data, resulting in augmented data with high diversity but low affinity, as well as the generation of images that are irrelevant to the given class (Fig. 1). By contrast, our augmentation precisely distributes class probabilities, resulting in augmented images with both high affinity and diversity. Consequently, our method hinders the training of the discriminator and balances the generator and the discriminator, yielding better-generated images (Fig. 1). Our contributions can be summarized as follows:

- We observe that existing data augmentation approaches provide either diversity or affinity for imbalanced data.
- We find that assigning classifier output with sufficient entropy to samples can be interpreted as interclass data augmentation that increases the diversity per class. We thus propose a simple yet effective label augmentation method that produces augmented data with high affinity and diversity using a pretrained classifier.
- We propose Softlabel-GAN, which uses our novel label augmentation, for conditional image synthesis in an imbalanced data regime. To the best of our knowledge, this is the first study that investigates label augmentation in learning cGANs with imbalanced data.
- We demonstrate, on several imbalanced datasets, that our method outperforms the other methods. Furthermore, the experiments show our advantages with intra-class fidelity and diversity.

2. Related Work

Image generation with limited data aims to improve the training stability and generation quality without the immense amount of data. Collecting large high-quality datasets is not always possible because of the tremendous annotation cost, data constraints, and privacy. In some cases, we only collect a few examples for each class, *e.g.*, photos of a specific landmark or illustrations of a specific artist. Since training GANs without a huge amount of data is crucial, several studies [16, 21, 25, 51, 64] paid attention to the data efficiency aspect. Some approaches [25, 64] are able to learn from limited data by using data augmentation (*i.e.*, label-preserving data augmentation). In contrast to data augmentation-based approaches, another line of research [7,26,36,53] employs semi- and self-supervised learning to reduce the cost of human annotation. Recent works [42, 44] design architecture-specific methods for image generation with limited data. We introduce a data augmentation approach, which achieves high-affinity and high-diversity augmentation on imbalanced data.

Non-label-preserving augmentation is a set of data augmentation. While the label-preserving augmentation directly modifies an input image, the non-label-preserving augmentation modifies an input label. Of which, label smoothing [54] and Mixup [61] are popularly used. Label smoothing [54] replaces hard targets with soft targets by taking the weighted average of the original targets, avoiding overconfidence on several tasks [9, 57]. Online label smoothing [60] quantifies class similarities and then assigns class-wise soft labels, unlike our method, which assigns instance-wise soft labels. Mixup [61] trains a network on convex combinations of the samples and their labels to improve accuracy and robustness to hyperparameters. These methods [54, 61] often provide incorrect information to cGANs owing to blindly distributing class probabilities, resulting in the high diversity yet low affinity of augmented data. In this work, we present content-aware label augmentation for imbalanced data, which builds augmented data with high affinity and diversity.

Pretrained recognition models have been widely used in training GANs. To generate images considering their contents, the high-level features extracted from a pretrained model as an alternative to the human visual perception are

Table 1. Comparison of balanced and imbalanced datasets. For each dataset, we indicate the range of the number of images in each class, a ratio (largest class samples/smallest class samples), overall samples, and the number of images per class and the ratio after our label augmentation. The number of samples per class on the imbalanced dataset is various and significantly less than that on the balanced one. A higher ratio indicates more imbalance. Our augmentation actually increases the minor class samples (last column).

Dataset	#Samples per class	Ratio	#Samples	#Augmented samples	Augmented ratio
Balanced datasets					
CIFAR-10 [28]	5000	1×	50000	—	—
CIFAR-100 [28]	500	1×	50000	—	—
Imbalanced datasets					
AnimeFace [1]	17–161	9.5×	14490	109–792	7.2×
Oxford-102 Flowers [41]	40–258	6.5×	8189	50–315	6.3×
Imbalanced CIFAR-10	65–1720	26.5×	3208	149–2232	14.9×
Imbalanced CIFAR-100	6–36	6×	2993	35–176	5×
Imbalanced Tiny ImageNet	9–586	65.1×	47602	60–680	11.3×
Stanford Cars [27]	24–68	2.8×	8144	73–263	3.6×

used [8, 22, 31]. The feature distance between two images with the pretrained VGG [52] has been widely used in style transfer [22] and super-resolution [31]. Knowledge distillation [15] transfers knowledge from a teacher model to a student model that solves the same task as the teacher solves for model compression [6, 32, 47, 58]. A concurrent work [10] proposes a CLIP [43]-based knowledge distillation method and exploits huge external knowledge for image generation. In contrast, we aim to share knowledge among classes to enhance minority classes.

3. Preliminary Knowledge

3.1. Imbalanced dataset

Depending on the number of samples in each class, any dataset can be classified as either a balanced or imbalanced dataset [5, 20, 34]. While a balanced dataset possesses classes with roughly the same number of samples in each, an imbalanced dataset possesses some classes with a few samples. As we can see in Tab. 1, the ratios between the major and minor classes of imbalanced datasets are much higher than those of balanced datasets.

On the basis of the above definition [34], we collect some imbalanced datasets to verify our method. We use AnimeFace [1], Oxford-102 Flowers [41], imbalanced CIFAR-10, imbalanced CIFAR-100, imbalanced Tiny ImageNet, and Stanford Cars [27] in our experiments. Note that we construct imbalanced CIFAR-10/100 and Tiny ImageNet from the original ones [28, 59] (see Sec. 5.1 for more details). For comparison, we list the number of classes and the number of samples per class for balanced and imbalanced datasets in Tab. 1, showing that the imbalanced datasets used in our experiments are more challenging.

3.2. Conditional GANs

Conditional GANs aim to model the conditional distribution of a target dataset using a generator $G : \mathbb{R}^{d_z} \times \mathbb{R}^k \to \mathbb{R}^d$ and a discriminator $D : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$, where d and d_z are the dimensions of an image and a latent variable, respectively. Here, the class label $\boldsymbol{y} \in \mathbb{R}^k$ with k being the number of classes indicates the probabilities of an instance belonging to each class, including one-hot vectors. The generator G maps condition $\boldsymbol{y}^f \in \mathbb{R}^k$ and latent vector $\boldsymbol{z} \in \mathbb{R}^{d_z}$ from a prior distribution $p(\boldsymbol{z})$ to output $\boldsymbol{x}^f = G(\boldsymbol{z}, \boldsymbol{y}^f) \in \mathbb{R}^d$. The discriminator D learns to distinguish between the generated distribution p and the target distribution q. The discriminator receives either a pair of a real sample $\boldsymbol{x}^r \in \mathbb{R}^d$ and a corresponding label $\boldsymbol{y}^r \in \mathbb{R}^k$ or a fake pair $(\boldsymbol{x}^f, \boldsymbol{y}^f)$. The objective functions of cGANs are

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{\boldsymbol{y}^r \sim q(\boldsymbol{y}), \boldsymbol{x}^r \sim q(\boldsymbol{x}|\boldsymbol{y})} \left[f_D(-D(\boldsymbol{x}^r, \boldsymbol{y}^r)) \right] \\ &+ \mathbb{E}_{\boldsymbol{y}^f \sim p(\boldsymbol{y}), \boldsymbol{z} \sim p(\boldsymbol{z})} \left[f_D(D(G(\boldsymbol{z}, \boldsymbol{y}^f), \boldsymbol{y}^f)) \right], \quad (1) \\ \mathcal{L}_G =& \mathbb{E}_{\boldsymbol{y}^f \sim p(\boldsymbol{y}), \boldsymbol{z} \sim p(\boldsymbol{z})} \left[-D(G(\boldsymbol{z}, \boldsymbol{y}^f), \boldsymbol{y}^f) \right], \quad (2)\end{aligned}$$

where $f_D(\cdot) = \max(0, 1 + \cdot)$ is the hinge loss [33, 37, 55]. Conventional cGANs optimize the above functions, leading to the generated distribution being close to $q(\boldsymbol{x}|\boldsymbol{y})$ on a well-constructed dataset. In contrast, we aim to learn a distribution that is close to $q(\boldsymbol{x}|\boldsymbol{y})$ on an imbalanced dataset.

4. Proposed Method

4.1. Label augmentation for imbalanced data

Imbalanced data possesses a special property, whereas some classes have a certain number of samples (*i.e.*, major classes) while others do not (*i.e.*, minor classes). The existence of minor classes leads to the failure of straightforward data augmentation, which directly applies transformations to training images. The reason can be explained as follows. Label-preserving data augmentation uses samples within each class to augment data. This strategy works well for major classes while it cannot provide enough diversity given minor classes. Namely, the augmented data have high affinity yet low diversity. Obviously, they easily trigger learning short-cut features.

To address the inadequacy of label-preserving data augmentation, we focus on increasing the affinity and diversity of augmented data. Needless to say, all training data would share some features such as color and shape. Inspired by the above observation, our method is designed to allow a class to implicitly borrow samples from other similar classes as augmented samples rather than reusing samples within the class. An appropriate way is non-label-preserving augmentation, in which a single image is shared by multiple classes. However, naive label augmentation methods automatically distribute the probability. Label smoothing [54] makes new

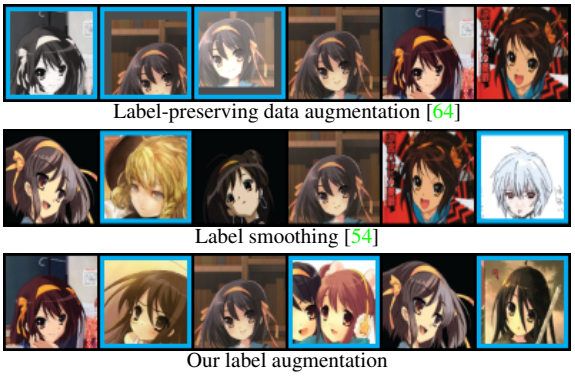


Figure 2. Training samples from the same class after applying each augmentation method (*i.e.*, images that assign probability to the class). The blue rectangles mean augmented samples, and the others are original samples of the class. Unlike other methods, our label augmentation imports similar images from other classes, resulting in augmented data with high affinity and diversity.

labels dissociate from the image content. Mixup [61] expands the class distribution by using convex combinations even for dissimilar class pairs.

We thus develop a distribution manner that distributes probabilities associated with each input image. Namely, we employ a pretrained image classifier for our label augmentation. Assigning the predictions of the classifier to samples as new labels enables the discriminator to consider the relationships between classes in training. Our label augmentation, therefore, facilitates learning with the proper information and balancing the generator and the discriminator. As shown in Fig. 2, while typical data augmentation methods complete data augmentation within each class, label augmentation does classes by importing instances from other classes. Therefore, label augmentation can be interpreted as inter-class data augmentation.

We compare augmented data obtained by different augmentation methods (Fig. 2). Label-preserving data augmentation [64] only reproduces images similar to original images, resulting in augmented data with high affinity and low diversity. Label smoothing [54] imports images different from the original images, resulting in low affinity and high diversity. In contrast to the above methods, our label augmentation imports images (from other classes) that are similar to the original class images (*e.g.*, the characteristics of the same hair color and similar painting style), resulting in augmented data with high affinity and diversity.

4.2. Softlabel-GAN

We propose Softlabel-GAN by incorporating our label augmentation into the discriminator of a cGAN model. More precisely, we feed the probability vectors of input images to the discriminator instead of one-hot ground truths. In what follows, we will formally present our label augmentation as well as how to combine it with the discriminator.

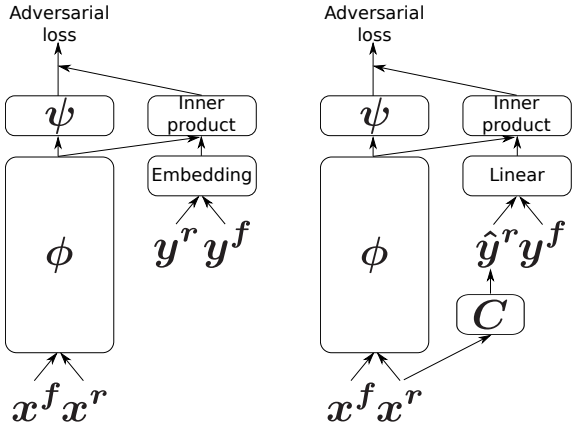


Figure 3. Discriminator architectures for cGANs: a projection discriminator and a discriminator with our label augmentation. (a) The typical conditional discriminator receives a pair of an image and a label: $(\boldsymbol{x}^f, \boldsymbol{y}^f)$ or $(\boldsymbol{x}^r, \boldsymbol{y}^r)$. (b) Our discriminator receives a pair of an image and a probability vector obtained from a pretrained classifier $\hat{\boldsymbol{y}}^r = C(\boldsymbol{x}^r)$: $(\boldsymbol{x}^f, \boldsymbol{y}^f)$ or $(\boldsymbol{x}^r, \hat{\boldsymbol{y}}^r)$.

Label augmentation definition. Let $T : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d \times \mathbb{R}^k$ be a label augmentation function. We define $T(\boldsymbol{x}, \boldsymbol{y}) = (T_{\boldsymbol{x}}(\cdot), T_{\boldsymbol{y}}(\cdot))$ where $T_{\boldsymbol{x}}$ is an image prediction function (*i.e.*, predicting a modified image) and $T_{\boldsymbol{y}}$ is a label prediction function (*i.e.*, predicting a new label for given image). Our $T_{\boldsymbol{x}}$ is the identity function. Unlike $T_{\boldsymbol{y}}$ used in [54, 61], which directly predicts a new label from a given label $\boldsymbol{y}$, our $T_{\boldsymbol{y}}$ predicts a new label $\hat{\boldsymbol{y}}$ from a given image $\boldsymbol{x}$.

As discussed above, we aim to distribute class probabilities to the given image $\boldsymbol{x}$ precisely, indicating that $\hat{\boldsymbol{y}}$ is assigned to multiple classes. Therefore, we use a pretrained classifier $C : \mathbb{R}^d \to \mathbb{R}^k$ as $T_{\boldsymbol{y}}$. The new label $\hat{\boldsymbol{y}}$ is obtained using C as $\hat{\boldsymbol{y}} = C(\boldsymbol{x})$. In other words, $T(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, \hat{\boldsymbol{y}})$.

Integration of label augmentation and discriminator. We now integrate $\hat{\boldsymbol{y}}$ obtained by our augmentation into Eq. (1), defining the objective of Softlabel-GAN as

$$\begin{aligned}\mathcal{L}_D = &\mathbb{E}_{\boldsymbol{y}^r \sim q(\boldsymbol{y}), \boldsymbol{x}^r \sim q(\boldsymbol{x}|\boldsymbol{y})} \left[f_D(-D(\boldsymbol{x}^r, \hat{\boldsymbol{y}}^r)) \right] \\ &+ \mathbb{E}_{\boldsymbol{y}^f \sim p(\boldsymbol{y}), \boldsymbol{z} \sim p(\boldsymbol{z})} \left[D(G(\boldsymbol{z}, \boldsymbol{y}^f), \boldsymbol{y}^f) \right]. \end{aligned} \quad (3)$$

We also employ Eq. (2) as the objective function of the generator for our training scheme. The generator only takes one-hot inputs $\boldsymbol{y}^f$ in both the training and testing phases. Our discriminator takes $\hat{\boldsymbol{y}}$ for real samples and $\boldsymbol{y}^f$ for fake samples, which are sampled from the uniform distribution. We apply our label augmentation to only real samples because our augmentation aims to correct the class imbalance in the dataset. Figure 3 illustrates the difference between our discriminator and the widely used projection discriminator [38]. Note that in Fig. 3, we omit the generator of Softlabel-GAN for simplicity because we maintain the generator of projection-based GANs.

Our method increases the diversity of the augmented

data while maintaining high affinity by importing similar images from other classes. This is because it assigns higher probabilities to proper classes (*e.g.*, correct or similar classes) and lower probabilities to improper classes (*e.g.*, irrelevant or dissimilar classes) according to the perception of a pretrained classifier. As opposed to our method, the data augmentation-based methods [25,64] limit the diversity due to only reusing a few samples inside each class. By augmenting data with high affinity and diversity, our method prevents just memorizing training data.

Generally, we can use another function as T. Label smoothing [54] (*i.e.*, $T(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{y}(1-\alpha) + \alpha \mathbf{1}/k)$ and Mixup [61] (*i.e.*, $T(\boldsymbol{x}, \boldsymbol{y}) = (\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}', \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{y}')$ have label prediction functions $T_{\boldsymbol{y}}(\boldsymbol{y}) = \boldsymbol{y}(1-\alpha) + \alpha \mathbf{1}/k$ and $T_{\boldsymbol{y}}(\boldsymbol{y}) = \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{y}'$, respectively. Unlike our method taking $T_y : \mathbb{R}^d \to \mathbb{R}^k$, naive label augmentations take $T_y : \mathbb{R}^k \to \mathbb{R}^k$ (*i.e.*, predicting a new label from a given label). Augmenting a class without considering actual image contents results in the low-affinity augmented data.

We checked that our label augmentation works as data augmentation. The number of samples that assigned a probability above a threshold of 0.01 to a class was counted as the number of samples belonging to the class. Our label augmentation indeed increases the minor class size (Tab. 1, last column).

4.3. Implementation details

We use BigGAN [4] to examine Softlabel-GAN. We build Softlabel-GAN upon BigGAN [4] by integrating our label augmentation and DiffAugment [64] with three transformations: translation (within $[\frac{-1}{8}, \frac{1}{8}]$ of the image size), color (including random brightness within $[-0.5, 0.5]$, contrast within $[0.5, 1.5]$, and saturation within $[0, 2]$), and cutout (masking with a random square of half image size).

We use SpinalNet [23] as pretrained classifier C because it achieves stat-of-the-art performance on fine-grained datasets and empirically works well on imbalanced data. We train SpinalNet on a target dataset with the entropy regularization term. Empirically, the weight of the term is set to 0.3 for all datasets. We then use the trained classifier to realize our label augmentation. The classifier feeds on input before applying DiffAugment. Note that our method does not require a perfect classifier because we aim to assign probabilities to similar classes.

We convert input probability vectors by a fully connected layer without a bias term instead of an embedding layer to accept $\hat{\boldsymbol{y}} \in \mathbb{R}^k$ (*i.e.*, a non-one-hot vector). Then, we use the discriminator $D(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})\boldsymbol{V}\boldsymbol{y} + \psi(\phi(\boldsymbol{x}))$ with the feature extractor $\phi : \mathbb{R}^d \to \mathbb{R}^l$, the discriminator head $\psi : \mathbb{R}^l \to \mathbb{R}$, and the weights of the linear layer $\boldsymbol{V} \in \mathbb{R}^{l \times k}$ (Fig. 3).

For the experiments with the resolution of 32×32, we set the latent dimension $d_{\boldsymbol{z}} = 128$. We use a minibatch size of 128 and a learning rate of 2×10^{-4} for both the generator and discriminator. For the experiments with higher resolution, we use the hierarchical latent space with 20 dimensions for each latent variable and the shared embedding with 128 dimensions. We use minibatch sizes of 512 and 32 for the resolutions of 64×64 and 128×128, respectively. We use the learning rates of 1×10^{-4} and 4×10^{-4} for the generator and discriminator, respectively.

5. Experiments

5.1. Datasets

AnimeFace [1] is constructed by extracting face regions from the images of anime characters obtained from the web. It consists of 176 characters (*i.e.*, classes), where each class contains between 17 and 161 images (128×128).

Oxford-102 Flowers [41] consists of 102 flower classes. The smallest class has 40 images and the largest one has 258 images. All images are resized to 128×128.

Imbalanced CIFAR-10 is an imbalanced subset of CIFAR-10 [28]. The original CIFAR-10 contains 50,000 32×32 images as the training set. The building procedure for this dataset consists of three steps. First, we shuffle the order of class labels. Then, to define the frequency of each class, we consider a histogram that has the same number of bins as the number of classes and approximates the range $[0, 3)$ in a lognormal distribution with a standard deviation of 3. Each sorted class corresponds to one bin, and we finally randomly pick up samples so that the overall class frequency follows this histogram. The dataset contains 3208 color images, with 65–1720 images per class.

Imbalanced CIFAR-100 is an imbalanced subset of 32×32 CIFAR-100. We build it in the same manner as imbalanced CIFAR-10 and use a χ^2-distribution with 3 degrees of freedom instead of a lognormal distribution. The dataset consists of 2993 images with 6–36 images per class.

Imbalanced Tiny ImageNet is a subset of Tiny ImageNet 128×128 [59] (200 classes). We take it in the same manner as imbalanced CIFAR-10/100 with the range $[1, 4)$ in a Pareto distribution with a shape parameter, $\alpha = 2$. It contains 47602 images with 41–483 images per class.

Stanford Cars Dataset [27] consists of 196 classes with 24–68 images per class. All images are resized to 128×128. The dataset provides 8144 images.

5.2. Compared methods and evaluation metrics

Compared methods. We use BigGAN [4] as a base model and carefully integrate data and label augmentations into it. Since BigGAN [4] cannot work properly without data augmentation as seen later in Sec. 5.4, we employ DiffAugment [64] for all the compared methods.

For comparison (Sec. 5.3), we compare Softlabel-GAN with DiffAug-GAN (*i.e.*, BigGAN [4] with DiffAug-

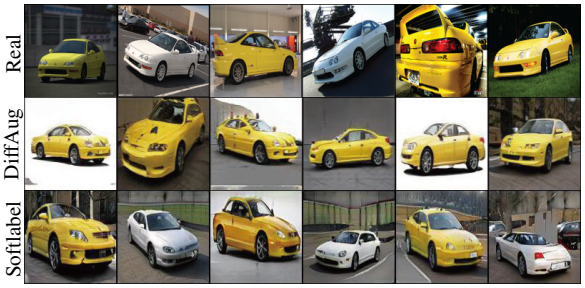


Figure 4. Generated examples on Stanford Cars [27]. DiffAug-GAN cannot generate the color variations of target classes.

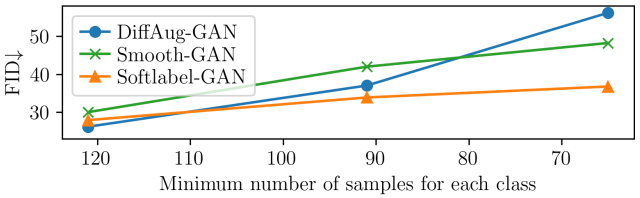


Figure 5. As the number of samples in the minor classes decreases, the FIDs obtained by DiffAug-GAN and Smooth-GAN increase considerably, implying that those methods worsen. In contrast, Softlabel-GAN achieves a relatively consistent performance regardless of the minimum number of samples for each class.

ment [64]). We also use Smooth-GAN (*i.e.*, BigGAN [4] with DiffAugment [64] and label smoothing [54] with α of 0.1) as a strong baseline with naive label augmentation.

For detailed analysis (Sec. 5.4), we use two ablated models: Softlabel-GAN^{--}, which is our model without both augmentations (*i.e.*, BigGAN [4]), and Softlabel-GAN^{-}, which is our model without data augmentation (*i.e.*, BigGAN [4] with our label augmentation). Additionally, we prepare Smooth-GAN^{-} (*i.e.*, BigGAN [4] with label smoothing [54]).

Evaluation metrics. We employ Inception Score (IS) [46], Fréchet Inception Distance (FID) [14], and LPIPS [63] diversity score. In addition, we use Precision, Recall, Density, and Coverage [30, 39]. We also calculate intra-class metrics: intra-FID [38], intra-LPIPS, intra-Precision, intra-Recall, intra-Density, and intra-Coverage to more extensively evaluate the quality within each class. Intra-FID, intra-LPIPS, intra-Precision, intra-Recall, intra-Density, and intra-Coverage are the averages of the FID, LPIPS, Precision, Recall, Density, and Coverage calculated for each class, respectively. For FID and intra-FID, we sample 10K generated images. For LPIPS and intra-LPIPS, we sample 100 generated images for each class.

5.3. Experimental results

Qualitative comparison. Figures 1 and 4 provide the examples generated by Softlabel-GAN and the baselines, showing that our method succeeds in the plausible and diverse image generation on imbalanced training data. On AnimeFace (Fig. 1), all methods can produce plausible images. However, from the perspective of the distribution of generated images, DiffAug-GAN produces only a few modes regardless of latent variables, resulting in less diverse images, and Smooth-GAN generates images with a wrong class. On Stanford Cars (Fig. 4), DiffAug-GAN generates images with low diversity. By contrast, our method generates the color variations of car models. We provide more examples in Supplementary Material.

Quantitative comparison. Table 2 provides quantitative results on the six datasets. Our method outperforms the others in FID and intra-FID on all datasets, demonstrating the superiority of Softlabel-GAN in overall and per-class performance. Table 2 also shows that our method achieves higher LPIPS and intra-LPIPS scores than other methods in most cases, indicating that our generated images are more diverse. Tight standard deviations of intra-FID and intra-LPIPS of our method indicate consistent performance in each class. Unlike in the case of FID, which neglects the minor classes, Tab. 2 demonstrates the superiority of our method in terms of intra-class metrics, which considers the quality of minor classes. Smooth-GAN sometimes achieves the highest diversity score because images that ignore a given class condition result in a higher diversity score, but not actual diversity. Unlike Smooth-GAN, Softlabel-GAN generates images always consistent with a given class (see Figs. 1 and 7). As shown in Fig. 1 and Tab. 2, our method achieves diverse image generation. We note that KD-DLGAN [10] achieves a FID of 11.63 on Stanford Cars.

Next, we evaluated the effects of the number of minor class samples on the performance of the compared methods. To this end, we train the compared models on imbalanced CIFAR-10 using variable sizes of minor class samples (*i.e.*, 120, 90, and 65 samples, respectively). Figure 5 indicates that while DiffAug-GAN and Smooth-GAN significantly deteriorate their performance as the number of samples decreases, our method achieves stable performance even with the limited amount of data. When compared with DiffAug-GAN, our method reduces performance degradation by 70%. These observations clearly show the benefits of our method in boosting the robustness to minor classes.

5.4. Detailed analysis

We evaluated the necessity of DiffAugment [64]; see the first three rows of Tab. 3. For this ablation study, we use AnimeFace and Oxford-102 at the resolution of 64×64. Softlabel-GAN achieves better FID over the ablation models (Tab. 4). We can see the necessity of the DiffuAugment baseline. In addition, our label augmentation further brings a performance gain over the baseline.

To confirm the contribution of each augmentation, we explore the FID and LPIPS (computed on each class) ob-

Table 2. Quantitative results on the six benchmark datasets.

Method	FID↓	LPIPS↑	intra-FID↓	intra-LPIPS↑	Precision↑	Recall↑	Intra-Prec↑	Intra-Rec↑	Density↑	Coverage↑	Intra-Dens↑	Intra-Cov↑
	AnimeFace											
DiffAug-GAN	25.09	0.5049	66.25±17.00	0.4018±0.0266	0.877	0.231	**0.967**	0.027	1.220	0.538	**1.437**	0.975
Smooth-GAN	22.46	0.5111	64.40±15.87	0.4211±0.0226	0.885	0.319	0.946	0.077	1.271	0.589	1.372	0.985
Softlabel-GAN	**19.14**	**0.5183**	**57.43**±16.56	**0.4627**±0.0314	**0.890**	**0.454**	0.952	**0.225**	**1.442**	**0.659**	1.365	**0.988**
	Oxford-102 Flowers											
DiffAug-GAN	28.70	0.5826	159.51±57.66	0.3964±0.0566	0.795	0.435	**0.848**	0.051	0.484	0.408	0.527	0.850
Smooth-GAN	23.36	**0.5961**	141.47±47.02	0.4313±0.0463	0.798	0.547	0.798	0.205	0.567	0.478	0.571	0.924
Softlabel-GAN	**20.97**	0.5793	**126.32**±42.69	**0.4696**±0.0407	**0.815**	**0.585**	0.796	**0.300**	**0.613**	**0.513**	**0.612**	**0.927**
	Imbalanced CIFAR-10											
DiffAug-GAN	57.24	0.2090	112.96±26.62	0.1647±0.0154	0.742	**0.468**	0.557	**0.343**	0.435	0.393	0.277	0.569
Smooth-GAN	66.75	0.2017	125.31±35.36	0.1751±0.0197	0.697	0.399	0.502	0.317	0.392	0.382	0.238	0.567
Softlabel-GAN	**54.59**	**0.2058**	**109.79**±24.72	**0.1773**±0.0167	**0.756**	0.408	**0.590**	0.337	**0.477**	**0.438**	**0.308**	**0.590**
	Imbalanced CIFAR-100											
DiffAug-GAN	37.70	**0.2774**	209.54±37.19	0.1543±0.0381	**0.830**	0.416	**0.742**	0.080	**0.703**	0.680	**0.502**	0.902
Smooth-GAN	34.36	0.2570	205.59±34.82	0.1704±0.0382	0.772	0.500	0.714	0.112	0.619	0.701	0.476	0.914
Softlabel-GAN	**32.70**	0.2438	**201.61**±32.50	**0.1948**±0.0352	0.770	**0.577**	0.699	**0.262**	0.635	**0.712**	0.446	**0.948**
	Stanford Cars											
DiffAug-GAN	8.99	0.5664	95.00±12.62	0.5437±0.0207	0.869	0.616	0.921	0.310	**1.389**	**0.856**	**1.247**	**1.000**
Smooth-GAN	7.91	**0.5863**	103.55±12.56	0.5436±0.0171	0.857	0.657	0.894	0.436	1.217	0.851	0.990	0.999
Softlabel-GAN	**7.35**	0.5855	**89.04**±11.64	**0.5452**±0.0168	**0.884**	**0.664**	**0.927**	**0.455**	1.228	0.851	0.972	**1.000**

	Imbalanced TinyImagenet												
	IS↑	FID↓	LPIPS↑	intra-FID↓	intra-LPIPS↑	Precision↑	Recall↑	Intra-Prec↑	Intra-Rec↑	Density↑	Coverage↑	Intra-Dens↑	Intra-Cov↑
DiffAug-GAN	4.33	159.74	0.5496	332.61±34.55	0.2630±0.0319	0.341	0.000	0.139	0.000	0.080	0.017	0.022	0.049
Smooth-GAN	4.48	151.10	0.5573	334.39±33.84	0.3991±0.0544	0.352	0.003	0.128	0.001	0.075	0.020	0.016	0.070
Softlabel-GAN	**14.58**	**53.22**	**0.6631**	**238.83**±53.74	**0.5755**±0.0501	**0.648**	**0.647**	**0.475**	**0.270**	**0.357**	**0.264**	**0.212**	**0.634**

Table 3. Ablation study on the imbalanced datasets using FID.

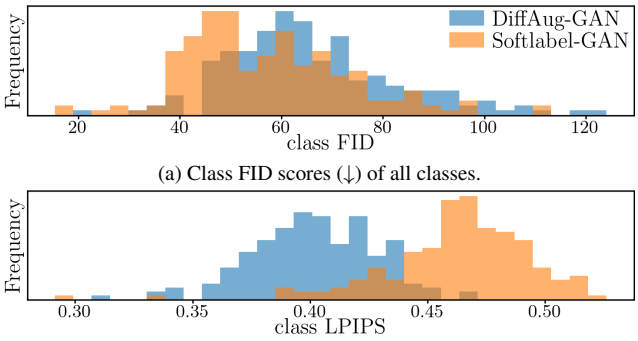
Method	AnimeFace	Oxford-102	CIFAR-10	CIFAR-100
Softlabel-GAN	**17.26**	**46.32**	**36.72**	**49.99**
Softlabel-GAN^{--}	36.16	84.36	79.16	80.99
Softlabel-GAN^{-}	32.89	60.31	72.81	73.94
Smooth-GAN^{-}	40.45	78.20	83.68	72.80

Table 4. Comparison with Mixup and Oversample using FID.

Method	AnimeFace	Cars	Oxford102	CIFAR-10	CIFAR-100
Softlabel-GAN	**19.14**	**7.35**	**20.97**	**54.59**	**32.70**
Mixup	24.87	14.53	39.13	56.36	36.32
Oversample	26.07	9.83	29.36	59.03	40.83

Table 5. FID scores in the experiments at 256×256 resolution.

	DiffAug-GAN	Smooth-GAN	Softlabel-GAN
AnimeFace	20.5066	20.6889	**17.2493**
Stanford Cars	10.1542	23.3695	**9.8605**



(a) Class FID scores (↓) of all classes.

(b) Class LPIPS scores (↑) of all classes.

Figure 6. Histogram of class FID and class LPIPS scores on AnimeFace. The mode to the left in FID and the mode to the right in LPIPS indicate good performance. The narrow distribution means achieving consistent performance. The performance of DiffAug-GAN varies by class more than Softlabel-GAN.

Figure 7. Comparison of GANs with label augmentation. All images are generated with the same class, but some images contain a different class from the the other images in Smooth-GAN. Each class has four samples.

tained by both methods in depth. Figure 6 shows the histograms of FID and LPIPS for each class on AnimeFace. DiffAug-GAN has a larger variance of class FID and LPIPS, varying the performance by each class data. It has the distributions of FID and LPIPS with longer right and left tails than Softlabel-GAN, respectively. Softlabel-GAN wins 152 classes out of 176 against DiffAug-GAN in FID and wins 175 classes in LPIPS. To sum up, the proposed method reduces standard deviation in addition to improving performance, indicating that label augmentation and data augmentation make different contributions. Along with qualitative and quantitative results (Sec. 5.3), our method is promising in dealing with imbalanced data.

Figure 7 shows the examples generated by cGANs with label augmentation, *i.e.*, Softlabel-GAN and Smooth-GAN. Note that we use the same class to generate images (*i.e.*, two first rows belong to a class and two last rows belong to another class). Unlike Softlabel-GAN, Smooth-GAN generates images of a different class from a conditional class.

Table 6. FID scores in experiments with the ADC-GAN [19] baseline.

	DiffAug-GAN	Smooth-GAN	Softlabel-GAN
AnimeFace	35.0864	23.1834	**21.0494**
Stanford Cars	20.6485	23.4987	**14.5688**

Table 7. Comparison with the diffusion model with classifier-free guidance (CFG) on TinyImageNet in FID, intra-KID, intra-LPIPS, intra-Precision (i-P), intra-Recall (i-R), intra-Density (i-D), and intra-Coverage (i-C).

Method	FID	intra-KID	intra-LPIPS	i-P	i-R	i-D	i-C
CFG [18]	**21.92**	0.065	**0.271**	0.783	0.419	0.732	0.867
w/ our label augmentation	22.04	**0.064**	0.268	**0.795**	**0.426**	**0.773**	**0.880**

Table 8. Quantitative results on ImageNetLT [34].

Method	FID	LPIPS	intra-FID	intra-LPIPS	i-P	i-R	i-D	i-C
DiffAug-GAN	**48.19**	**0.6641**	371.1±29.20	0.615±0.015	0.232	0.632	0.045	0.600
Softlabel-GAN	49.17	0.6601	**352.0**±29.05	**0.643**±0.011	**0.251**	**0.723**	**0.055**	**0.707**

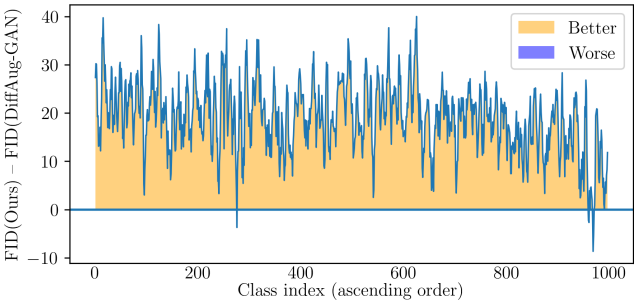


Figure 8. Differences between class FID scores of DiffAug-GAN and Softlabel-GAN. The large gains in minor classes and small gains in major (rightmost) classes show that our method effectively contributes to minor classes.

We count the generated images with a class different from a conditional class. First, we generate 50 images per class on AnimeFace and compute the color histogram of each generated image. The color histogram describes the number of pixels in each color range over all pixels of an image. Next, we calculate the average of the histogram correlations among images with the same class and use the inter-quartile range rules to detect images with a different class from a given condition. DiffAug-GAN, Smooth-GAN, and Softlabel-GAN generate 11, 52, and 12 images that differ from the given conditions, respectively. DiffAug-GAN and our method can generate the correct class, while Smooth-GAN fails. This is because Smooth-GAN blindly distributes probabilities without considering the given images. This observation clearly confirms the outperformance of our label augmentation against other methods. In the experiments with 256 resolution, we also observe the superiority of our Softlabel-GAN over the baselines as shown in Tab. 5. The experiments with an ADC-GAN [19] baseline (Tab. 6) show that our method works on another architecture. We also conduct the experiment with a Vision-aided GAN [29] baseline. Vision-aided GAN achieves FIDs of 15.77 on AnimeFace, 14.48 on Stanford Cars, and 26.88 on Oxford102, and Vision-aided GAN with our label augmentation achieves FIDs of 13.61 on AnimeFace, 11.01 on Stanford Cars, and 16.84 on Oxford102. We also report fine-grained metrics for the experiments in Supplementary Material. For the applicability of our method to another generative model, we integrate our label augmentation into diffusion models. We compare classifier-free guidance (CFG) [17, 18, 40] and CFG with our label augmentation on the imbalanced TinyImageNet dataset. In this experiment, we use intra-Kernel Inception Distance (KID) [3] instead of intra-FID due to the huge inference costs of diffusion models. In Tab. 7, our method shows that marginal improvements over a diffusion model with CFG in fine-grained metrics. Table 8 shows the experimental results on ImageNetLT [34]. Our method demonstrates the improvements in fine-grained metrics on the large-scale dataset. The comparison of per-class performance in Fig. 8 clearly shows our method provides performance gains in minor classes.

We discuss the impact of classification accuracy on generation quality. The accuracies of the pretrained classifiers in the experiments for AnimeFace, Oxford102, imbalanced Tiny ImageNet, and Cars are 81.2%, 99.9%, 84.4%, and 99.9%, respectively. An accuracy of less than 90% is not sophisticated but useful for training the cGANs. We further analyze the insensitivity of the classifier performance to the generation quality in Supplementary Material.

6. Conclusion

We investigated the problem of image generation in an imbalanced data regime. To prevent overlooking minor classes in conventional approaches, we introduced label augmentation to increase diversity while maintaining affinity. Furthermore, we proposed Softlabel-GAN by incorporating our label augmentation into the discriminator. Owing to the use of classifier predictions as a discriminator's class condition, Softlabel-GAN enables us to extract the features from other class samples, resulting in more focus on minor classes. Comprehensive benchmarking on imbalanced datasets shows that our method outperforms other methods and is less affected by the number of samples of each class. Our limitation is that our method outperforms conventional methods on only imbalanced datasets and does not outperform them on balanced datasets, as conventional methods perform well with sufficient training data.

Acknowledgements. This work was supported by the Institute for AI and Beyond of the University of Tokyo, the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), ROIS NII Open Collaborative Research 2023_23FC01, JSPS KAKENHI Grant Numbers JP22K17947, JP23KJ0381, JP23H03449, and JP22H05015.

References

[1] Animeface dataset. www.nurs.or.jp/~nagadomi/animeface-character-dataset/. 1, 3, 5

[2] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023. 1

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 8

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2018. 5, 6

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002. 3

[6] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NeurIPS*, pages 742–751, 2017. 3

[7] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised GANs via auxiliary rotation loss. In *CVPR*, pages 12146–12155, 2019. 2

[8] Anoop Cherian and Alan Sullivan. Sem-GAN: Semantically-consistent image-to-image translation. In *WACV*, pages 1797–1806, 2019. 3

[9] Jan Chorowski and Navdeep Jaitly. Towards better decoding and language model integration in sequence to sequence models. In *Proceedings of Interspeech*, pages 523–527, 2017. 2

[10] Kaiwen Cui, Yingchen Yu, Fangneng Zhan, Shengcai Liao, Shijian Lu, and Eric P Xing. KD-DLGAN: Data limited image generation via knowledge distillation. In *CVPR*, pages 3872–3882, 2023. 3, 6

[11] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 1

[13] Raphael Gontijo-Lopes, Sylvia Smullin, Ekin Dogus Cubuk, and Ethan Dyer. Tradeoffs in data augmentation: An empirical study. In *ICLR*, 2021. 1

[14] 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*, pages 6626–6637, 2017. 6

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[16] Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. Improved techniques for training single-image GANs. In *WACV*, pages 1300–1309, 2021. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020. 8

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

[19] Liang Hou, Qi Cao, Huawei Shen, Siyuan Pan, Xiaoshuang Li, and Xueqi Cheng. Conditional GANs with auxiliary discriminative classifier. In *ICML*, pages 8888–8902, 2022. 8

[20] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)*, pages 111–117, 2000. 1, 3

[21] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive D: adaptive pseudo augmentation for GAN training with limited data. In *NeurIPS*, pages 21655–21667, 2021. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 3

[23] HM Kabir, Moloud Abdar, Seyed Mohammad Jafar Jalali, Abbas Khosravi, Amir F Atiya, Saeid Nahavandi, and Dipti Srinivasan. Spinalnet: Deep neural network with gradual input. *arXiv preprint arXiv:2007.03347*, 2020. 5

[24] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 1

[25] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, pages 12104–12114, 2020. 1, 2, 5

[26] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama. OS-SGAN: Open-set semi-supervised image generation. In *CVPR*, pages 11185–11193, 2022. 2

[27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013. 3, 5, 6

[28] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. 3, 5

[29] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *CVPR*, pages 10651–10662, 2022. 8

[30] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, pages 3927—-3936, 2019. 6

[31] 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 *CVPR*, pages 4681–4690, 2017. 3

[32] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. GAN compression: Efficient architectures for interactive conditional GANs. *IEEE TPAMI*, 44(12):9331–9346, 2022. 3

[33] Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017. 3

[34] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019. 3, 8

[35] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 1

[36] Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *ICML*, pages 4183–4192, 2019. 2

[37] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 1, 3

[38] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *ICLR*, 2018. 4, 6

[39] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, pages 7176–7185, 2020. 6

[40] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 8

[41] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 3, 5

[42] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *CVPR*, pages 18434–18443, 2023. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[44] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R Venkatesh Babu. Improving GANs for long-tailed data through group spectral regularization. In *ECCV*, pages 426–442, 2022. 2

[45] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, pages 14866—-14876, 2019. 1

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, pages 2234–2242, 2016. 6

[47] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

[48] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *NeurIPS*, pages 17480–17492, 2021. 1

[49] Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2022. 1

[50] Artem Sevastopolskiy, Yury Malkov, Nikita Durasov, Luisa Verdoliva, and Matthias Nießner. How to boost face recognition with stylegan? In *ICCV*, pages 20924–20934, 2023. 1

[51] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. SinGAN: Learning a generative model from a single natural image. In *CVPR*, pages 4570–4580, 2019. 2

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[53] Jiaze Sun, Binod Bhattarai, and Tae-Kyun Kim. MatchGAN: a self-supervised semi-supervised conditional generative adversarial network. In *ACCV*, 2020. 2

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 1, 2, 3, 4, 5, 6

[55] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *NeurIPS*, pages 5523–5533, 2017. 3

[56] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for GAN training. *IEEE TIP*, 30:1882–1897, 2021. 1, 2

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 6000–6010, 2017. 2

[58] Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. PPCD-GAN: Progressive pruning and class-aware distillation for large-scale conditional GANs compression. In *WACV*, pages 2436–2444, 2022. 3

[59] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny ImageNet challenge. 3, 5

[60] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE TIP*, 30:5984–5996, 2021. 2

[61] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1, 2, 4, 5

[62] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, pages 7354–7363, 2019. 1

[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[64] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In *NeurIPS*, pages 7559–7570, 2020. 1, 2, 4, 5, 6

[65] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for GAN training. *arXiv preprint arXiv:2006.02595*, 2020. 1, 2