SYNAuG: Exploiting Synthetic Data for Data Imbalance Problems

Moon Ye-Bin [* 1] **Nam Hyeon-Woo** [* 1] **Wonseok Choi** [2] **Nayeong Kim** [3] **Suha Kwak** [3 2] **Tae-Hyun Oh** [1 2 3]

Abstract

Data imbalance in training data often leads to biased predictions from trained models, which in turn causes ethical and social issues. A straightforward solution is to carefully curate training data, but given the enormous scale of modern neural networks, this is prohibitively labor-intensive and thus impractical. Inspired by recent developments in generative models, this paper explores the potential of synthetic data to address the data imbalance problem. To be specific, our method, dubbed SYNAuG, leverages synthetic data to equalize the unbalanced distribution of training data. Our experiments demonstrate that, although a domain gap between real and synthetic data exists, training with SYNAuG followed by fine-tuning with a few real samples allows to achieve impressive performance on diverse tasks with different data imbalance issues, surpassing existing task-specific methods for the same purpose.

1. Introduction

Modern machine learning resides in the over-parameterization regime (Neyshabur et al., 2019; Allen-Zhu et al., 2019). Deep Neural Networks (DNNs) overfit the given abundant and diverse labeled data and show outstanding performance on the test data (Deng et al., 2009; Lin et al., 2014). However, DNNs are vulnerable to biased features and distribution, *e.g.*, shortcut learning (Geirhos et al., 2020; Hermann & Lampinen, 2020; Jacobsen et al., 2019), which is caused by the data distribution. Prior works have focused on the model architecture and algorithms in general (Sambasivan et al., 2021) to tackle this problem.

While the advancement of architectures and algorithms is undeniably crucial, the fundamental solution lies in prioritizing the quality and quantity of data. In long-tail recog-

nition (Cui et al., 2019), we can enhance the accuracy of classes having a few number of data points by collecting more data for those classes. It is important to acknowledge that collecting such data demands substantial resources in terms of human effort, time, and financial investment.

It leads us to ask the question, **"Can recent diffusion models be leveraged to address the data challenge?"** The success of generative diffusion models (Rombach et al., 2022) has been remarkable. These models demonstrate their ability to generate highly realistic images. Moreover, text-to-image diffusion models offer remarkable controllability, enabling users to generate images with short sentences. For example, we can generate images of waterbirds on the water or land by changing the prompt to contain "on the water" or "on the land." Likewise, diffusion models could be a plausible resource for the quality and quantity of data, which is the fundamental solution to data-caused issues.

In this work, our objective is to investigate the efficacy of synthetic data in addressing data-related challenges. For these experiments, we propose SYNAuG, the pipeline to augment the data and train the models. Initially, we address the class imbalance by leveraging text-to-image diffusion models to ensure uniformity in the number of samples across all classes. Subsequently, we apply off-the-shelf task-specific algorithms to train the classifier. Following training on this combined dataset comprising both original and synthetic data, we further optimize the last layer using uniformly sub-sampled original data.

Our investigation into the proposed SYNAuG method reveals several key findings. Firstly, while synthetic data generated by recent generative models can partially substitute real samples, a few real samples are crucial. We observe varying degrees of performance degradation when real samples are replaced with synthetic ones, highlighting the incompleteness of synthetic data as a perfect substitute for real-world data. Secondly, we identify the importance of mitigating the domain gap between real and synthetic data. To address this issue, we propose several methods, including employing data augmentation techniques across real and synthetic data, and implementing fine-tuning strategies. Lastly, despite the aforementioned concerns, SYNAuG addresses data imbalance issues such as long-tailed distributions, model fairness, and robustness to spurious correlation,

---

[*]Equal contribution [1]Dept. of Electrical Engineering, POSTECH, Pohang, South Korea [2]Grad. School of Artificial Intelligence, POSTECH, Pohang, South Korea [3]Dept. of Computer Science and Engineering, POSTECH, Pohang, South Korea. Correspondence to: Tae-Hyun Oh <taehyun@postech.ac.kr>.
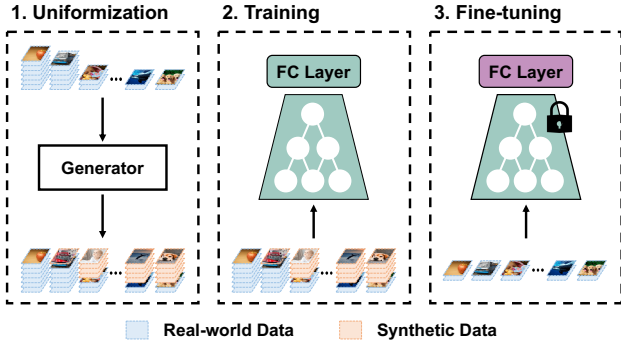
Figure 1: **Overview of SYNAuG process.** Given the imbalanced real-world data with the class labels, we first uniformize the imbalanced real data distribution by generating the synthetic samples that are conditioned on the class label. Second, we train a model with the uniformized training data. Finally, we fine-tune the last layer with the uniformly subsampled real-world data.

showing potential for improving performance. This offers valuable insights into the application of generative models for addressing real-world challenges.

2. Background

Data imbalance can lead to suboptimal generalization and many challenges in practical application scenarios, *e.g.*, finance, healthcare, and autonomous driving. The data imbalance problem is a common source of different imbalance sub-problems: long-tailed recognition, model fairness, and model robustness to spurious correlation.

Long-tailed recognition. Long-tailed distribution is inherent to real-world (Cui et al., 2019; Zhang et al., 2023). There are prior work in the realm of re-balancing classes, including re-sampling (Shen et al., 2016; Park et al., 2022) and re-weighting (Samuel & Chechik, 2021; Ren et al., 2020; Cui et al., 2019), designing loss functions (Ryou et al., 2019; Lin et al., 2017; Ren et al., 2020), model selection (Wang et al., 2021), and language-guided feature augmentation (Ye-Bin et al., 2023).

Model fairness. In fairness (Narayanan, 2018; Hardt et al., 2016; Feldman et al., 2015), researchers have tackled the issue of model bias, where accuracy varies based on sensitive attributes such as race, age, and ethnicity. Model fairness is also related to data imbalance because the number of samples of some sensitive groups is lower than that of the major groups. Similar to long-tailed recognition, fairness has predominantly been tackled using loss weighting (Jung et al., 2023) and batch sampling (Kamiran & Calders, 2012; Roh et al., 2021).

Spurious correlation. The spurious correlation problem is related to the robustness of models against misleading correlations. DNNs are susceptible to falling into short-cuts that capture the most frequently observed patterns in a class regardless of true causality; it is called spurious correlation or shortcut problems (Geirhos et al., 2020; Kirichenko et al., 2023). The spurious correlation problem is also dealt with similar approaches to the above two tasks: reweighting (Sagawa et al., 2020a; Kim et al., 2022), sampling (Idrissi et al., 2022; Sagawa et al., 2020b), and post-calibration (Liu et al., 2021; Kirichenko et al., 2023).

Summary of data imbalance problems. While researchers have developed algorithms for each task separately, three different tasks primarily arising from data imbalance have been tackled in the shared perspective, *i.e.*, up-weight loss values or sampling probabilities of minor groups using group or sensitive information. However, most have focused more on algorithmic parts; thus, the inherent imbalance may still remain. In this work, we shed light on the overlooked convention to go beyond the given bounded dataset. We exploit the synthetic data from the generative foundation models (Rombach et al., 2022; Saharia et al., 2022; Nichol et al., 2022) to take flexibility and controllability.

Using synthetic data in machine learning tasks. Synthetic data is beneficial to overcome the lack of data and sensitive issues of data, *e.g.*, licensing and privacy concerns. Several approaches (Jahanian et al., 2022; Oh et al., 2018; Zhang et al., 2021) have started to leverage synthetic data for their tasks of interest, but do not consider imbalance problems. Recently, deep generative models (Rombach et al., 2022; Saharia et al., 2022; Nichol et al., 2022) have shown promising results in generating realistic and high-quality samples, stemming from the goal of modeling the real data distribution. In this work, we explore the use of a pre-trained foundation diffusion model (Rombach et al., 2022) to observe whether synthetic data is effective in mitigating data imbalance problems.

3. Exploiting Synthetic Data

We propose SYNAuG, which leverages synthetic data to mitigate data imbalance problems from a data perspective. As illustrated in Fig. 1, we first uniformize the imbalanced data by generating synthetic data, then train the model on the uniformized data, and finally fine-tune the last layer with a few original data uniformly subsampled from each class. We exploit a recent powerful generative model, Stable Diffusion (Rombach et al., 2022), to generate synthetic data of corresponding classes or attributes with the controllable prompt. Since they are trained on a large number of web data, it would be considered to cover and model a wide distribution of the real world. Exploiting these favorable properties, we generate supporting data to alleviate the im-

Algorithm 1 SYNAuG

Require: Real data $\mathcal{R}$, Generator G, image encoder f and linear classifier g parameterized with $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$, and training algorithms $\texttt{Alg}^M(\cdot)$ and $\texttt{Alg}(\cdot)$ with and without Mixup, respectively.
  1. Training data: $\mathcal{D} = \texttt{Uniformization}(\mathcal{R}, \mathcal{S})$, where $\mathcal{S}$ is synthetic data generated from G.
  2. Training f and g : $f, g = \texttt{Alg}^M(f, g, \mathcal{D})$ ▷ Both $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$ are updated
  3. Fine-tuning g : $g = \texttt{Alg}(g, \mathcal{R})$ ▷ Last layer is fine-tuned, *i.e.*, only $\boldsymbol{\theta}_g$ is updated, with the subset of real data, which of distribution is uniform.

---

balance of the data distribution.

While uniformizing with synthetic data is simple and effective, there is still room to improve its performance because of the domain gap. To improve further by mitigating the domain gap, we propose utilizing two simple methods. First, we propose to leverage Mixup (Zhang et al., 2018) during training to augment samples to be interpolated samples between real and synthetic samples, *i.e.*, domain Mixup. Second, we propose to fine-tune a classifier on the subsampled uniform original data from the original training data after the first training stage. The fine-tuned classifier would lead to more accurate recognition of target data. Algorithm 1 is a pseudo code of SYNAuG.

Details of image generation and uniformization. We can produce high-quality synthetic images across various classes by employing Stable Diffusion (Rombach et al., 2022). We can also achieve data controllability based on a text prompt containing a class name to represent class semantics. The synthetic data is utilized to populate and uniformize each training dataset, including CIFAR100-LT (Cao et al., 2019), ImageNet100-LT (Jiang et al., 2021), UTKFace (Zhang et al., 2017), and Waterbirds (Sagawa et al., 2020a).

To generate synthetic images for CIFAR100-LT, we use the prompt "a photo of {modifier} {class}," leveraging a set of 20 modifiers recommended by ChatGPT (Ouyang et al., 2022) for each class. ImageNet100-LT has 100 classes, and the image resolution is 256×256, ranging from a maximum of 1,280 to a minimum of 6 images per class. However, merely 3 classes possess over 500 samples per class. Consequently, we have added synthetic data to ensure 500 samples for the remaining classes. We input the prompt "a photo of {class}, ..." during the generation process.

UTKFace is composed of 23,708 images with age, gender, and race labels. Age is from 0 to 116; gender is male or female; races are White, Black, Asian, Indian, and others. We use race annotation as group labels, serving as sensitive attributes, except for others. Additionally, gender annotation functions as class labels. For generating synthetic data,

we provide a prompt incorporating race, gender, and age, structured as "Portrait face photo of {race} {gender}, {age} year old ...".

Waterbirds is synthesized using the CUB dataset (Welinder et al., 2010) and the backgrounds. The seabirds and waterfowl are grouped as waterbirds, while the other bird types are categorized as landbirds. For generating synthetic data, we provide the prompt composed of water or land, such as "The photo of water bird in land ...".

4. Experiments

We evaluate our method for three sub-tasks: long-tailed recognition, model fairness, and model robustness to spurious correlation. We provide key findings and the effectiveness of SYNAuG for data imbalance problems.

4.1. Experimental Settings

Long-tailed recognition. We employ two long-tail datasets: CIFAR100-LT (Cao et al., 2019) and ImageNet100-LT (Jiang et al., 2021). The test sets for them are the same as the original one. The classes in the long-tailed datasets are divided into three groups: Many-shot (more than 100 samples), Medium-shot (20-100 samples), and Few-shot (less than 20 samples). The imbalance factor (IF) corresponds to the skewness of the training data; high IF means high skewness and a more challenging setting. We evaluate under the standard IFs of 100, 50, and 10, following the prior work (Alshammari et al., 2022). We use ResNet32 for CIFAR100-LT and ResNet50 for ImageNet100-LT.

Model fairness. We employ UTKFace (Zhang et al., 2017) composed of 23,708 images with age, gender, and race labels. We use race annotation as a sensitive attribute (group label) and gender as the class label. We use fairness metrics, which have been proposed to measure the fairness performance of models: Demographic Parity (DP) (Feldman et al., 2015), Equal Opportunity (EO) (Hardt et al., 2016; Jung et al., 2022), and Equalized Odds (ED) (Hardt et al., 2016). These metrics are based on the difference in the performance of the learned classifiers depending on groups, *i.e.*, the sensitive attributes. Lower values of fairness metrics indicate that the model is fairer.

Model robustness to spurious correlation. We use the Waterbirds dataset (Sagawa et al., 2020a), which is a synthetic dataset created by combining images of birds from the CUB dataset (Welinder et al., 2010) with backgrounds. The birds are grouped into two categories: waterbirds, which include seabirds and waterfowl, and landbirds. The Land and water background are spuriously correlated to landbirds and waterbirds, respectively. We report the result over 5 independent runs using the code from DFR (Kirichenko et al., 2023). We reproduce the BaseModel and DFR and

(a) Class-wise replacement
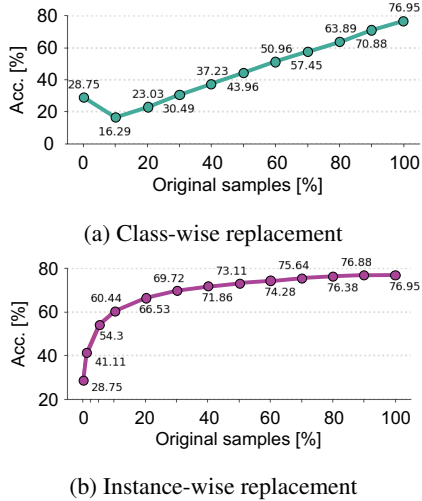

(b) Instance-wise replacement

Figure 2: **Replacement test.** To investigate the effect on model performance when using original and synthetic data together, we replace the original data with synthetic ones in two ways: (a) class-wise and (b) the same ratio of instances across all classes. We use CIFAR100, which has 500 samples per class and 100 classes.

report the performance.[1]

4.2. Experimental Results

Is synthetic data effective? Yes but not equal to real data. In Table 2, we check the effectiveness of synthetic data compared to real data by replacing some of the original data with synthetic data in two ways of settings. The original dataset has the same amount of data in each class. In the **first setting**, we completely replace original data belonging to specific classes with synthetic data. Note that certain classes have no real sample but only synthetic samples. In the **second setting**, we uniformly replace the original data with synthetic data, which means all classes have the same amount of original and synthetic data.

To isolate the effects of synthetic data, we do not apply Mixup and last-layer fine-tuning. The results reveal contrasting patterns between the two settings. In the first setting, we observe a linear decline in performance as the number of classes lacking original data increases (See Fig. 2a). Conversely, the second setting shows a logarithmic decrease in performance as more original data are uniformly replaced with synthetic data (See Fig. 2b). Remarkably, using only 1% of real data in the second setting yields a performance of 41.11%, which is comparable to the 43.96% achieved using 50% real data in the first setting. It suggests that synthetic data help improve performance even with a small amount

---

[1] https://github.com/PolinaKirichenko/deep_feature_reweighting

---

(a) Binary domain classification
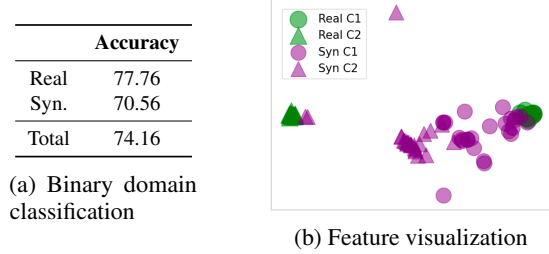

(b) Feature visualization

Figure 3: **Domain gap between real and synthetic data.** We test domain gap empirically with (a) binary domain classification and (b) feature visualization. For classification, we use 2.5k samples for each real and synthetic domain and train only one fully-connected layer on the features extracted from pre-trained model. For visualization, the features are extracted from the pre-trained model on CIFAR100. C1 and C2 denote different classes.

Table 1: **Method to reduce domain gap.** We use CIFAR100-LT. Each component, Modifier, Mixup, Re-train, and Finetune, means we use the class-related modifiers in the prompt, use Mixup augmentation during training, and re-train or finetune the last layer after training, respectively. (e) stands for our SYNAuG.

	Modifier	Mixup	Re-train	Finetune	IF		
					100	50	10
(a)					52.41	56.99	66.34
(b)	✓				53.54	57.09	66.66
(c)	✓	✓			55.45	58.69	66.84
(d)	✓	✓	✓		57.31	60.34	67.90
(e)	✓	✓		✓	**58.59**	**61.36**	**69.01**

of original data, but still including at least a small amount of original data is crucial. It may imply a complementary domain gap between synthetic and real data.

Is there a domain gap between synthetic and real data? Yes. The above experiments also show the domain gap may still exist even with high-quality synthetic data. To check the presence of a domain gap, we conduct domain classification and visualization of the features from both real and synthetic data (See Fig. 3). As shown in Fig. 3a, the domain classification performance is 74.16%. Considering that 50% means no domain gap, the result implies that there is an apparent domain difference classifiable. As shown in Fig. 3b, the features of Syn C2 are closer to Syn C1 than Real C2. This observation provides empirical evidence of a domain gap existing between real and synthetic data.

Do we need to reduce the domain gap? Yes. We propose several methods to reduce the domain gap between synthetic and real data: Modifiers generate diverse synthetic samples, Mixup interpolates between synthetic and real samples, whereby the domain gap is mitigated by bridging two

Method	# step	Many	Medium	Few	All
CE CVPR'19		**61.85**	15.83	0.29	32.06
SYNAuG	3	48.23	46.29	40.07	45.10
	10	53.89	**49.49**	43.87	**49.34**
	50	52.91	48.63	**45.27**	49.12
	100	53.03	49.20	44.47	49.12
	300	54.11	47.71	44.73	49.06

Figure 4: **Ablation study according to sample quality. (Top)** quality of the generated samples according to the number of steps, **(Bottom)** long-tailed recognition performance (%) according to the different times of steps for generating synthetic data, which affects sample quality. We use ImageNet100-LT with ResNet50.

different domain data, and we fine-tune the last layer to get an appropriate hyperplane for real samples. Table 1 shows the effectiveness of our methods. When we use modifiers in the prompt, we observe the performance gain (See (a) and (b) in Table 1). We can achieve further improvement by utilizing Mixup (See (c) in Table 1). As shown in Table 1-(d,e), we can achieve an additional improvement by adjusting the classifier towards the targeted real data and found that fine-tuning is more effective than re-training.

Is more realistic synthetic data more effective? Not really. In Fig. 4, we evaluate SYNAuG on ImageNet100-LT. We conduct an ablation study to investigate the impact of data quality of SYNAuG by controlling the diffusion step parameter of Stable Diffusion (Rombach et al., 2022), which is known to affect the quality of generated images. As shown in Fig. 4-(Top), generation quality is low with a minimal number of steps, but visually, the differences become less noticeable as the step count increases to a certain level. Figure 4-(Bottom) shows the quantitative results. Compared to the CE method trained on the original long-tailed data, while the accuracy of the Many class is degraded, we achieve a large improvement in the Medium, Few, and even All cases regardless of the synthetic image quality. However, there is a certain level of quality that exhibits a surge point in performance. The quality difference becomes negligible when the step value exceeds a certain threshold. This may hint that there exists a favorable gap that can complement real data.

Does synthetic data help with data imbalance problems? Indeed. For the long-tailed recognition problem, there are significant improvements when we use synthetic data, as

Table 2: **Long-tailed recognition performance on CIFAR100-LT.** We compare our SYNAuG with recent works. We report the Top-1 accuracy (%) with different imbalance factors, *i.e.*, IF={100, 50, 10}.

Method	IF=100				50	10
	Many	Medium	Few	All		
CE CVPR'19	68.31	36.88	4.87	37.96	43.54	59.50
SSD ICCV'21	-	-	-	46.0	50.5	62.3
PaCo ICCV'21	-	-	-	52.0	56.0	64.2
RISDA AAAI'21	-	-	-	50.16	53.84	62.38
CE + CMO CVPR'22	70.4	42.5	14.4	43.9	48.3	59.5
LDAM + CMO CVPR'22	61.5	48.6	28.8	47.2	51.7	58.4
RIDE (3 experts) + CMO CVPR'22	-	-	-	50.0	53.0	60.2
Weight Balancing CVPR'22	72.60	51.86	32.63	53.35	57.71	68.67
SYNAuG	**74.06**	**56.63**	**42.83**	**58.59**	**61.36**	**69.01**

shown in Table 2. When we compare the accuracy between CE and SYNAuG, performance is improved regardless of the skewness of the training data. CE stands for the method trained with the Cross-Entropy loss on the original data. The model performance is also improved on the fairness problem in Table 3a and the spurious correlation problem in Table 4. Our results demonstrate that synthetic data is helpful for data imbalance problems.

Is synthetic data more helpful than the state-of-the-art algorithmic methods? Surprisingly, yes. In Table 2, we compare with recent prior arts: SSD (Li et al., 2021) and PaCo (Cui et al., 2021) for self-supervised learning, RISDA (Chen et al., 2022) and CMO (Park et al., 2022) for data augmentation, and Weight Balancing (Alshammari et al., 2022) for the rebalance classifier.

In Table 2, our method outperforms most of the computing methods. This is a stunning result in that it suggests that relieving the imbalance from the data point of view is simple but more effective than the conventional complex algorithmic methods.

Is uniformization with synthetic data better than other data augmentation methods? In Table 5, compared to the popular image augmentation methods, Mixup (Zhang et al., 2018) and Cutmix (Yun et al., 2019), our performance is much higher. The results are consistent with the one naïvely applying Mixup to imbalanced data cases is known to be detrimental (Ye-Bin et al., 2023); thus, we distinctively apply Mixup after uniformizing data distribution, which makes a noticeable difference.

In Table 3c, we evaluate the effect of data augmentations when we apply Mixup and Cutmix after uniformization. In this ablation study, we do not apply fine-tuning for clear comparison. Both augmentations improve the accuracy of ERM; Mixup also works in the fairness metrics. Compared to ERM with Mixup, SYNAuG shows higher accuracy and better fairness metrics. SYNAuG with Mixup outperforms more in accuracy and fairness metrics compared to ERM

Table 3: **Fairness performance.** (a) accuracy (↑) and model fairness (↓), (b) compatibility with fairness algorithms, Group-DRO and Re-Sampling (RS), (c) ablation study with data augmentation, Mixup and Cuxmix, and (d) ablation study using the prior about sensitive attributes.

Method	Accuracy	DP	ED	EO
		ResNet18		
ERM	93.9	0.0817	0.0632	0.0779
SYNAuG	**94.1**	**0.0600**	**0.0462**	**0.0434**
		ResNet50		
ERM	94.0	0.07272	0.05850	0.07269
SYNAuG	**94.4**	**0.05432**	**0.03936**	**0.05472**

(a) **Fairness performance.**

Method	Accuracy	DP	ED	EO
ERM	93.9	0.08174	0.06320	0.07790
+ Group-DRO	93.9	0.07266	0.05819	0.06954
+ RS	93.8	**0.06360**	**0.04663**	**0.05808**
SYNAuG	**94.0**	0.06995	0.05630	0.06825
+ Group-DRO	93.9	0.07110	0.05433	0.07045
+ RS	93.6	0.06786	0.06439	0.08120

(b) **Ablation with other algorithms**

Method	Accuracy	DP	ED	EO
ERM	93.9	0.08174	0.06320	0.07790
+ Mixup	93.9	0.07266	0.05819	0.06954
+ CutMix	94.7	0.08265	0.06245	0.08333
SYNAuG	94.0	0.06995	0.05630	0.06825
+ Mixup	94.2	**0.06400**	**0.03793**	**0.04658**
+ CutMix	**94.9**	0.07430	0.04745	0.06319

(c) **Augmentation ablation**

Method	Accuracy	DP	ED	EO
SYNAuG*	**94.0**	0.07342	0.05805	0.07253
SYNAuG	**94.0**	**0.06995**	**0.05630**	**0.06825**

*Not use the sensitivity attribute

(d) **Sensitivity ablation**

Table 4: **Robustness to spurious correlation on Waterbirds.** SYNAuG outperforms DFR consistently in worst-group accuracy.

Method	Waterbirds	
	Worst	Mean
BaseModel	$73.7_{\pm 3.04}$	$90.4_{\pm 0.21}$
SYNAuG	$79.9_{\pm 2.22}$	$91.5_{\pm 0.98}$
$\text{DFR}^{\text{Tr}}_{\text{Val}}$ ICLR'23	$91.2_{\pm 1.92}$	$93.1_{\pm 0.91}$
SYNAuG (re-train)	$92.9_{\pm 0.45}$	$93.6_{\pm 0.35}$
SYNAuG (fine-tuning)	$93.2_{\pm 0.42}$	$94.8_{\pm 0.11}$

Table 5: **Comparison with the baselines.** We use CIFAR100-LT. The second column denotes the data type used in uniformization.

Method	Additional Data Type	IF		
		100	50	10
CE CVPR'19	N/A	37.96	43.54	59.50
Mixup ICLR'18	N/A	36.21	41.09	56.82
Cutmix ICCV'19	N/A	37.91	43.54	60.27
Intra-class Image Translation	Syn.	47.87	53.33	64.95
Inter-class Image Translation	Syn.	47.17	51.33	64.11
Class Distribution Fitting	Syn.	51.53	55.60	65.60
Web crawled images	Real	54.06	56.40	63.86
SYNAuG	Syn.	**58.59**	**61.36**	**69.01**

the diffusion model with the samples in each class to model the class-wise distribution, named *Class Distribution Fitting*, and 4) As a strong baseline, we collect the real data from the internet instead of generating synthetic images, *i.e.*, *Web crawled images*.

In Table 5, we compare our method with proposed baselines. Compared to the case that uses real-world web data, it shows that the generated images are of sufficient quality to mitigate the class imbalance problem. Also, we evaluate additional baselines, which apply the variant methods during the generation process. While they are better than training only with the original long-tailed data (CE method), the performance is lower than SYNAuG. While simple, SYNAuG is noticeably effective.

Is SYNAuG compatible with other algorithms to make a synergy? Yes! In Table 3b, we evaluate whether our method can be compatible with two algorithms, Group-DRO (Sagawa et al., 2020a) and Re-Sampling (RS). Note that we do not apply Mixup and fine-tuning in this experiment. Group-DRO and RS improve the fairness metrics of ERM at the same time. While Group-DRO and RS improve SYNAuG in some fairness metrics, developing a fairness algorithm with synthetic data might be a promising direction toward a fair model.

In Table 4, SYNAuG generates samples not to be correlated with spurious features, which improves the performance in BaseModel both on worst and mean accuracies.

with Mixup. Although Mixup or Cutmix is applied regardless of the data type, real or synthetic, it allows domain gap reduction and performance improvement.

How good is SYNAuG for generating samples? Much better than web crawled images. We present other variants of generation methods as baselines: 1) Motivated by the recent work (He et al., 2023) using the few-shot original samples as guidance during the generation process, we first introduce *Intra-class Image Translation*, where we use the original samples from the original training data as a class-wise guidance image for generation, 2) Inspired by the M2m (Kim et al., 2020) translating an image of the major class to the minor class for leveraging the diversity of the majority information, we introduce *Inter-class Image Translation*, where we use random samples in the dataset as guidance regardless of the class, 3) As an advanced version motivated by DreamBooth (Ruiz et al., 2023), we fine-tune
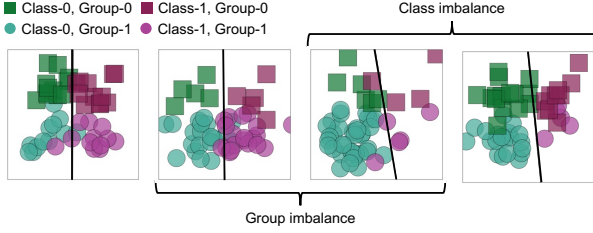
Figure 5: **Influence of the class and group imbalance on classifier.** The 2D data are sampled from the normal distributions with four different means and the same covariance. We simulate 4 different experiments with the latent group imbalance (sensitive attributes) by adjusting the number of data in each group. We train classifiers for the classes and visualize the learned classifiers (bold black lines). The fairer the classifiers, the more vertically aligned. The classifier trained on the class imbalance is more unfair than the one on the group imbalance.

When applying DFR, the synthetic data can increase the worst and mean accuracy consistently. We also observe that fine-tuning is more effective compared to re-train, which is consistent with Table 1.

Can we improve fairness performance even when we do not know about sensitive attributes? Favorably, yes. In Fig. 5, we empirically observe that the group imbalance with the class imbalance amplifies unfair classifiers. While both class and group imbalance contribute to the unfair classifiers, class imbalance contributes more than group imbalance.

In Table 3d, we compare the performance between original SYNAuG and variant SYNAuG*. For SYNAuG, we augment the data to mitigate the class imbalance across the sensitive attribute; the female and male ratio of each sensitive attribute becomes equal. For SYNAuG*, we augment the synthetic data to mitigate the class imbalance regardless of sensitive attributes. Although exploiting the knowledge of sensitive attributes is more effective, SYNAuG* shows better fairness metrics compared to ERM.

5. Conclusion

We propose SYNAuG to investigate the effectiveness of synthetic data in data imbalance problems. The key findings are the importance of at least a few real-world samples and the existence of a domain gap for synthetic images from a recent diffusion model. Despite the above limitations of synthetic data, SYNAuG is effective on data imbalance problems, including long-tailed recognition, model fairness, and robustness to spurious correlations. Our study suggests the importance of controlling imbalance from the data perspective. We believe that addressing data controllability is a promising research direction to resolve the early bottleneck

in machine learning model development. While we focus on the data perspective, improving the model in multiple views is necessary to solve data imbalance effectively.

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

Alshammari, S., Wang, Y.-X., Ramanan, D., and Kong, S. Long-tailed recognition via weight balancing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Chen, X., Zhou, Y., Wu, D., Zhang, W., Zhou, Y., Li, B., and Wang, W. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *KDD*, pp. 259–268, 2015.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., and Qi, X. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations (ICLR)*, 2023.

Hermann, K. and Lampinen, A. What shapes feature representations? exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351, 2022.

Jacobsen, J.-H., Behrmann, J., Zemel, R., and Bethge, M. Excessive invariance causes adversarial vulnerability. In *International Conference on Learning Representations*, 2019.

Jahanian, A., Puig, X., Tian, Y., and Isola, P. Generative models as a data source for multiview representation learning. *International Conference on Learning Representations (ICLR)*, 2022.

Jiang, Z., Chen, T., Mortazavi, B. J., and Wang, Z. Self-damaging contrastive learning. In *International Conference on Machine Learning (ICML)*, 2021.

Jung, S., Chun, S., and Moon, T. Learning fair classifiers with partially annotated group labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jung, S., Park, T., Chun, S., and Moon, T. Re-weighting based group fairness regularization via classwise robust optimization. In *International Conference on Learning Representations (ICLR)*, 2023.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33:1–33, 2012.

Kim, J., Jeong, J., and Shin, J. M2m: Imbalanced classification via major-to-minor translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Kim, N., Hwang, S., Ahn, S., Park, J., and Kwak, S. Learning debiased classifier with biased committee. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

Li, T., Wang, L., and Wu, G. Self supervision to distillation for long-tailed visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.

Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In *FAccT*, volume 1170, 2018.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *International Conference on Machine Learning (ICML)*, 2022.

Oh, T.-H., Jaroensri, R., Kim, C., Elgharib, M., Durand, F., Freeman, W. T., and Matusik, W. Learning-based video motion magnification. In *European Conference on Computer Vision (ECCV)*, 2018.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Park, S., Hong, Y., Heo, B., Yun, S., and Choi, J. Y. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. Fairbatch: Batch selection for model fairness. *International Conference on Learning Representations (ICLR)*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ryou, S., Jeong, S.-G., and Perona, P. Anchor loss: Modulating loss scale based on prediction difficulty. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations (ICLR)*, 2020a.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2020b.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

Samuel, D. and Chechik, G. Distributional robustness loss for long-tail learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Shen, L., Lin, Z., and Huang, Q. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016.

Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations (ICLR)*, 2021.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Ye-Bin, M., Kim, J., Kim, H., Son, K., and Oh, T.-H. Textmania: Enriching visual feature by text-driven manifold augmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.-F., Barriuso, A., Torralba, A., and Fidler, S. Datasetgan: Efficient labeled data factory with minimal human effort. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.