Impartial Adversarial Distillation: Addressing Biased Data-Free Knowledge Distillation via Adaptive Constrained Optimization

Donping Liao[1]*, Xitong Gao[2]*, Chengzhong Xu[1]†

[1] State Key Lab of IoTSC, Department of Computer and Information Science, University of Macau, Macau SAR, China
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
yb97428@um.edu.mo, xt.gao@siat.ac.cn, czxu@um.edu.mo

Abstract

Data-Free Knowledge Distillation (DFKD) enables knowledge transfer from a pretrained teacher to a light-weighted student without original training data. Existing works are limited by a strong assumption that samples used to pretrain the teacher model are balanced, which is, however, unrealistic for many real-world tasks. In this work, we investigated a pragmatic yet under-explored problem: *how to perform DFKD from a teacher model pretrained from imbalanced data.* We observe a seemingly counter-intuitive phenomenon, *i.e.*, adversarial DFKD algorithms favour minority classes, while causing a disastrous impact on majority classes. We theoretically prove that a biased teacher could cause severe disparity on different groups of synthetic data in adversarial distillation, which further exacerbates the mode collapse of a generator and consequently degenerates the overall accuracy of a distilled student model. To tackle this problem, we propose a class-adaptive regularization method, aiming to encourage impartial representation learning of a generator among different classes under a constrained learning formulation. We devise a primal-dual algorithm to solve the target optimization problem. Through extensive experiments, we show that our method mitigates the biased learning of majority classes in DFKD and improves the overall performance compared with baselines. Code will be available at https://github.com/ldpbuaa/ipad.

Introduction

Large pretrained neural networks achieve unprecedented success in many modern deep learning applications (He et al. 2016; Kenton and Toutanova 2019). Despite the ground-breaking performance, their over-parameterized and compute-intensive nature hinder the application on resource-constrained devices. This prompts the rapid advancement of research on compressing a large-sized teacher to an efficient student (Hinton et al. 2015; Zhao et al. 2019; Gao et al. 2018; Wang et al. 2019). However, in many scenarios, granting access to original training data is usually infeasible due to privacy concerns (Taigman et al. 2014) or intellectual property protection (Wu et al. 2016). To address this limitation, data-free knowledge distillation (DFKD) (Lopes, Fenu, and Starner 2017; Micaelli and Storkey 2019; Chen et al. 2019; Fang et al. 2019) has emerged as an effective tool to transfer knowledge without raw training data. An influx of new methods was proposed to address diversity (Fang et al. 2021b; Han et al. 2021), fidelity (Yin et al. 2020) and efficiency (Fang et al. 2022), among which adversarial DFKD (Micaelli and Storkey 2019; Choi et al. 2020; Fang et al. 2021b) has attracted particular interest. This approach can not only enhance the performance of student by generating boundary-aware samples (Micaelli and Storkey 2019), but it also improves the distillation efficiency compared with class-wise data synthesis methods (Luo et al. 2020).

Nevertheless, existing works generally rely on a strong assumption that the teacher is learned from a deliberately or artificially balanced dataset. This assumption may be very limited and unrealistic, as pretraining data in realistic scenarios may have unknown marginal probabilities among classes. For instance, a practical consideration is that large-scale real-world data is likely imbalanced due to the difficulty to cover rare species, scarce scenarios and uncommon events (He and Garcia 2009). Models learned on such data are potentially biased, even with the help of re-balancing techniques (Cui et al. 2019). Therefore, a pressing and practical problem that has been largely overlooked arises: *How can we transfer knowledge from a biased teacher without access to the original training data?*

In this paper, we push the frontier of DFKD by extending its applicability to this interesting but challenging setting. This problem presents an opportunity to harness the power of DFKD in class-imbalanced learning, which has recently gained a surge of research interest (Zhang et al. 2021). Yet from the optimization perspective, our work reveals that dataset bias can have an unexpected and detrimental impact on DFKD. As evinced by our empirical and theoretical studies, due to the interplay between a biased teacher and an adversarially-optimized generator, there can be a significant performance disparity among *different classes* for the distilled student. Moreover, the goal of privacy protection in DFKD means that the original data distribution is not available and may expose sensitive information about data creators. Therefore, previous techniques used for data re-balancing that depend on such information are not directly applicable.

To overcome these challenges, we propose impartial

*These authors contributed equally to this work.
†Corresponding author.

adversarial distillation (IPAD) to handle this new problem while accounting for both biased learning and privacy preservation in DFKD. Briefly, we introduce a constraint term to the vanilla DFKD formulation so that the discrepancy between the teacher and student of each class does not excessively surpass the average value. The constraint forces the generator to pay attention to majority classes that are easier to collapse rather than simply synthesizing adversarial examples to fool the student as in standard adversarial DFKD. Our method has the following advantages: first, it does not require the original data distribution as priors to reweigh samples and therefore less likely to disclose private information in original training data; second, the constrained optimization problem can be efficiently solved by a primal-dual method with minuscule overhead while exhibiting simple interpretation. Our contributions are as follows:

- **Problem.** To our knowledge, we initiate the first analysis on adversarially distilling knowledge from a *biased* teacher in a data-free manner. Empirically, we observe and interpret a new phenomenon that causes the prior adversarial DFKD method to suffer from this setting.
- **Analysis.** We provide a principled theoretical analysis to unravel underlying reasons. Our analysis implies that under biased adversarial DFKD, the synthetic samples predicted as minority class provably yield a larger discrepancy between the teacher and student, which corroborates our empirical observations.
- **Algorithm.** We propose to adaptively impose constraints on class-wise KL divergence. We solve this constrained optimization problem with a primal-dual updating scheme without requiring additional information beyond the standard DFKD objectives.
- **Performance.** From extensive experiments, we show that our method can address the mode collapse problem of majority classes and improve the overall distillation accuracy.

Related Work

Data-free Knowledge Distillation

Data-free knowledge distillation aims to transfer knowledge from a pretrained teacher to a compact student network without requiring raw training data. Lopes *et al.*(Lopes, Fenu, and Starner 2017) proposed the first work to leverage activation statistics to reconstruct pseudo training data. The central aspect of DFKD is determining how to generate meaningful pseudo samples that can facilitate successful knowledge transfer. To achieve this goal, existing works explored non-adversarial (Lopes, Fenu, and Starner 2017; Chen et al. 2019; Luo et al. 2020; Nayak et al. 2019; Wang 2021; Yoo et al. 2019) or adversarial methods (Micaelli and Storkey 2019; Choi et al. 2020; Fang et al. 2021b; Han et al. 2021; Do et al. 2022). Non-adversarial methods seek to devise different objectives to synthesize data to approximate the original data or feature distribution. ZSKD (Nayak et al. 2019) models the softmax outputs as Dirichlet distribution and construct a class similarity matrix as the surrogate of concentration parameters to capture the data prior. DAFL (Chen et al. 2019) regards the pretrained teacher as a discriminator and proposed a one-host loss to encourage the generator to synthesize high-confident samples. KegNet (Yoo et al. 2019) leverages a decoder to further regularize the reconstruction error of input variable to learn meaningful latent representation for conditional generator. Adversarial DFKD methods reformulate the knowledge transfer process as an adversarial learning problem to exploit boundary-aware synthetic samples. Micaelli *et al.* proposed adversarial belief matching (ABM) (Micaelli and Storkey 2019), a concept that has spurred additional research aimed at further enhancing diversity (Fang et al. 2021b; Han et al. 2021) and aligning feature map statistics (Yin et al. 2020). However, these works primarily rely on artificially-balanced datasets, while our work takes a step further by relaxing such limitation and substantially broadening the practicality of DFKD.

Prior research has revealed a potential issue with adversarial learning in the DFKD setting, specifically that it could result in the generator experiencing catastrophic forgetting. DFKD-Mem (Binici et al. 2022b) utilizes a memory bank to replay previously generated samples to help the student recall past knowledge. To further reduce memory footprint, Pre-DFKD (Binici et al. 2022a) employs a variational autoencoder (Kingma and Welling 2013) to model the distribution of synthetic samples. MAD (Do et al. 2022) introduces an additional generator with momentum updating to serve as a proxy of the temporal ensemble of old versions. SpaceshipNet (Yu et al. 2023) proposes channel-wise feature exchange and spatial activation region regularization to improve the feature diversity of generator and representation consistency between teacher and student. Patel *et al.* (Patel, Mopuri, and Qiu 2023) address the forgetting problem by aligning the gradient update of knowledge acquisition and knowledge retention objectives inspired by model-agnostic meta-learning (Finn, Abbeel, and Levine 2017). These studies demonstrate that the distribution shift issue (Taori et al. 2020) of synthesized samples could hinder knowledge transfer in DFKD. Distinct from these above works, this work reveals that label shift (Azizzadenesheli et al. 2018; Wu et al. 2021), which arises from a pre-trained teacher, can further exacerbate mode collapse in a generator.

Learning from Imbalanced Data

Real-world datasets often suffer from class imbalance, which becomes more prominent with scale. This issue results in long-tailed distributions and poses significant challenges for a wide range of tasks (Van Horn et al. 2018; Gupta, Dollar, and Girshick 2019). The problem has been previously addressed with resampling (Chawla et al. 2002) and reweighing (Cui et al. 2019) of the training data. Resampling involves constructing a virtually-balanced training dataset by oversampling the tail classes. Reweighing methods emphasize the tail classes by assigning them larger weights in a loss function. Recent works also provide new perspectives with decoupled training (Kang et al. 2019), data augmentation (Li et al. 2021), ensemble methods (Cai, Wang, and Hwang 2021) and causal inference (Tang, Huang, and Zhang 2020). While existing works focused on improving performance, they largely ignored how to transfer the

knowledge from the pretrained model to a light-weighted student for faster inference.

Beyond classification tasks, Rangwani *et al.* (Rangwani et al. 2022) explored the training of generative adversarial networks (GANs) on imbalanced data and uncovered that the tail classes exhibit class-specific mode collapse. Their results suggest that if the data imbalance issue arises in adversarial learning, it is important to address mode collapse of the generator. This issue becomes even more critical in data-free scenarios where the prior data distribution is unavailable, as studied in this work.

Preliminaries

Adversarial Data-free Knowledge Distillation

Let $T(\boldsymbol{x})$ denote a teacher pretrained on a dataset $\mathcal{D}_{\text{train}}$ and $S(\boldsymbol{x};\boldsymbol{\theta})$ be a student network with learnable parameters θ. Let $G(\boldsymbol{z};\boldsymbol{\phi})$ be a generator parameterized by weights $\boldsymbol{\phi}$ to yield pseudo training data with an input random noise vector $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. In DFKD, we aim to transfer the knowledge from T to S without original training data. Adversarial Belief Matching (ABM) (Micaelli and Storkey 2019) proposes to facilitate the learning process with synthetic data from a generator and formulates a min-max game:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{x}=G(\boldsymbol{z};\boldsymbol{\phi})}[\mathcal{L}_{\text{KD}}(T(\boldsymbol{x}), S(\boldsymbol{x};\boldsymbol{\theta}))], \tag{1}$$

where $\mathcal{L}_{\text{KD}}(\cdot)$ is the knowledge distillation loss to measure the prediction discrepancy between T and S. $\mathcal{L}_{\text{KD}}(\cdot)$ is usually instantiated by Kullback-Leibler (KL) divergence (Micaelli and Storkey 2019). Optimizing Equation (1) in an alternating way encourages G to create samples that maximize the discrepancy between the teacher and student's predictions, while the student is optimized to absorb the teacher's knowledge by imitating its predictions on synthetic data.

Adversarial DFKD from a Biased Teacher

Precursory DFKD methods assume the teacher is learnt from a balanced dataset. We extend this problem under imbalanced pretraining data, which we refer to as *biased* DFKD. As preliminary investigation, we use a WRN40-2 $\rightarrow$ WRN16-2 (Zagoruyko and Komodakis 2016) configuration, where the teacher is pretrained on CIFAR-100 with imbalance ratio $r = 0.01$. We adopt a simple two-stage training strategy from (Kang et al. 2019). More exploration of other rebalancing strategies is presented in Appendix C. We first train the teacher with Cross-Entropy loss and then fix the backbone network and fine-tune the classification head with class-balanced data sampling. Afterwards, we adopt ABM to perform DFKD and report the accuracy values in Table 1. We present a more detailed illustration of the class-wise accuracy differences between the teacher and the distilled student model in Figure 1. It turns out that adversarial DFKD surprisingly delivers accuracy improvements on the few-shot (minority) classes while causing a disastrous effect on the many-shot (majority) ones. This conflicts with the common intuition that compressing a pretrained neural network to a smaller size has an adverse impact on the minority classes (Tran et al. 2022). Since minority classes are under-represented in a pretrained network due to limited training data, they are generally considered hard examples. Recent studies also show that when the capacity of a neural network is reduced, samples from minority classes tend to be misclassified or forgotten either in supervised (Tran et al. 2022) or unsupervised learning (Jiang et al. 2021). The anomalous phenomenon triggers our exploration of an intriguing question: *why does adversarial DFKD favour minority classes while harming majority classes?*

	Many	Medium	Few	Overall
Pretrained teacher	62.0	58.1	41.5	53.9
DFKD student	53.8	53.5	42.0	49.8
ΔAccuracy	-8.2	-4.6	+0.5	-4.1

Table 1: An overview of accuracy and change in accuracy (ΔAccuracy) values (%) for different groups of data after adversarial DFKD on a imbalanced CIFAR-100 dataset with imbalance ratio $r = 0.01$.

We speculate the accuracy plunge of majority classes may stem from the generated data used for optimizing the student. Figure 2 provides a strong support for this hypothesis, where we observe the synthetic samples from the generator exhibit severe disparity among classes[1], *i.e.*, the occurrence of minority classes overwhelms majority classes throughout the entire adversarial distillation phase. This suggests that the adversarial formulation of DFKD with a biased teacher network results in a disparate impact on different groups of synthetic data. To gain intuition, in Figure 3, we provide a conceptual interpretation of this phenomenon with a binary classification task. In standard balanced DFKD, the generator yields synthetic samples that are distributed evenly among classes and learns to be consistent with the original data distribution. In contrast, for biased DFKD, the pseudo samples close to the teacher's inherently-biased decision boundary tend to be far away from the distribution of the majority class. Such samples are less likely to deliver meaningful representation of the majority class. This suggests that the generator prioritizes the synthesis of minority samples while disregarding majority samples, potentially resulting in a biased generator. As a result, the generator only examines a partial manifold of the teacher's representation space, extracting "incomplete" knowledge to the student.

Theoretical Analysis

While the conceptual interpretation mentioned above offers an intuitive explanation of the "distribution mismatch" in synthetic samples, it does not explain why the generator tends to prioritize the minority class from a theoretical standpoint. To gain a deeper understanding of this phenomenon, we explore this issue in a binary classification problem with input data sampled from a Gaussian mixture model (Bishop and Nasrabadi 2006).

[1]Synthetic samples do not have ground-truth labels; we use the classification results from the teacher or student to assign them predicted labels.

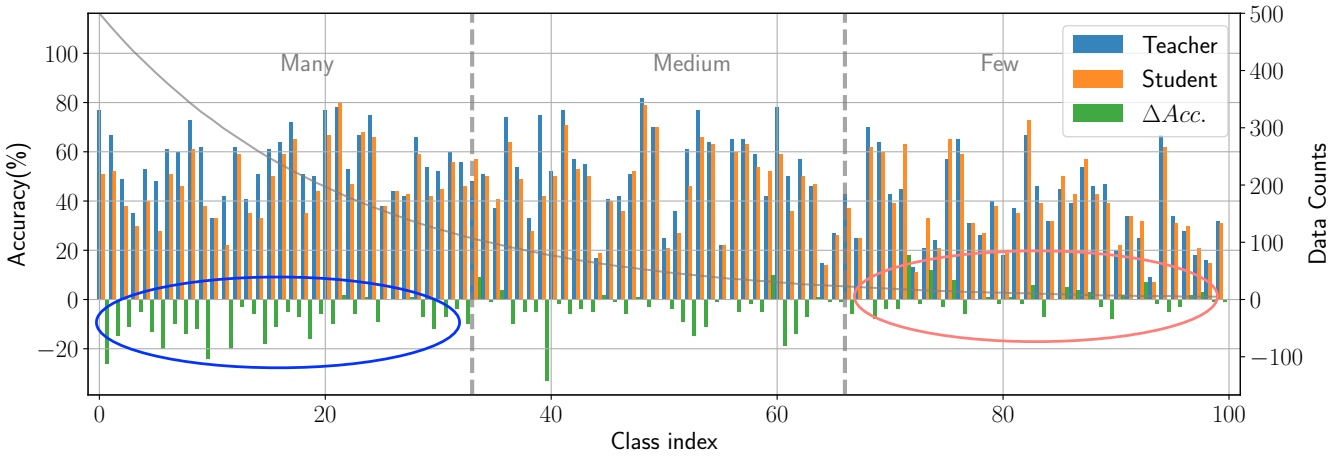


Figure 1: Comparison of class-wise accuracy change between a teacher pretrained on imbalanced CIFAR-100 dataset and a student after adversarial DFKD. The gray solid line represents the number of samples of each class, which can be quantified by the right vertical axis. The blue ellipse highlights the dramatic accuracy drop of majority classes, while the orange one indicates slight performance boosts of minority classes. Here, different shots (Many/Medium/Few) refer to the data groups in the original training dataset rather than synthetic samples.

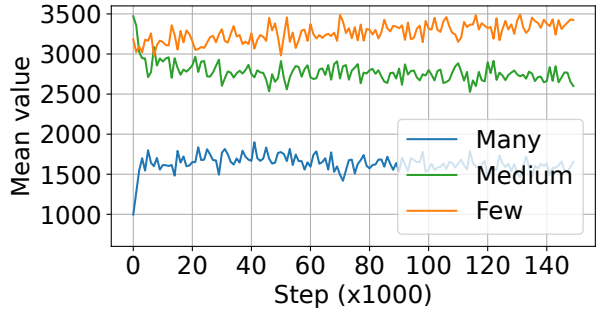
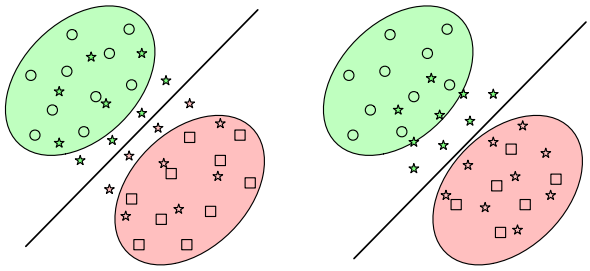


Figure 2: Biased adversarial DFKD has a disparate impact on different groups of synthetic data. We observe the number of synthetic samples predicted as minority classes by student prevails that of majority classes in the distillation phase.



(a) Standard DFKD (b) Biased DFKD

Figure 3: Comparing standard and biased adversarial DFKD on a binary classification task, where solid line represents the teacher's decision boundary, which is fixed throughout the distillation phase. The shapes ○ and □ respectively denote the original data of the majority and minority class, and ☆ represents synthetic samples. We illustrate the concept that in biased DFKD, synthetic samples from minority class are closer to the true distribution, while the opposite happens on the majority class.

Problem Definition We start by defining a binary classifier $f : \mathcal{X} \to \mathcal{Y}$ from input data $\mathcal{X} \in \mathbb{R}^d$ to output targets $\mathcal{Y} = \{-1, +1\}$. The overall empirical risk of f is defined as $\mathcal{R}(f) = \mathbb{P}\{f(x) \neq y\}$, which is the probability that f misclassifies the input data x. We use $\mathcal{R}(f; y)$ to denote the empirical risk for a class y.

We simulate two data clusters following an i.i.d Gaussian mixture distribution with identical variance $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ but different means centered at $-\boldsymbol{\mu}$ and $\boldsymbol{\mu}$, which gives the dataset $\mathcal{D}$:

$$\boldsymbol{x} \sim \begin{cases} \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) & \text{if } y = +1, \\ \mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) & \text{if } y = -1, \end{cases} \tag{2}$$

where $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu) \in \mathbb{R}^d$ and $\mu > 0$.

To simulate class imbalance, we assume the class "-1" is the majority class, and it samples N times as many data points as the minority class "+1". Then we consider an imbalanced classification problem with the overall empirical risk:

$$\mathcal{R}(f) = N\mathcal{R}(f; -1) + \mathcal{R}(f; +1), \tag{3}$$

where we assume f is instantiated as a linear classifier:

$$f(x) = \text{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b). \tag{4}$$

Before introducing the main theorem, we present two lemmas for characterizing the optimal classifier trained on an imbalanced dataset. Brief explanations of these lemmas can help to shed light on the reason for the aforementioned phenomenon. The following lemmas describe the property of

weights and bias term of the optimal binary classifier, along with the implication of class-wise empirical risks.

Lemma 1 (**equivalence of optimal weights**). *Consider the imbalanced classification problem defined in Equation (2), an optimal linear classifier f which minimizes the empirical risk:*

$$\boldsymbol{w}^*, b^* = \arg\min_{\boldsymbol{w},b} \mathbb{P}\{N\mathcal{R}(f;-1) + \mathcal{R}(f;+1)\} \quad (5)$$

satisfies that $w_1^* = w_2^* = \cdots = w_d^*$, i.e.*, $\boldsymbol{w}^* = w \cdot \mathbf{1}$.*

Lemma 1 shares some similarities with Lemma D.1 in (Tsipras et al. 2018). The property of optimal weights is a direct consequence of the isotropic property of input data among different dimensions. Namely, for features x_i and x_j in a given class, they follow the identical Gaussian distribution. Therefore, swapping different entries in $\boldsymbol{w}$, *e.g.*, w_i and w_j, should not change the output. According to Lemma 1, we can rewrite Equation (4) as:

$$\begin{aligned} f(x) &= \text{sign}(\langle w \cdot \boldsymbol{I}, \boldsymbol{x}\rangle + b) \\ &= \text{sign}(\langle \boldsymbol{I}, w \cdot \boldsymbol{x}\rangle + b). \end{aligned} \quad (6)$$

Here, $w \cdot \boldsymbol{x}$ is equivalent to a scaling transformation (preprocessing) on each dimension of input data $\boldsymbol{x}$. To simplify derivation, we assume $w = 1$, and the linear classifier reduces to:

$$f(x) = \text{sign}\left(\sum_{i=1}^{d} x_i + b\right). \quad (7)$$

Lemma 2 (**negativity of optimal bias**). *For the imbalanced classification problem defined in Equation (3), the optimal classifier has a negative bias term $b = -\frac{\log N}{4d\mu} < 0$. Thus, the empirical risks of the two classes are:*

$$\mathcal{R}(f;-1) = \mathbb{P}\left\{x < -\frac{\sqrt{d}\mu}{\sigma} - \frac{\log N}{4d\mu} \mid x \sim \mathcal{N}(0,1)\right\}$$

$$\mathcal{R}(f;+1) = \mathbb{P}\left\{x < -\frac{\sqrt{d}\mu}{\sigma} + \frac{\log N}{4d\mu} \mid x \sim \mathcal{N}(0,1)\right\},$$

where $\mathcal{N}(0,1)$ is the standard normal distribution. As a result, the minority class "+1" has a larger empirical risk:

$$\mathcal{R}(f;-1) < \mathcal{R}(f;+1). \quad (8)$$

Lemma 2 quantifies the optimal risks of the majority and minority classes. It also implies a biased decision boundary (towards the minority class) contingent on the bias term $b < 0$ as a result of learning from imbalanced data. With Lemmas 1 and 2, we are now ready to introduce the following theorem. For biased adversarial DFKD, we denote binary classifiers f_T and f_S as the teacher and student model, respectively.

Theorem 1. *Assume the teacher and student share the optimal weights introduced in Lemma 1, while the student has a different margin $\Delta b > 0$ on the bias term b compared with the teacher. Define the prediction discrepancy between teacher and student conditioned on a class $\tilde{y} \in \{-1, +1\}$ as $\mathcal{R}_{\mathcal{D}}(f_T, f_S; \tilde{y}) = \mathbb{P}\{f_T(x) \neq f_S(x) | y = \tilde{y}\}$. Then the discrepancy for the two classes satisfies:*

$$\mathcal{R}_{\mathcal{D}}(f_T, f_S; +1) > \mathcal{R}_{\mathcal{D}}(f_T, f_S; -1) \quad (9)$$

Theorem 1 demonstrates that the synthetic samples following the distribution of minority class "+1" provably yield a larger discrepancy between the teacher and student model. For this reason, the generator is encouraged to synthesize more minority pseudo samples to maximize the KL divergence, which corroborates our empirical observation in Figure 2. Note that Equation (9) is *not* limited by the optimal bias term $b = -\frac{\log N}{4d\mu}$ but only requires a negative bias term $b < 0$. Namely, the biased decision boundary is the root cause of imbalanced discrepancy among classes. For an in-depth proof, please see Appendix B. As a result, the generator tends to over-emphasize the minority classes, which, however, incurs mode collapse on majority classes. In Appendix F, we confirm this intuition by visualizing synthetic samples. We observed that synthetic images predicted as majority classes share homogeneous patterns due to mode collapse, whereas those for minority classes present discernible visual diversity.

For the above reasons, the distilled student are trained with *collapsed* and *limited* synthetic samples of majority classes. Consequently, the student cannot learn meaningful representation for majority classes, but acquires a biased representation of the synthetic dataset, which ultimately leads to the compromised performance of the majority classes.

The IPAD Method

Theorem 1 indicates the class-specific mode collapse of the adversarial generator has a close relationship with the imbalanced discrepancy among different classes. To mitigate this problem and address degenerate synthetic data of majority classes, we propose to rectify the learning objective of the generator in adversarial training stage, following the motivation by constrained learning (Chamon et al. 2022):

$$\max_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{x}=G(\boldsymbol{z};\boldsymbol{\phi})}[\mathcal{L}_{\text{KL}}(T(\boldsymbol{x}), S(\boldsymbol{x}))] \; s.t. \; \forall i \in [1, \ldots, K]:$$

$$\mathcal{L}_{\text{KL}}(T(\boldsymbol{x}_i), S(\boldsymbol{x}_i)) - \bar{\mathcal{L}}_{\text{KL}}(T(\boldsymbol{x}), S(\boldsymbol{x})) \leq \epsilon. \quad (10)$$

Here, $\bar{\mathcal{L}}_{\text{KL}}(T(\boldsymbol{x}), S(\boldsymbol{x}))$ is the mean KL divergence between teacher and student for all classes, and ϵ is a tolerance constant that decides the tightness of the constraint. To simplify notations, we denote $\mathcal{L}_{\text{KL}}(T(\boldsymbol{x}), S(\boldsymbol{x}))$ and $\mathcal{L}_{\text{KL}}(T(\boldsymbol{x}_i), S(\boldsymbol{x}_i))$ as $\mathcal{L}_{\text{KL}}(\boldsymbol{x})$ and $\mathcal{L}_{\text{KL}}(\boldsymbol{x}_i)$, respectively. To solve the constrained optimization problem, we leverage the Lagrange multipliers method (Boyd and Vandenberghe 2004) and introduce non-negative dual variables $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ associated with each class. We define the Lagrangian function:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\lambda}) = \; & \mathbb{E}_{\boldsymbol{x}=G(\boldsymbol{z};\boldsymbol{\phi})}[\mathcal{L}_{\text{KL}}(\boldsymbol{x})] \\ & + \frac{1}{K}\sum_{i=1}^{K} \lambda_i \left(\mathcal{L}_{\text{KL}}(\boldsymbol{x}_i) - \frac{1}{K}\sum_{j=1}^{K} \mathcal{L}_{\text{KL}}(\boldsymbol{x}_j) - \epsilon\right) \end{aligned} \quad (11)$$

T → S	**ResNet34 $\overset{\diamondsuit}{\rightarrow}$ ResNet18**		**WRN40-2 $\overset{\diamondsuit}{\rightarrow}$ WRN40-1**		**WRN40-2 $\overset{\diamondsuit}{\rightarrow}$ WRN16-2**		**ResNet34 $\overset{\heartsuit}{\rightarrow}$ VGG16**	
CIFAR-100 ($r = 0.01$)								
Acc. (%)	Shot Acc.	Overall	Shot Acc.	Overall	Shot Acc.	Overall	Shot Acc.	Overall
Teacher‡	65.2, 62.4, 44.2	57.3	62.0, 58.1, 41.5	53.9	62.0, 58.1, 41.5	53.9	65.2, 62.4, 44.2	57.3
Student‡	61.4, 58.2, 44.2	54.6	59.1, 50.7, 36.0	48.6	62.0, 56.0, 41.2	53.1	61.5, 49.2, 27.7	46.1
DAFL	56.4, 55.6, 41.0	51.0	51.2, 50.7, 39.0	47.0	54.9, 52.1, 40.0	49.0	49.9, 49.2, 33.5	44.2
ABM	53.9, 59.2, 44.4	52.5	51.0, 50.9, 40.1	47.3	55.3, 53.5, 42.4	50.4	50.4, 50.6, 36.6	45.9
CMI	58.7, 58.2, 43.6	53.5	51.7, 51.4, 40.3	47.8	55.4, 54.1, 42.6	50.7	52.4, 51.2, 38.0	47.2
MAD	54.0, 60.4, 44.3	52.9	51.8, 51.8, 40.7	48.1	54.2, 54.0, 42.1	50.1	50.8, 50.8, 39.1	46.9
IPAD	59.9, 60.7, 44.1	**54.9**	55.0, 53.9, 40.1	**49.7**	59.3, 55.5, 41.8	**52.2**	57.7, 52.4, 37.5	**49.2**
Tiny-ImageNet ($r = 0.2$)								
Teacher‡	66.0, 51.3, 46.2	54.5	57.7, 45.5, 41.8	48.3	57.7, 45.5, 41.8	48.3	66.0, 51.3, 46.2	54.5
Student‡	63.5, 51.9, 45.7	53.7	55.3, 45.1, 41.7	47.4	55.8, 44.2, 41.5	47.1	60.8, 48.5, 42.4	50.5
DAFL	25.9, 15.8, 15.9	19.2	30.3, 16.5, 18.3	21.7	30.1, 22.9, 14.2	22.4	39.3, 23.4, 18.0	26.9
ABM	53.3, 44.9, 46.1	48.1	36.9, 31.7, 30.1	32.9	39.3, 34.2, 34.8	36.1	44.2, 35.0, 35.6	38.3
CMI	56.5, 49.6, 46.0	50.7	35.2, 32.9, 31.8	33.3	38.2, 36.0, 36.5	36.9	45.0, 38.4, 36.9	40.1
MAD	55.0, 49.1, 46.8	50.3	37.2, 31.4, 31.9	33.5	37.4, 37.4, 36.8	37.2	45.4, 39.3, 37.1	40.6
IPAD	63.3, 50.5, 45.8	**53.2**	44.0, 33.8, 30.2	**36.0**	43.2, 38.1, 34.5	**38.6**	50.3, 42.9, 36.4	**43.2**

Table 2: A performance overview of DFKD methods with different imbalanced pretraining datasets and varying imbalance ratio r. The symbol ‡ represents the train-from-scratch accuracy on imbalanced data followed by rebalanced finetuning (Kang et al. 2019). We evaluate homogeneous ($\diamondsuit$) and heterogeneous ($\heartsuit$) model architecture configurations of teacher and student. Here, "Shot Acc." denotes the Many/Medium/Few shot accuracy.

Then $\mathcal{L}(\phi, \boldsymbol{\lambda})$ can be minimized by alternating optimization of the primal ϕ and dual $\boldsymbol{\lambda}$ variable:

Primal Update: With a fixed $\boldsymbol{\lambda}$, the minimization of $\mathcal{L}(\phi, \boldsymbol{\lambda})$ *w.r.t.* ϕ is equivalent to a constrained version of the original objective with class-specific adjustment terms.

Dual Update: Once we obtained an updated generator, we can perform dual updating on $\boldsymbol{\lambda}$ for several steps by taking dual gradient ascent (Boyd et al. 2011) of the following objective:

$$\max_{\boldsymbol{\lambda}} \frac{1}{K} \sum_{i=1}^{K} \lambda_i \mathbb{E}_{x=G(z)} \left(\mathcal{L}_{\mathrm{KL}}(\boldsymbol{x}_i) - \frac{1}{K} \sum_{j=1}^{K} \mathcal{L}_{\mathrm{KL}}(\boldsymbol{x}_j) - \epsilon \right).$$

The parameter of $\boldsymbol{\lambda}$ is updated as follows:

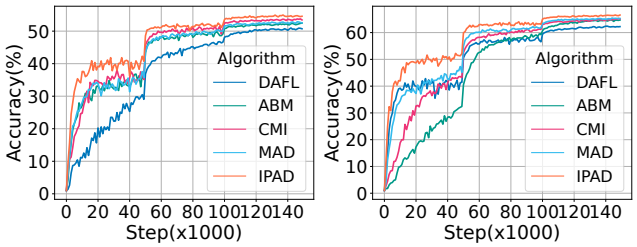
$$\boldsymbol{\lambda} \leftarrow \left[\boldsymbol{\lambda} + \eta_D \frac{1}{K} \sum_{i=1}^{K} \left(\mathcal{L}_{\mathrm{KL}}(\boldsymbol{x}_i) - \frac{1}{K} \sum_{j=1}^{K} \mathcal{L}_{\mathrm{KL}}(\boldsymbol{x}_j) - \epsilon \right) \right]_+.$$

Here, $[\cdot]_+$ clips $\boldsymbol{\lambda}$ to ensure non-negativity of the dual variable $\boldsymbol{\lambda} \in \mathbb{R}_+^K$, and η_D denotes the learning rate of dual update. Solving Equation (10) under constraints enjoys a clear interpretation of the class-balancing effect compared with the original adversarial knowledge distillation objective. Namely, if the KL divergence of a specific class exceeds the mean value too much, it receives a penalty to down-weigh its influence; inversely, it could be adjusted by a compensation term if smaller than the mean value. Such regularization prevents the generator's optimization from being dominated by minority classes. It is important to note that our method neither requires the composition of original training data nor intends to disclose such information by estimating a proxy data distribution. It thus strictly adheres the principle of privacy preservation in data-free knowledge transfer. In Appendix E, we present more details of the proposed method.

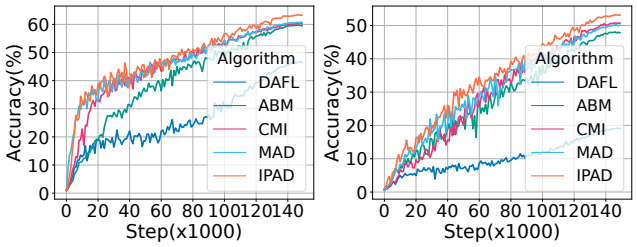
Experiments

Experiment Settings

Datasets We evaluate IPAD against the prevalent DFKD methods on multiple image classification datasets that are frequently used in recent DFKD literature (Do et al. 2022; Fang et al. 2021b,a), including CIFAR-100 (Krizhevsky, Hinton et al. 2009), Tiny-ImageNet (Le and Yang 2015), Food101 (Bossard, Guillaumin, and Van Gool 2014), Places365 (Zhou et al. 2017) and ImageNet (Deng et al. 2009). To simulate imbalanced data for pretraining the teacher, we follow (Kang et al. 2019) to create an imbalanced subset from the original dataset by subsampling data in each class according to an exponential function $n_i = N_{max} \cdot r^{i/K}$. Here, i is the class index and K is the total number of classes. N_{max} denotes the maximum number of data for the head class. Importantly, $r \in (0, 1)$ determines the imbalance ratio, and a smaller r indicates more pronounced data imbalance. For example, $r = 0.01$ means the tail class possesses only 1% data points compared with the head class. We choose multiple configurations of r, ranging from 0.2 to 0.01 for different datasets to cover both *moderated* and *ag-*



(a) CIFAR-100 ($r = 0.01$) (b) CIFAR-100 ($r = 0.05$)



(c) Food101 ($r = 0.1$) (d) Tiny-ImageNet ($r = 0.2$)

Figure 4: A comparison of convergence curves on different datasets under the ResNet34 $\rightarrow$ ResNet18 setup.

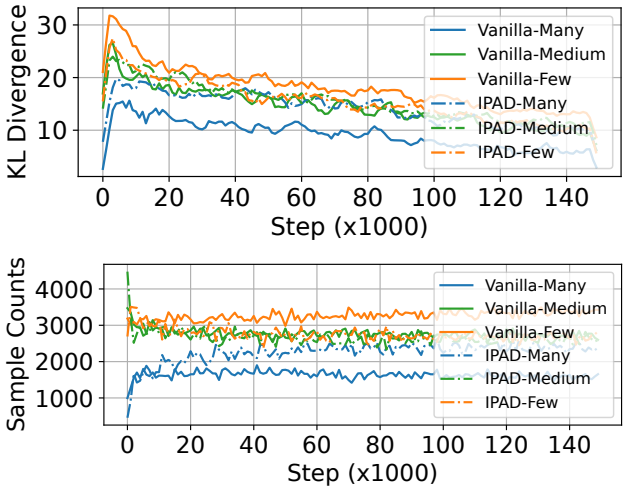


Figure 5: An illustration of class-balancing effect of IPAD on each data groups under WRN40-2 $\rightarrow$ WRN16-2 setup on CIFAR-100 dataset. Top: KL divergence. Bottom: sample counts.

gressive imbalance levels.

Competing methods. We compare our method with existing non-adversarial (DAFL (Chen et al. 2019)) and adversarial methods (ABM (Micaelli and Storkey 2019), MAD (Do et al. 2022) and CMI (Fang et al. 2021b)). As these methods are not designed for tackling biased DFKD, we append an entropy-maximization regularization term (Micaelli and Storkey 2019) to their original learning objective to introduce a balancing effect to provide them a fair comparison. Detailed experiement setups are outlined in Appendix D.

Main Results

Table 2 reports the accuracy of distilled students on different datasets with varying imbalance ratios. Generally, our method outperforms other competing methods by improving the accuracy of many shot classes. In the ResNet34 $\rightarrow$ ResNet18 case, it even matches or surpasses the student's train-from-scratch accuracy on original data. Figure 4 presents the accuracy curves of all compared methods, where IPAD generally converges faster than its competitors.

Discussion

Is non-adversarial DFKD a direct remedy? Since adversarial optimization causes the imbalanced learning problem in biased DFKD, it is natural to ask: *shall we relinquish the adversarial paradigm to circumvent this issue?* Unfortunately, we discover that the non-adversarial method DAFL, in most experiments, exhibits inferior performance than adversarial baselines. On CIFAR-100 dataset, we find the DAFL's accuracy plateaus prematurely compared with its adversarial counterparts. It fails even more catastrophically when scaling to larger datasets such as Tiny-ImageNet. This aligns with the results in standard DFKD (Micaelli and Storkey 2019; Do et al. 2022; Fang et al. 2022) that non-adversarial methods usually struggle to achieve on-par performance of adversarial methods. Since the generator utilized in DFKD draws inspiration from the generative adversarial network (GAN) (Goodfellow et al. 2020), it's likely unable to explore the teacher's representation space sufficiently without the aid of adversarial supervision signals. This is also in line with the experiments in CMI (Fang et al. 2021b) that ablating the adversarial loss component leads to clear performance degradation.

IPAD features a clear class-balancing effect. In Fig. 5, we illustrate the class-balancing effect of IPAD and its influence on the synthetic samples. While the vanilla adversarial DFKD exhibits a notable gap of the KL divergence between the "Many" and "Few" classes, IPAD has shown great efficacy in reducing this gap due to its use of constrained optimization. Furthermore, it serves as an powerful tool for addressing the imbalance in synthetic data, ultimately leading to improved performance in majority classes. We defer more experiments on larger datasets and ablations to Appendix F.

Conclusion

In this work, we have observed an interesting phenomenon that in biased adversarial DFKD, majority classes from original training data "involuntarily convert" to minority groups of synthetic samples. We have empirically and theoretically revealed the underlying reasons and proposed a solution, IPAD, to effectively tackle this problem. We believe our findings, along with the observations from early works (Binici et al. 2022b,a; Do et al. 2022), show that adversarial DFKD, albeit effective, has some pitfalls that are raised either by distribution shift or label shift, and thus, must be carefully handled with in practice.

Acknowledgements

This research is supported by Science and Technology Development Fund of Macau SAR (Nos. 0081/2022/A2, 0123/2022/AFJ, and 0015/2019/AKP), Guangdong Basic and Applied Basic Research Foundation (No. 2020B515130004), National Natural Science Foundation of China (No. 62376263), Basic Research Program of Shenzhen (No. JCYJ20230807140507015). This work was carried out in part at SICC, which is supported by SKL-IOTSC, University of Macau.

References

Azizzadenesheli, K.; Liu, A.; Yang, F.; and Anandkumar, A. 2018. Regularized Learning for Domain Adaptation under Label Shifts. In *International Conference on Learning Representations*.

Binici, K.; Aggarwal, S.; Pham, N. T.; Leman, K.; and Mitra, T. 2022a. Robust and Resource-Efficient Data-Free Knowledge Distillation by Generative Pseudo Replay. *arXiv preprint arXiv:2201.03019*.

Binici, K.; Pham, N. T.; Mitra, T.; and Leman, K. 2022b. Preventing catastrophic forgetting and distribution mismatch in knowledge distillation via synthetic data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 663–671.

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J.; et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1): 1–122.

Boyd, S. P.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Cai, J.; Wang, Y.; and Hwang, J.-N. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 112–121.

Chamon, L. F.; Paternain, S.; Calvo-Fullana, M.; and Ribeiro, A. 2022. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chen, H.; Wang, Y.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Xu, C.; Xu, C.; and Tian, Q. 2019. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3514–3522.

Choi, Y.; Choi, J.; El-Khamy, M.; and Lee, J. 2020. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 710–711.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Do, K.; Le, H.; Nguyen, D.; Nguyen, D.; Harikumar, H.; Tran, T.; Rana, S.; and Venkatesh, S. 2022. Momentum Adversarial Distillation: Handling Large Distribution Shifts in Data-Free Knowledge Distillation. *arXiv preprint arXiv:2209.10359*.

Fang, G.; Bao, Y.; Song, J.; Wang, X.; Xie, D.; Shen, C.; and Song, M. 2021a. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems*, 34: 11920–11932.

Fang, G.; Mo, K.; Wang, X.; Song, J.; Bei, S.; Zhang, H.; and Song, M. 2022. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6597–6604.

Fang, G.; Song, J.; Shen, C.; Wang, X.; Chen, D.; and Song, M. 2019. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*.

Fang, G.; Song, J.; Wang, X.; Shen, C.; Wang, X.; and Song, M. 2021b. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Gao, X.; Zhao, Y.; Dudziak, Ł.; Mullins, R.; and Xu, C.-z. 2018. Dynamic channel pruning: Feature boosting and suppression. *arXiv preprint arXiv:1810.05331*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Han, P.; Park, J.; Wang, S.; and Liu, Y. 2021. Robustness and diversity seeking data-free knowledge distillation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2740–2744. IEEE.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9): 1263–1284.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Jiang, Z.; Chen, T.; Mortazavi, B. J.; and Wang, Z. 2021. Self-damaging contrastive learning. In *International Conference on Machine Learning*, 4927–4939. PMLR.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational {Bayes}. In *Int. Conf. on Learning Representations*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technical report*.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *Technical report*, 7(7): 3.

Li, S.; Gong, K.; Liu, C. H.; Wang, Y.; Qiao, F.; and Cheng, X. 2021. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5212–5221.

Lopes, R. G.; Fenu, S.; and Starner, T. 2017. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*.

Luo, L.; Sandler, M.; Lin, Z.; Zhmoginov, A.; and Howard, A. 2020. Large-scale generative data-free distillation. *arXiv preprint arXiv:2012.05578*.

Micaelli, P.; and Storkey, A. J. 2019. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, 32.

Nayak, G. K.; Mopuri, K. R.; Shaj, V.; Radhakrishnan, V. B.; and Chakraborty, A. 2019. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, 4743–4751. PMLR.

Patel, G.; Mopuri, K. R.; and Qiu, Q. 2023. Learning to Retain while Acquiring: Combating Distribution-Shift in Adversarial Data-Free Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7786–7794.

Rangwani, H.; Jaswani, N.; Karmali, T.; Jampani, V.; and Babu, R. V. 2022. Improving GANs for Long-Tailed Data Through Group Spectral Regularization. In *European Conference on Computer Vision*, 426–442. Springer.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1701–1708.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33: 1513–1524.

Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599.

Tran, C.; Fioretto, F.; Kim, J.-E.; and Naidu, R. 2022. Pruning has a disparate impact on model accuracy. *arXiv preprint arXiv:2205.13574*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2018. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wang, K.; Gao, X.; Zhao, Y.; Li, X.; Dou, D.; and Xu, C.-Z. 2019. Pay attention to features, transfer learn faster CNNs. In *International conference on learning representations*.

Wang, Z. 2021. Data-free knowledge distillation with soft targeted transfer set synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10245–10253.

Wu, R.; Guo, C.; Su, Y.; and Weinberger, K. Q. 2021. Online adaptation to label distribution shift. *Advances in Neural Information Processing Systems*, 34: 11340–11351.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8715–8724.

Yoo, J.; Cho, M.; Kim, T.; and Kang, U. 2019. Knowledge extraction with no observable data. *Advances in Neural Information Processing Systems*, 32.

Yu, S.; Chen, J.; Han, H.; and Jiang, S. 2023. Data-Free Knowledge Distillation via Feature Exchange and Activation Region Constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24266–24275.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2021. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhao, Y.; Gao, X.; Bates, D.; Mullins, R.; and Xu, C.-Z. 2019. Focused quantization for sparse CNNs. *Advances in Neural Information Processing Systems*, 32.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.