Hybrid Data-Free Knowledge Distillation

Jialiang Tang[1,2,3], **Shuo Chen**[4*], **Chen Gong**[5*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, China
[3]Jiangsu Key Laboratory of Image and Video Understanding for Social Security, China
[4]Center for Advanced Intelligence Project, RIKEN, Japan
[5]Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
tangjialiang@njust.edu.cn, shuo.chen.ya@riken.jp, goodgongchen@gmail.com

Abstract

Data-free knowledge distillation aims to learn a compact student network from a pre-trained large teacher network without using the original training data of the teacher network. Existing collection-based and generation-based methods train student networks by collecting massive real examples and generating synthetic examples, respectively. However, they inevitably become weak in practical scenarios due to the difficulties in gathering or emulating sufficient real-world data. To solve this problem, we propose a novel method called **H**ybr**i**d **D**ata-**F**ree **D**istillation (HiDFD), which leverages only a small amount of collected data as well as generates sufficient examples for training student networks. Our HiDFD comprises two primary modules, *i.e.*, the teacher-guided generation and student distillation. The teacher-guided generation module guides a Generative Adversarial Network (GAN) by the teacher network to produce high-quality synthetic examples from very few real-world collected examples. Specifically, we design a feature integration mechanism to prevent the GAN from overfitting and facilitate the reliable representation learning from the teacher network. Meanwhile, we drive a category frequency smoothing technique via the teacher network to balance the generative training of each category. In the student distillation module, we explore a data inflation strategy to properly utilize a blend of real and synthetic data to train the student network via a classifier-sharing-based feature alignment technique. Intensive experiments across multiple benchmarks demonstrate that our HiDFD can achieve state-of-the-art performance using 120 times less collected data than existing methods. Code is available at https://github.com/tangjialiang97/HiDFD.

Introduction

The success of Deep Neural Networks (DNNs) (He et al. 2016; Hao et al. 2024) is usually accompanied by significant computational and storage demands, which hinders their deployment on practical resource-limited devices. Knowledge Distillation (KD) (Hinton, Vinyals, and Dean 2015; Miles and Mikolajczyk 2024) has served as an effective compression technology that transfers knowledge from a complex pre-trained teacher network to improve the performance of a lightweight student network. However, in practice, the training data of the teacher network is usually inaccessible due to privacy concerns and only the pre-trained teacher network itself can be used to learn the student network. This is because users may prefer sharing a pre-trained black box DNN rather than disclosing their sensitive data. In such cases, vanilla KD methods can hardly train a reliable student network owing to the absence of original training data. To address this issue, various Data-Free Knowledge Distillation (DFKD) approaches (Binici et al. 2022; Chen et al. 2019, 2021b; Tang et al. 2023) have been developed to enable training the student network without using any original data.

Among existing DFKD methods, collection-based approaches (Chen et al. 2021b; Tang et al. 2023) can achieve satisfactory performance by amassing numerous real examples to train the student network. However, it is still difficult for the collection-based methods to train a reliable student network in practical tasks, *e.g.*, medical image classification because gathering sufficient training examples can be challenging. On the other hand, generation-based methods (Yin et al. 2020; Chen et al. 2019) leverage the teacher network to guide a generative model (Creswell et al. 2018) in producing fake examples, thereby successfully training the student network without reliance on real examples. Nevertheless, the synthesized examples may exhibit low quality in the absence of real data supervision, leading to suboptimal student performance, especially for many challenging recognition tasks on ImageNet (Deng et al. 2009). The inherent constraints of both collection-based and generation-based DFKD methods prompt an essential question: *Can we train an effective generative model only using a small number of collected examples and then learn reliable student networks with the hybrid data comprising both collected and synthetic examples?*

To answer the above question under the practical data-free distillation scenario, we need a generative model that not only possesses powerful generative capabilities but also has the ability to acquire valuable knowledge from the teacher network. Recent studies (Cui et al. 2023; Rangwani, Mopuri, and Babu 2021) suggest that the Generative Adversarial Network (GAN) (Mirza and Osindero 2014) can easily learn from pre-trained models and then generate high-quality synthetic examples, so we employ this great approach as our generative module. The standard GAN consists of a generator and a discriminator trained in an adversarial manner,

*Corresponding authors: Chen Gong, Shuo Chen.

where the generator attempts to produce fake examples to deceive the discriminator while the discriminator strives to distinguish between real and fake examples. However, the collected data in practice tasks like medical image classification has two inherent characteristics that may impede the training of the GAN, namely: 1) *Limited data quantity*, as capturing medical images requires expensive and complex equipment; and 2) *Imbalanced class distribution*, where certain diseases (*e.g.*, "vascular lesions") are more rarely than others (*e.g.*, "nevus"). When training on the collected data with limited examples and imbalanced class distribution, the discriminator is susceptible to overfitting (Huang et al. 2022; Jiang et al. 2021). It implies that the discriminator tends to memorize all real examples and almost perfectly distinguish them from fake examples, resulting in the disappearance of the gradient for the generator. Moreover, the generator training is dominated by a few classes occupying the majority of examples, which prevents it from generating diverse examples. Therefore, it is critical to overcome the overfitting issue of discriminator and data imbalance issue of generator when training with scarce collected examples.

In this paper, we propose a novel approach called **H**ybr**i**d **D**ata-**F**ree **D**istillation (HiDFD), which learns reliable student networks on the hybrid data comprising synthetic examples and very few real collected examples. Our HiDFD is composed of two pivotal modules of teacher-guided generation and student distillation. In the teacher-guided generation module, we aim to solve the critical issues in the GAN mentioned above, and thus generating high-quality synthetic examples. Specifically, we propose a feature integration mechanism to aggregate the features of both the collected and synthetic examples between the teacher network and GAN. Such an integration mechanism not only mitigates the overfitting of the discriminator, which forcibly distinguishes those closely resembling examples, but also transfers valuable representations to guide the discriminator to capture category dependencies. Meanwhile, we also develop a new technique called category frequency smoothing to alleviate the imbalanced training of the generator. In the student distillation module, we develop a data inflation operation to adjust the contribution of collected examples among the hybrid data when training the student network. Finally, we design a classifier-sharing-based strategy to closely align the features of student network with those of teacher network to enhance student performance. Thanks to effectively transferring knowledge from the teacher network to both the GAN and student network, our HiDFD can successfully train reliable student networks using very few collected real-world examples. The contributions of our HiDFD are summarized as follows:

- By considering the difficulties in gathering or emulating real-world data, we propose a novel data-free distillation method called HiDFD, which only requires a small number of collected data to generate high-quality synthetic examples for training the student network.
- We design a teacher-guided generation module to effectively tackle the critical issues of discriminator overfitting and imbalanced learning in generating synthetic examples, which empowers the distillation module to learn reliable student networks from the teacher network.
- Our HiDFD can achieve State-Of-The-Art (SOTA) performance using only 1/120 (5,000/600,000) of examples required by existing collection-based DFKD methods.

Related Works

In this section, we review the relevant works, including knowledge distillation and generative models.

Knowledge Distillation

Traditional KD methods (Chen et al. 2021a; Li et al. 2023) learn a compact and reliable student network by encouraging it to mimic a variety of knowledge, *i.e.*, softened logits (Zhao et al. 2022a), intermediate features (Chen et al. 2022), and representation relationships (Peng et al. 2019), from a large teacher network using ample original training data. However, in practical applications, these approaches might be ineffective because the original data is usually unavailable due to privacy concerns.

To address the above issue, generation-based (Tran et al. 2024; Wang et al. 2024a,b) and collection-based (Chen et al. 2021b; Tang et al. 2023) DFKD methods have been proposed to train student networks using synthetic and collected data, respectively. The generation-based methods utilize the teacher network to guide a generator in producing examples from statistics in the teacher network or random noise. However, the resulting student network still achieves suboptimal performance due to the flawed synthetic examples. Conversely, collection-based methods assume that there are numerous easily accessible examples in the real-world, and they acquire an oversized collected data (*e.g.*, 600,000 examples on CIFAR10) to train the student network. In practical tasks, it is hard to gather so many examples, and thus they still fail to train reliable student networks.

In this paper, our HiDFD only utilizes a small collected data that contains fewer examples than the original data, which initially guides the GAN in training on such collected data by the teacher and then trains the student on adequate data composed of the synthetic and collected examples.

Generative Models

Recent advances in generative models, including Variational AutoEncoders (Kingma and Welling 2013; Zhao, Song, and Ermon 2019), diffusion models (Ho, Jain, and Abbeel 2020; Mei and Patel 2023), and GAN (Hou et al. 2021; Mirza and Osindero 2014) have significantly propelled the data generation. This paper focuses on the powerful GAN due to its ability to learn from pre-trained models (Cui et al. 2023; Rangwani, Mopuri, and Babu 2021). The traditional GAN (Goodfellow et al. 2020) consists of a generator and a discriminator, where the generator produces fake examples to deceive the discriminator, and the discriminator tries to accurately distinguish between real and fake examples. Recently, Auxiliary Discriminative Classifier GAN (ADCGAN) (Hou et al. 2022) captures dependencies between generated examples and class labels by encouraging the discriminator to

classify synthetic examples into specific categories, which effectively improves the quality of synthetic data.

In our method, we hope the GAN can produce high-quality synthetic examples that are easily classifiable, and thus training a precise student network. Therefore, we adopt ADCGAN as the foundational generative model. The ADCGAN composed of a generator $\mathcal{N}_{\mathrm{G}} : \mathcal{Z} \times \mathcal{Y} \rightarrow \mathcal{X}$ maps a noise-label pair $(\boldsymbol{z}, y)$ to a fake example $\mathcal{N}_{\mathrm{G}}(\boldsymbol{z}, y) \in \mathcal{X}$ that can be precisely predicted as $y \in \mathcal{Y}$; and a discriminator $\mathcal{N}_{\mathrm{D}} : \mathcal{X} \rightarrow \{0, 1\}$ determines whether the input example is real (*i.e.*, 1) or fake (*i.e.*, 0), which also has a classifier $\Psi_{\mathrm{D}} : \mathcal{X} \rightarrow \mathcal{Y}^+ \cup \mathcal{Y}^-$ ($y^+ \in \mathcal{Y}^+$ and $y^- \in \mathcal{Y}^-$ denote the labels for real and fake examples, respectively). Mathematically, the objective functions for the discriminator and generator in the ADCGAN are defined as $\mathcal{L}_{\mathrm{adc\_d}}$ and $\mathcal{L}_{\mathrm{adc\_g}}$, respectively, as follows:

$$\begin{cases} \mathcal{L}_{\mathrm{adc\_d}} = -\mathcal{L}_{\mathrm{d}} + \mathbb{E}_{\boldsymbol{x},y\sim P_{X,Y}}[\log \Psi_{\mathrm{D}}(y^+ \mid \boldsymbol{x})] \\ \qquad\qquad + \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log \Psi_{\mathrm{D}}(y^- \mid \boldsymbol{x})], \\ \mathcal{L}_{\mathrm{adc\_g}} = \mathcal{L}_{\mathrm{g}} - \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log \Psi_{\mathrm{D}}(y^+ \mid \boldsymbol{x})] \\ \qquad\qquad + \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log \Psi_{\mathrm{D}}(y^- \mid \boldsymbol{x})], \end{cases} \tag{1}$$

where $\mathcal{L}_{\mathrm{d}} = \mathbb{E}_{\boldsymbol{x}\sim P_X}[\log \mathcal{N}_D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim Q_X}[\log(1 - \mathcal{N}_D(\boldsymbol{x}))]$ and $\mathcal{L}_{\mathrm{g}} = \mathbb{E}_{\boldsymbol{x}\sim Q_X}[\log(1 - \mathcal{N}_D(\boldsymbol{x}))]$ are the loss functions for the standard GAN, P and Q denote the distribution of real collected data and fake synthetic data, respectively. $\Psi_{\mathrm{D}}(y^+ \mid \cdot)$ (resp. $\Psi_{\mathrm{D}}(y^- \mid \cdot)$) denotes the probability that the input example is classified as the label y and real (resp. fake) simultaneously by the following classifier of the discriminator. Formally, $\Psi_{\mathrm{D}}(y^+ \mid \boldsymbol{x}) = \frac{\exp\left(\Psi_{\mathrm{D}}^+(y)\cdot\Phi_{\mathrm{D}}(\boldsymbol{x})\right)}{\sum_{\bar{y}\in\mathcal{Y}^+}\exp\left(\Psi_{\mathrm{D}}^+(\bar{y})\cdot\Phi_{\mathrm{D}}(\boldsymbol{x})\right)+\sum_{\bar{y}\in\mathcal{Y}^-}\exp\left(\Psi_{\mathrm{D}}^-(\bar{y})\cdot\Phi_{\mathrm{D}}(\boldsymbol{x})\right)}$, where Φ_{D} represents the shared feature extractor between the original discriminator $\mathcal{N}_{\mathrm{D}}$ and the classifier Ψ_{D}. Ψ_{D}^+ (resp. Ψ_{D}^-) captures the dependencies between the category labels and real (resp. fake) data. Notably, DeGAN (Addepalli et al. 2020) also trains a GAN using collected data, but it still requires a large number of collected examples and can only utilize synthetic examples to train the target model.

Approach

Data-free distillation aims to train a compact student network $\mathcal{N}_{\mathrm{S}}$ using a pre-trained teacher network $\mathcal{N}_{\mathrm{T}}$ without accessing the teacher's original training data $\mathcal{D}_{\mathrm{o}}$. Both $\mathcal{N}_{\mathrm{T}}$ and $\mathcal{N}_{\mathrm{S}}$ consist of a feature extractor Φ and classifier Ψ, where the subscripts T and S indicate "teacher" and "student", respectively. Existing collection-based DFKD methods (Tang et al. 2023) usually rely on the collected data $\mathcal{D}_{\mathrm{c}}$ with overwhelming examples searched based on the categories of the original data. Here, the data amount $|\mathcal{D}_{\mathrm{c}}| \gg |\mathcal{D}_{\mathrm{o}}|$, which is hard to satisfy in practice tasks. To overcome this limitation, we propose a more practical method that only requires a small number of collected examples for DFKD, *i.e.*, the data amount $|\mathcal{D}_{\mathrm{c}}| \leq |\mathcal{D}_{\mathrm{o}}|$. To this end, we develop a hybrid framework to generate abundant synthetic examples from very few collected examples, and then we integrate them as the hybrid data for training the reliable student network.

Motivation of the Hybrid Learning

Formally, we denote the distribution of the collected data $\mathcal{D}_{\mathrm{c}}$ and synthetic data $\mathcal{D}_{\mathrm{s}}$ as P and Q, respectively, while the distribution of the hybrid data $\mathcal{D} = \mathcal{D}_{\mathrm{c}} \cup \mathcal{D}_{\mathrm{s}}$ is represented as $U = \alpha P + (1 - \alpha)Q$. Here $\alpha = |\mathcal{D}_{\mathrm{c}}| \,/\, (|\mathcal{D}_{\mathrm{c}}| + |\mathcal{D}_{\mathrm{s}}|)$ represents the proportion of collected examples in the hybrid data. In general, the synthetic and collected examples usually exhibit a significant distribution gap. This can cause substantial fluctuations during the training of the student network on hybrid data, ultimately leading to poor performance (Wang, Zhang, and Wang 2024). Therefore, it is essential to align the distribution of synthetic data with that of collected data, thereby forming reliable hybrid data. Here, the synthetic data is generated under the supervision of the collected data, so we assume that synthetic data, collected data, and hybrid data have the same support set $\mathcal{X}$. Then, the distribution gap between the reliable hybrid data and synthetic data can be characterized by the Total Variation Distance (TVD), which is defined as

$$\mathrm{TVD}(U, Q) = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} |U(\boldsymbol{x}) - Q(\boldsymbol{x})|, \tag{2}$$

where $U(\boldsymbol{x}) \in (0, 1)$ and $Q(\boldsymbol{x}) \in (0, 1)$ measure the distribution probability of $\boldsymbol{x}$ in the hybrid and synthetic data, respectively. Here $\mathrm{TVD}(\cdot, \cdot) \geq 0$, and $\mathrm{TVD}(U, Q) = \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} |U(\boldsymbol{x}) - Q(\boldsymbol{x})| \leq \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}}(|U(\boldsymbol{x})| + |Q(\boldsymbol{x})|) = 1$. Based on the triangle inequality (Steerneman 1983) of TVD, we easily have that

$$\mathrm{TVD}(U, Q) \leq \mathrm{TVD}(U, P) + \mathrm{TVD}(P, Q). \tag{3}$$

Then, given that $U = \alpha P + (1 - \alpha)Q$ with parameter α controlling the weight of collected data, we can compute $\mathrm{TVD}(U, P)$ as

$$\begin{aligned} \mathrm{TVD}(U, P) &= \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} |U(\boldsymbol{x}) - P(\boldsymbol{x})| \\ &= \frac{1}{2}\sum_{\boldsymbol{x}\in\mathcal{X}} |\alpha P(\boldsymbol{x}) + (1-\alpha)Q(\boldsymbol{x}) - P(\boldsymbol{x})| \\ &= \frac{1}{2}(1-\alpha)\sum_{\boldsymbol{x}\in\mathcal{X}} |Q(\boldsymbol{x}) - P(\boldsymbol{x})| \\ &= (1-\alpha)\mathrm{TVD}\left(Q, P\right). \end{aligned} \tag{4}$$

By invoking the symmetry of TVD and Eq. (3), we obtain

$$\mathrm{TVD}(U, Q) \leq (2 - \alpha)\mathrm{TVD}(P, Q). \tag{5}$$

Here Eq. (5) reveals that the **high-quality synthetic data** $\mathcal{D}_{\mathrm{s}}$ and the **mix proportion** α are two critical factors influencing the distribution gap $\mathrm{TVD}(U, P)$.

The above observation inspires us to employ two modules to align the distribution of synthetic data with that of collected data, as shown in Fig. 1(c). In the teacher-guided generation module, we employ the teacher network to guide the GAN to enhance the quality of synthetic data, which solves its intrinsic issues when trained on the small and imbalanced collected data, including the overfitting of discriminator and imbalanced learning of generator:

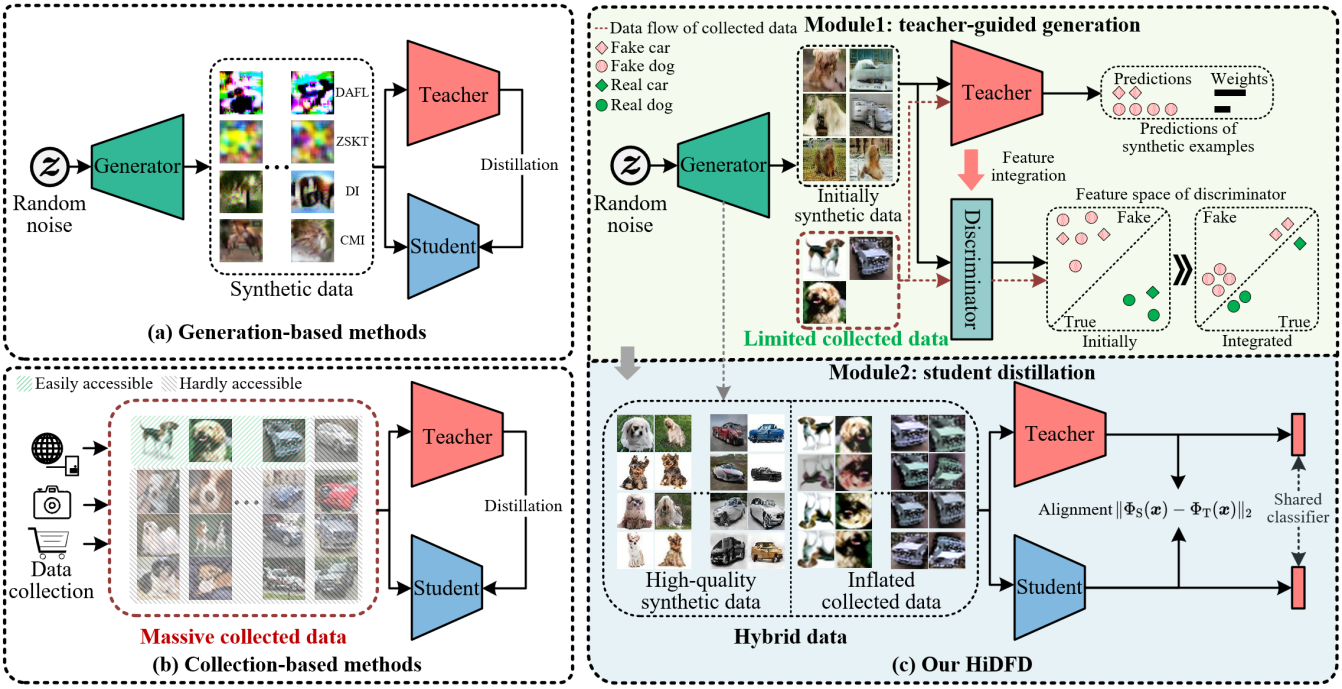


Figure 1: The diagram of (a) generation-based methods (Fang et al. 2021; Yin et al. 2020; Chen et al. 2019; Micaelli and Storkey 2019), (b) collection-based methods (Chen et al. 2021b; Tang et al. 2023), and (c) our HiDFD. In HiDFD, the teacher-guided generation module employs the teacher network to guide the training of the GAN on limited collected data. Subsequently, the student distillation module closely aligns the features of the student network with those of the teacher network on the hybrid data comprising high-quality synthetic examples and properly inflated collected examples.

Discriminator Overfitting. When trained with very few collected data, the discriminator is prone to be overconfident in determining fake examples, *i.e.*, $\mathbb{E}_{\boldsymbol{x}\sim Q_X}[\mathcal{N}_{\mathrm{D}}(\boldsymbol{x})]$ tends to be 0. As a result, the gradient of $\mathcal{L}_{\mathrm{g}}$ in Eq. (1), which specialized in promoting generator to produce high-quality examples, may become ineffective, namely

$$\nabla_{\mathcal{N}_{\mathrm{G}}}\mathbb{E}_{\boldsymbol{x}\sim Q_X}[\log(1-\mathcal{N}_{\mathrm{D}}(\boldsymbol{x}))]=\mathbb{E}_{\boldsymbol{x}\sim Q_X}\left[-\frac{\nabla_{\mathcal{N}_{\mathrm{G}}}\mathcal{N}_{\mathrm{D}}(\boldsymbol{x})}{1-\mathcal{N}_{\mathrm{D}}(\boldsymbol{x})}\right]\approx 0, \tag{6}$$

as the parameters of $\mathcal{N}_{\mathrm{D}}$ and $\mathcal{N}_{\mathrm{G}}$ are independent of each other, and Eq. (6) is proved by (Arjovsky and Bottou 2022). Meanwhile, the discriminator also has a classifier that provides valuable category dependencies for the generator by precisely predicting input examples, and thus promoting the generator to generate classifiable examples. However, multi-class classification is more challenging than binary determination of true and fake. Given very few collected examples, the discriminator is difficult to learn powerful representations for its classifier to achieve precise classification.

Imbalanced Generator Learning. Given the optimal classifier Ψ_{D}^*[1] of the discriminator, optimizing the generator to produce the classifiable examples[2] is equivalent to

$$\max_{\mathcal{N}_{\mathrm{G}}}[\mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}\log(\frac{p(\boldsymbol{x},y)}{q(\boldsymbol{x},y)})]\Rightarrow\min_{\mathcal{N}_{\mathrm{G}}}\mathrm{KL}(Q_{X,Y}\|P_{X,Y})\,, \tag{7}$$

where KL represents the Kullback-Leibler divergence, and the proof of Eq. (7) is provided in Appendix. The above Eq. (7) indicates that optimizing the generator will force the joint distribution $Q_{X,Y}$ of synthetic data toward the $P_{X,Y}$ of the imbalanced collected data, inevitably resulting in synthetic examples with poor diversity.

In student distillation, we properly inflate the collected examples to construct the hybrid data with a moderate mix proportion α for effectively training the student network.

Teacher-Guided Generation

In this section, we promote GAN to generate high-quality examples by solving its critical issues guided by the teacher network. To mitigate the discriminator overfitting, we design a feature integration mechanism to force the aggregation between the features of both real collected examples and fake synthetic examples. Specifically, we blend the boundaries between real and fake examples to increase the difficulty for the discriminator to accurately discriminate them, and thus preventing the discriminator from overconfidence, *i.e.*,

$$\begin{aligned}\mathcal{L}_{\mathrm{blend}}=\mathbb{E}_{\boldsymbol{x},y\sim P_{X,Y},\hat{\boldsymbol{x}},y\sim Q_{X,Y}}[\mathbb{I}(p>q)(\|\Phi_{\mathrm{T}}(\boldsymbol{x})-\Phi_{\mathrm{D}}(\hat{\boldsymbol{x}})\|_2\\ +\|\Phi_{\mathrm{T}}(\hat{\boldsymbol{x}})-\Phi_{\mathrm{D}}(\boldsymbol{x})\|_2)],\end{aligned} \tag{8}$$

where $\mathbb{I}(p>q)$ is an indicator function to control $\mathcal{L}_{\mathrm{blend}}$ be applied with a probability of q and its value is 1 if $p>q$ and 0 otherwise (p is sampled from $[0,1]$, q=0.7 and it is analyzed in Appendix). Meanwhile, we transfer the expressive features of the teacher network to enhance the representation

[1] $\Psi_{\mathrm{D}}^*(y^+|\,\boldsymbol{x})=\frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})}$, $\Psi_{\mathrm{D}}^*(y^-|\,\boldsymbol{x})=\frac{q(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})}$ (see Appendix).

[2] $\max_{\mathcal{N}_{\mathrm{G}}}[\mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log\Psi_{\mathrm{D}}^*(y^+|\boldsymbol{x})]-\mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log\Psi_{\mathrm{D}}^*(y^-|\boldsymbol{x})]]$.

ability of the discriminator, *i.e.*,

$$\begin{aligned}\mathcal{L}_{\text{trans}} =& \mathbb{E}_{\boldsymbol{x},y\sim P_{X,Y},\hat{\boldsymbol{x}},y\sim Q_{X,Y}}[(\|\Phi_{\text{T}}(\boldsymbol{x})-\Phi_{\text{D}}(\boldsymbol{x})\|_2 \\ &+\|\Phi_{\text{T}}(\hat{\boldsymbol{x}})-\Phi_{\text{D}}(\hat{\boldsymbol{x}})\|_2)].\end{aligned} \tag{9}$$

To alleviate the imbalanced learning of the generator, we regularize the GAN training across all categories. During generator training, we dynamically update the class frequencies $\{n_c^t\}_{c=1}^C$ (C represents the number of categories) at the beginning of iteration t via the following exponential moving average function with a weight $\gamma \in [0,1]$, namely

$$n_c^t = (1-\gamma)n_c^{t-1} + \gamma \bar{n}_c^{t-1}, \tag{10}$$

where $\bar{n}_c^{t-1}$ is the number of synthetic examples belonging to class c in iteration $t-1$, n_c^t is initially set as a constant, and γ=0.5 (analyzed in Appendix). Then, each class frequency $n_c^t \in \{n_c^t\}_{c=1}^C$ is normalized as

$$\hat{n}_c^t = \frac{n_c^t}{\sum_{j=1}^C n_j^t}. \tag{11}$$

Thereafter, the generator is regulated to produce balanced examples by minimizing the loss function:

$$\mathcal{L}_{\text{reg}} = \sum_{c=1}^{C} \frac{\boldsymbol{p}_{\text{T}}^c \log(\boldsymbol{p}_{\text{T}}^c)}{\hat{n}_c^t}, \tag{12}$$

where $\boldsymbol{p}_{\text{T}} = \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\text{SoftMax}(\mathcal{N}_{\text{T}}(\boldsymbol{x}))]$ is the average softmax vector output by the teacher network. The teacher network is well-trained on the original data, so it can precisely predict synthetic examples. In such a case, $\boldsymbol{p}_T^c$ can be regarded as the proportion of examples in category c within the synthetic data. In Eq. (12), the generation of examples in a category c with the lower (or higher) $\boldsymbol{p}_{\text{T}}^c$ is adjusted by the larger (or smaller) $1/\hat{n}_c^t$.

The loss functions of discriminator and generator in our teacher-guided GAN are summarized as

$$\begin{cases} \mathcal{L}_{\text{D}} = \mathcal{L}_{\text{adc\_d}} + \lambda_{\text{d}}(\mathcal{L}_{\text{blend}} + \mathcal{L}_{\text{trans}}), \\ \mathcal{L}_{\text{G}} = \mathcal{L}_{\text{adc\_g}} + \lambda_{\text{g}}\mathcal{L}_{\text{reg}}, \end{cases} \tag{13}$$

where $\mathcal{L}_{\text{adc\_d}}$ and $\mathcal{L}_{\text{adc\_g}}$ are defined in Eq. (1), and the trade-off parameters $\lambda_{\text{d}} > 0$ and $\lambda_{\text{g}} > 0$.

Student Distillation

In the teacher-guided generation module, we successfully trained an effective GAN for generating high-quality synthetic examples, which are then combined with collected examples to construct the hybrid data $\mathcal{D}$ for training the student network. However, directly composing the limited collected examples with numerous synthetic examples will result in a small mix ratio α (*i.e.*, a large distribution gap $\text{TVD}(U,Q)$) to disturb the training of the student network. Therefore, we inflate the collected data via example repeating to enlarge the α from $|\mathcal{D}_{\text{c}}| \,/\, (|\mathcal{D}_{\text{c}}| + |\mathcal{D}_{\text{s}}|)$ to $N \times |\mathcal{D}_{\text{c}}| \,/\, (N \times |\mathcal{D}_{\text{c}}| + |\mathcal{D}_{\text{s}}|)$, where N is the inflation factor. We adopt a moderate inflation factor of $N = \lfloor |\mathcal{D}_{\text{s}}|/|\mathcal{D}_{\text{c}}| \rfloor$ and further details are available in Extended Experiments.

Recent works (Chen et al. 2021b; Tang et al. 2023) indicate that the collected data usually contains many noisy examples, which may mislead the GAN to produce undesired synthetic examples with wrong labels. As a result, these potentially noisy examples will harm the training of the student network, particularly affecting its classifier. In DFKD, the teacher network is well-trained on the original data and possesses an accurate classifier. Recent studies (Tang et al. 2023; Chen et al. 2022) show that the teacher's classifier contains useful category information regarding the original data. Therefore, we share the classifier of the teacher network with the student network. Then, we closely align the feature of the student network with that of the teacher network as follows:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\Phi_{\text{S}}(\boldsymbol{x}) - \Phi_{\text{T}}(\boldsymbol{x})\|_2\right]. \tag{14}$$

By minimizing the $\mathcal{L}_{\text{align}}$, the feature of the student network is closely aligned with that of the teacher network, and the aligned feature is inputted into the shared classifier can produce predictions as accurately as the teacher network. The student network did not use any example labels during the training process, thereby avoiding the negative impact of potentially noisy labels. The whole algorithm of our proposed HiDFD is given in Appendix.

Experiments

In this section, we employ various DNNs commonly utilized in DFKD methods (Chen et al. 2021b; Tang et al. 2023) and conduct intensive experiments on different benchmark datasets to evaluate the effectiveness of our HiDFD.

Datasets and Implementation Details

Original Datasets. We evaluate the effectiveness of our HiDFD on popular datasets, including CIFAR (Krizhevsky 2009), CINIC (Darlow et al. 2018), and TinyImageNet (Le and Yang 2015), which are widely used by existing DFKD methods (Chen et al. 2019, 2021b). Additionally, we also conduct experiments on the large-scale ImageNet (Deng et al. 2009) and the practical medical image dataset HAM (Tschandl, Rosendahl, and Kittler 2018), which are challenging for existing DFKD methods.

Collected Datasets. When using CIFAR and CINIC as the original datasets, we search for examples from ImageNet. With TinyImageNet and ImageNet as the original datasets, we utilize WebVision (Li et al. 2017) as our source of collected data. Moreover, we collect examples from ISIC (Codella et al. 2018) when using HAM as the original dataset. We follow (Chen et al. 2021b) and sample a part of examples from the corresponding dataset as collected data $\mathcal{D}_{\text{c}}$. Here, we define the ratio between the collected data $\mathcal{D}_{\text{c}}$ and original data $\mathcal{D}_{\text{o}}$ as $\rho = |\mathcal{D}_{\text{c}}|/|\mathcal{D}_{\text{o}}|$. We construct small ($\rho$=0.1) and moderate ($\rho$=1.0) collected data for the experiments of collection-based DFKD methods. Notably, the original dataset is solely required for the pre-training of the teacher network. Detailed information regarding these datasets and the corresponding synthesized examples are provided in Appendix.

Implementation Details. All student networks in our HiDFD employ SGD with weight decay as $5\times\ 10^{-4}$ and momentum as 0.9 as the optimizer. The student networks are trained over 240 epochs with a learning rate of 0.05, which

Dataset	Arch	ACC^T	ACC^S	Generation-Based						Collection-Based							
				DAFL	DDAD	DI	PRE	CMI	SSNet	DeGAN		DFND		KD^3		HiDFD (ours)	
										ρ=0.1	ρ=1.0	ρ=0.1	ρ=1.0	ρ=0.1	ρ=1.0	ρ=0.1	ρ=1.0
CIFAR10	♢	95.70	95.20	92.22	93.08	93.26	93.25	94.84	**95.39**	90.39	91.95	48.82	85.82	65.70	93.37	94.74	95.11
	□	94.07	92.69	86.92	90.85	85.27	91.82	88.49	92.00	87.52	90.37	48.65	89.22	48.93	91.49	92.28	**93.14**
	△	95.70	92.69	83.36	89.76	90.24	91.53	86.63	92.03	86.40	89.69	49.48	90.60	65.10	93.05	92.90	**93.76**
CIFAR100	♢	78.05	77.10	74.47	73.64	61.32	74.19	77.04	77.41	53.20	62.94	21.45	64.73	26.96	72.90	76.93	**78.35**
	□	74.53	72.28	65.36	68.33	60.00	70.34	59.70	71.16	53.97	61.80	23.48	63.90	21.27	71.44	71.26	**74.18**
	△	78.05	72.28	45.28	68.59	61.07	67.49	61.80	72.38	46.82	56.44	23.86	64.54	25.25	72.46	73.44	**75.65**
CINIC	♢	86.62	85.09	60.54	80.10	78.57	77.56	78.47	83.47	57.59	76.78	24.53	80.94	39.35	82.68	85.62	**86.68**
	□	84.22	83.28	59.08	77.90	68.90	65.38	74.99	79.63	54.36	76.11	29.53	77.41	29.88	78.18	81.92	**82.27**
	△	86.62	83.28	44.62	77.63	59.52	63.23	75.46	80.30	54.43	74.40	33.40	79.33	71.57	80.28	81.90	**82.88**
Tiny-ImageNet	♢	66.44	64.87	52.20	59.84	6.98	50.15	64.01	64.04	25.74	49.11	26.36	60.09	20.26	63.63	65.96	**66.61**
	□	62.34	61.55	53.89	42.25	1.22	45.92	17.73	57.82	23.13	44.65	25.39	58.47	24.26	61.06	60.46	**62.69**
	△	66.44	61.55	52.46	44.20	2.27	47.22	20.57	59.16	21.09	48.12	25.53	58.18	27.37	61.98	61.67	**65.27**
HAM	♢	81.18	79.64	32.05	44.68	62.79	63.20	67.34	74.52	34.75	64.43	27.55	62.59	64.10	68.44	77.08	**81.52**
ImageNet	♢	73.27	67.00	1.92	1.46	1.14	1.60	1.84	5.74	22.28	43.96	28.99	45.66	35.02	55.05	65.36	**66.89**

Table 1: Accuracies (in %) of student networks trained by various methods on six image classification datasets. The columns "ACC^T" and "ACC^S" report the accuracies yielded by the teacher network and student network trained on the full original data, respectively. The best and the second-best results are highlighted in **bold** and underlined, respectively. The notations ♢, □, and △ represent the teacher-student pairs ResNet34-ResNet18, ResNet34-VGG13, and VGG16-VGG13, respectively.

is sequentially divided by 10 at the 150th, 180th, and 210th epochs. Meanwhile, the generator and discriminator in GAN utilize Adam for optimization with learning rates 1×10^{-4} and 4×10^{-4}, respectively, and both of them are trained over 500 epochs. Additionally, the hyper-parameters in Eq. (13) are configured as $\lambda_{\mathrm{d}} = 0.1$ and $\lambda_{\mathrm{g}} = 0.1$.

Experiments on Benchmark Datasets

In this section, we conduct comprehensive experiments on various benchmark datasets to evaluate the performance of our proposed HiDFD against SOTA generation-based (Chen et al. 2019; Zhao et al. 2022b; Yin et al. 2020; Binici et al. 2022; Fang et al. 2021; Yu et al. 2023) and collection-based (Addepalli et al. 2020; Chen et al. 2021b; Tang et al. 2023) DFKD methods. These methods are reproduced by using their official source codes.

Tab. 1 reports the results of the compared methods and our proposed HiDFD. Firstly, our proposed HiDFD using only a small quantity of collected examples (ρ=0.1) achieves comparable performance with those trained on the full original data. Secondly, when trained on the modestly sized collected data (ρ=1.0), our proposed HiDFD significantly outperforms compared methods on most tasks, especially on the challenging HAM and ImageNet. Thirdly, those generation-based methods, which utilize generative models to produce training examples without the supervision of real examples, tend to perform unsatisfactorily due to the deficiencies in their synthetic examples. These results demonstrate that our proposed HiDFD can train robust student networks by effectively generating training examples from limited real-world examples and properly utilizing all realistic examples.

Ablation Studies & Parametric Sensitivities

In this section, we evaluate the effectiveness of our method with a small collected data (ρ=0.1), where CIFAR and ImageNet serve as the original and collected datasets, respectively. Moreover, ResNet34 and ResNet18 are used as the teacher network and student network, respectively.

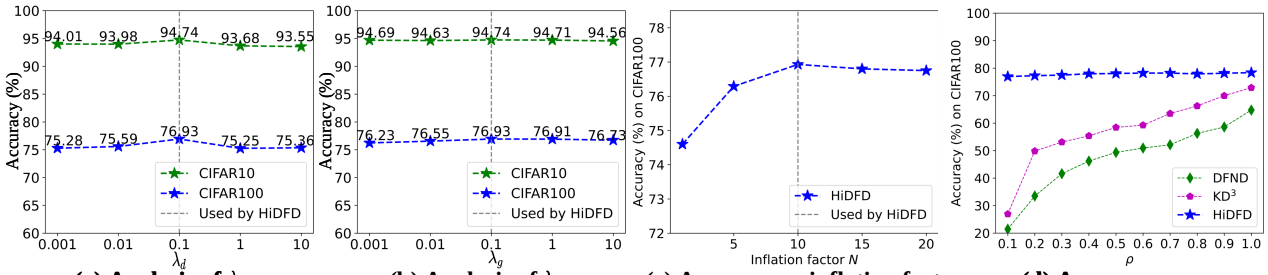
Ablation Studies. We evaluate three key operations ($\mathcal{L}_{\mathrm{blend}}$, $\mathcal{L}_{\mathrm{trans}}$, and $\mathcal{L}_{\mathrm{reg}}$) in teacher-guided generation and the classifier-sharing-based strategy in the student distillation. The experimental results are reported in Tab. 2, and the contributions of these components are analyzed as follows:

Type	Algorithm	CIFAR10	CIFAR100
Teacher-Guided Generation	w/o $\mathcal{L}_{\mathrm{blend}}$	92.87 (↓1.87)	74.18 (↓2.75)
	w/o $\mathcal{L}_{\mathrm{trans}}$	91.86 (↓2.88)	74.40 (↓2.53)
	w/o $\mathcal{L}_{\mathrm{reg}}$	92.76 (↓1.98)	73.95 (↓2.98)
	w/o $\mathcal{L}_{\mathrm{blend}}$, $\mathcal{L}_{\mathrm{trans}}$	91.10 (↓3.64)	71.02 (↓5.91)
	w/o $\mathcal{L}_{\mathrm{blend}}$, $\mathcal{L}_{\mathrm{reg}}$	90.77 (↓3.97)	71.83 (↓5.10)
	w/o $\mathcal{L}_{\mathrm{trans}}$, $\mathcal{L}_{\mathrm{reg}}$	91.42 (↓3.32)	72.10 (↓4.83)
	w/o $\mathcal{L}_{\mathrm{blend}}$, $\mathcal{L}_{\mathrm{trans}}$, $\mathcal{L}_{\mathrm{reg}}$	89.55 (↓5.19)	70.25 (↓6.68)
Student Distillation	OFAKD (Hao et al. 2023)	92.88 (↓1.86)	70.86 (↓6.07)
	VKD (Hinton, Vinyals, and Dean 2015)	92.69 (↓2.05)	66.96 (↓9.97)
	SemcKD (Chen et al. 2021a)	93.49 (↓1.25)	70.93 (↓6.00)
	CC (Peng et al. 2019)	92.63 (↓2.11)	69.52 (↓7.41)
	DKD (Zhao et al. 2022a)	92.95 (↓1.79)	68.25 (↓8.68)
	RKD (Park et al. 2019)	92.40 (↓2.34)	70.53 (↓6.40)
	CATKD (Guo et al. 2023)	92.49 (↓2.25)	68.69 (↓8.24)
	NKD (Yang et al. 2023)	93.26 (↓1.48)	65.31 (↓11.62)
	HiDFD (ours)	94.74	76.93

Table 2: Accuracies (in %) of ablation studies.

1) Teacher-Guided Generation. The feature blending $\mathcal{L}_{\mathrm{blend}}$ in Eq. (8) and feature transferring $\mathcal{L}_{\mathrm{trans}}$ in Eq. (9) for preventing the overfitting of discriminator and enhancing its representation ability. Meanwhile, the generator regulation $\mathcal{L}_{\mathrm{reg}}$ in Eq. (12) is also essential for maintaining the balanced training of the generator. Therefore, the omission of any components among them leads to a noticeable reduction in the performance of the student network. Particularly, training the student network only on synthetic examples without any guidance from the teacher network results in the poorest performance (as shown in the term "w/o $\mathcal{L}_{\mathrm{blend}}$, $\mathcal{L}_{\mathrm{trans}}$, $\mathcal{L}_{\mathrm{reg}}$"). These results indicate the importance of these operations for robust GAN training with limited collected examples, thereby generating high-quality examples for training reliable student networks.

2) Student Distillation. We examine the impact of replacing the classifier-sharing-based feature alignment with traditional KD methods (Hinton, Vinyals, and Dean 2015; Chen et al. 2021a). Both the student networks are trained on the hybrid data composed of collected and synthetic examples. We can find that the student networks trained by these meth-



(a) Analysis of λ_d **(b) Analysis of λ_g** **(c) Accuracy vs inflation factor** **(d) Accuracy vs ρ**

Figure 2: Parametric sensitivities of (a) λ_{d} and (b) λ_{g} in Eq. (13). Accuracies (in %) of the student networks trained with collected data with (c) varying inflation factors and (d) various quantities.

ods generally achieve suboptimal performance due to their inability to effectively handle the potentially noisy examples among the hybrid data. These results highlight the suitability of our training strategy for reliable student networks in the data-free distillation scenarios.

Parametric Sensitivity. There are two tuning parameters in our HiDFD, including λ_{d} and λ_{g} in Eq. (13). To analyze the sensitivities, we individually vary each parameter while keeping the others constant during training. The accuracies of the corresponding student networks are shown in Fig. 2(a) and Fig. 2(b). Despite the large fluctuations in these parameters, where $\lambda_{\mathrm{d}}, \lambda_{\mathrm{g}} \in \{0.001, 0.01, 0.1, 1, 10\}$, the accuracy curve of the student network remains relatively stable. These results indicate the robustness of our HiDFD against parameter variations. Additionally, the student network achieved the best performance when $\lambda_{\mathrm{d}} = \lambda_{\mathrm{g}} = 0.1$, so we adopted such parameter configuration in our method.

Extended Experiments

Experiments with Various Backbones. We evaluate our HiDFD across many widely used teacher-student pairs to assess its adaptability to different networks. The results are shown in Tab. 3, we can observe that our HiDFD consistently achieves satisfactory performance across different teacher-student pairs, where both the trained students perform comparably to those trained on the original data.

Experiments with Varying Inflation Factors. We report the accuracies of the student networks trained on collected data with various inflation factors in Fig. 2(c). The student network performs better with increasing N, and the best accuracy is observed at N=10. Furthermore, excessive inflation may reduce the diversity brought by synthetic data, so that the student network encounters performance degradation when N>10. Therefore, we adopt a moderate inflation factor of $N = \lfloor |\mathcal{D}_{\mathrm{s}}|/|\mathcal{D}_{\mathrm{c}}| \rfloor$. These experiments demonstrate that appropriately inflating the collected examples, which are crucial for reducing the distribution gap between synthetic and collected data, can effectively improve the performance of the student network.

Experiments on Collected Data with Various Data Quantities. We explore the impact of varying the volume of collected data on the performance of student networks, with ρ values ranging from 0.1 to 1. As shown in Fig. 2(d), student networks trained by the compared collection-based DFKD methods (Chen et al. 2021b; Tang et al. 2023) tend to underperform with small values of ρ. Conversely, our HiDFD consistently achieves satisfactory performance across a spectrum of ρ values. These results further demonstrate the effectiveness of our HiDFD in training reliable student networks leveraging limited collected data.

Dataset	Teacher	Student	ACC$^{\mathrm{S}}$	HiDFD	↑ or ↓
CIFAR10	ResNet32×4	ResNet110	93.37	95.04	↑1.67
	ResNet32×4	ShuffleNet	93.23	93.62	↑0.39
	ResNet110×2	ResNet116	93.21	94.83	↑1.62
	ResNet110×2	WRN40×2	94.86	95.35	↑0.49
CIFAR100	ResNet32×4	ResNet110	74.31	75.69	↑1.38
	ResNet32×4	ShuffleNet	72.60	75.03	↑2.43
	ResNet110×2	ResNet116	74.46	74.49	↑0.03
	ResNet110×2	WRN40×2	76.31	75.65	↓0.66

Table 3: Accuracies (in %) of various networks trained by our method (ρ=1.0).

Conclusion

In this paper, we proposed a new data-free distillation approach termed HiDFD to train the student networks on the hybrid data comprising high-quality synthetic examples and scarce collected examples, which well meets practical requirements. Our investigation reveals that bridging the distribution gap between the hybrid and synthetic data is crucial for training reliable student networks, and it implies that the quality of synthetic data and the weight of collected data are two key factors in reducing this gap. This observation inspired us to propose a novel hybrid distillation framework, where the teacher-guided generation module can effectively generate high-quality synthetic examples from the limited collected data by leveraging the teacher network to guide the GAN training process, and the student distillation module properly enhances the influence of collected examples within the hybrid data by inflating their frequency. Consequently, we can naturally define a classifier-sharing-based feature alignment to distill the student network, and we achieve state-of-the-art performance using significantly fewer examples than existing methods. The limitations and broader impacts of our HiDFD are discussed in Appendix.

Acknowledgments

This research is supported by NSF of China (Nos: 62336003, 12371510), and NSF for Distinguished Young Scholar of Jiangsu Province (No: BK20220080).

References

Addepalli, S.; Nayak, G. K.; Chakraborty, A.; and Radhakrishnan, V. B. 2020. Degan: Data-enriching gan for retrieving representative samples from a trained classifier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3130–3137.

Arjovsky, M.; and Bottou, L. 2022. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.

Binici, K.; Aggarwal, S.; Pham, N. T.; Leman, K.; and Mitra, T. 2022. Robust and resource-efficient data-free knowledge distillation by generative pseudo replay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6089–6096.

Chen, D.; Mei, J.-P.; Zhang, H.; Wang, C.; Feng, Y.; and Chen, C. 2022. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11933–11942.

Chen, D.; Mei, J.-P.; Zhang, Y.; Wang, C.; Wang, Z.; Feng, Y.; and Chen, C. 2021a. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7028–7036.

Chen, H.; Guo, T.; Xu, C.; Li, W.; Xu, C.; Xu, C.; and Wang, Y. 2021b. Learning student networks in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6428–6437.

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 (ICCV)*, 3514–3522.

Codella, N. C.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration. In *15th IEEE International Symposium on Biomedical Imaging (ISBI)*, 168–172.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine (SPM)*, 35(1): 53–65.

Cui, K.; Yu, Y.; Zhan, F.; Liao, S.; Lu, S.; and Xing, E. P. 2023. Kd-dlgan: Data limited image generation via knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3872–3882.

Darlow, L. N.; Crowley, E. J.; Antoniou, A.; and Storkey, A. J. 2018. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

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

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.

Guo, Z.; Yan, H.; Li, H.; and Lin, X. 2023. Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11868–11877.

Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; and Xu, C. 2023. One-for-all: bridge the gap btween heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 79570–79582.

Hao, Z.; Guo, J.; Wang, C.; Tang, Y.; Wu, H.; Hu, H.; Han, K.; and Xu, C. 2024. Data-efficient large vision models through sequential autoregression. In *Forty-first International Conference on Machine Learning (ICML)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 6840–6851.

Hou, L.; Cao, Q.; Shen, H.; Pan, S.; Li, X.; and Cheng, X. 2022. Conditional gans with auxiliary discriminative classifier. In *International Conference on Machine Learning (ICML)*, 8888–8902. PMLR.

Hou, L.; Yuan, Z.; Huang, L.; Shen, H.; Cheng, X.; and Wang, C. 2021. Slimmable generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7746–7753.

Huang, J.; Cui, K.; Guan, D.; Xiao, A.; Zhan, F.; Lu, S.; Liao, S.; and Xing, E. 2022. Masked generative adversarial networks are data-efficient generation learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 2154–2167.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021. Deceive d: Adaptive pseudo augmentation for gan training with limited data. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 21655–21667.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. *Master's Thesis, University of Tront*.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Van Gool, L. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1504–1512.

Mei, K.; and Patel, V. 2023. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9117–9125.

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

Miles, R.; and Mikolajczyk, K. 2024. Understanding the role of the projector in knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4233–4241.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3967–3976.

Peng, B.; Jin, X.; Liu, J.; Li, D.; Wu, Y.; Liu, Y.; Zhou, S.; and Zhang, Z. 2019. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5007–5016.

Rangwani, H.; Mopuri, K. R.; and Babu, R. V. 2021. Class balancing gan with a classifier in the loop. In *Uncertainty in Artificial Intelligence (UAI)*, 1618–1627. PMLR.

Steerneman, T. 1983. On the total variation and Hellinger distance between signed measures; an application to product measures. *Proceedings of the American Mathematical Society (AMS)*, 88(4): 684–688.

Tang, J.; Chen, S.; Niu, G.; Sugiyama, M.; and Gong, C. 2023. Distribution shift matters for knowledge distillation with webly collected images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tran, M.-T.; Le, T.; Le, X.-M.; Harandi, M.; Tran, Q. H.; and Phung, D. 2024. Nayer: Noisy layer data generation for efficient and effective data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23860–23869.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1): 1–9.

Wang, Y.; Qian, B.; Liu, H.; Rui, Y.; and Wang, M. 2024a. Unpacking the gap box against data-free knowledge distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Wang, Y.; Yang, D.; Chen, Z.; Liu, Y.; Liu, S.; Zhang, W.; Zhang, L.; and Qi, L. 2024b. De-confounded data-free knowledge distillation for handling distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12615–12625.

Wang, Y.; Zhang, J.; and Wang, Y. 2024. Do generated data always help contrastive learning? In *International Conference on Learning Representations (ICLR)*.

Yang, Z.; Zeng, A.; Li, Z.; Zhang, T.; Yuan, C.; and Li, Y. 2023. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. *arXiv preprint arXiv:2303.13005*.

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 (CVPR)*, 8715–8724.

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 (CVPR)*, 24266–24275.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022a. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11953–11962.

Zhao, H.; Sun, X.; Dong, J.; Manic, M.; Zhou, H.; and Yu, H. 2022b. Dual discriminator adversarial distillation for data-free model compression. *International Journal of Machine Learning and Cybernetics (IJMLC)*, 13(5): 1213–1230.

Zhao, S.; Song, J.; and Ermon, S. 2019. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5885–5892.

Supplementary Materials for Hybrid Data-Free Knowledge Distillation

Jialiang Tang[1,2,3], **Shuo Chen**[4*], **Chen Gong**[5*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, China
[3]Jiangsu Key Laboratory of Image and Video Understanding for Social Security, China
[4]Center for Advanced Intelligence Project, RIKEN, Japan
[5]Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
tangjialiang@njust.edu.cn, shuo.chen.ya@riken.jp, goodgongchen@gmail.com

Proofs

The Optimal Classifier of Discriminator

For a fixed generator, the optimal classifier Ψ_{D}^* of the discriminator in our employed Auxiliary Discriminative Classifier GAN (ADCGAN) can be formatted as follows:

$$\begin{aligned}\Psi_{\mathrm{D}}^*\left(y^+ \mid \boldsymbol{x}\right) &= \frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})},\\ \Psi_{\mathrm{D}}^*\left(y^- \mid \boldsymbol{x}\right) &= \frac{q(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})}.\end{aligned} \tag{1}$$

Proof.

$$\begin{aligned}&\max_{\Psi_{\mathrm{D}}} \mathbb{E}_{\boldsymbol{x},y\sim P_{X,Y}}[\log \Psi_{\mathrm{D}}(y^+|\boldsymbol{x})]+\\ &\qquad \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log \Psi_{\mathrm{D}}(y^-|\boldsymbol{x})] \qquad (2)\\ &\Rightarrow \max_{\Psi_{\mathrm{D}}} \mathbb{E}_{\boldsymbol{x},y\sim P^m_{X,Y}}[\log \Psi_{\mathrm{D}}(y|\boldsymbol{x})],\end{aligned}$$

with $p^m(\boldsymbol{x},y^+) = \frac{1}{2}p(\boldsymbol{x},y)$, $p^m(\boldsymbol{x},y^-) = \frac{1}{2}q(\boldsymbol{x},y)$, and $p^m(\boldsymbol{x}) = \sum_y p^m(\boldsymbol{x},y) = \frac{1}{2}p(\boldsymbol{x}) + \frac{1}{2}q(\boldsymbol{x})$.

$$\Rightarrow \max_{\Psi_{\mathrm{D}}} \mathbb{E}_{\boldsymbol{x}\sim P^m_X}\mathbb{E}_{y\sim P^m_{Y|X}}[\log \Psi_{\mathrm{D}}(y|\boldsymbol{x})] \tag{3}$$

$$\Rightarrow \min_{\Psi_{\mathrm{D}}} \mathbb{E}_{\boldsymbol{x}\sim P^m_X}\mathbb{E}_{y\sim P^m_{Y|X}}[-\log \Psi_{\mathrm{D}}(y|\boldsymbol{x})] \tag{4}$$

$$\Rightarrow \min_{\Psi_{\mathrm{D}}} \mathbb{E}_{\boldsymbol{x}\sim P^m_X}[H(p^m(y|\boldsymbol{x}))+\mathrm{KL}(p^m(y|\boldsymbol{x})\|\Psi_{\mathrm{D}}(y|\boldsymbol{x}))] \tag{5}$$

$$\Rightarrow \Psi_{\mathrm{D}}^*(y|\boldsymbol{x}) = \arg\min_{\Psi_{\mathrm{D}}} \mathrm{KL}(p^m(y|\boldsymbol{x})\|\Psi_{\mathrm{D}}(y|\boldsymbol{x})) \tag{6}$$

$$= p^m(y|\boldsymbol{x}) = \frac{p^m(\boldsymbol{x},y)}{p^m(\boldsymbol{x})} \tag{7}$$

Therefore, the optimal discriminative classifier of ADCGAN has the form of $\Psi_{\mathrm{D}}^*(y^+|\boldsymbol{x}) = \frac{p^m(\boldsymbol{x},y^+)}{p^m(\boldsymbol{x})} = \frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})}$ and $\Psi_{\mathrm{D}}^*(y^-|\boldsymbol{x}) = \frac{p^m(\boldsymbol{x},y^-)}{p^m(\boldsymbol{x})} = \frac{q(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})}$ that conclude the proof. □

Proof of Eq. (7)

Given the optimal classifier of the discriminator, at the equilibrium point, encouraging the generator to produce easily classifiable examples of our employed ADCGAN is equivalent to

$$\begin{aligned}&\max_{\mathcal{N}_{\mathrm{G}}}[\mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log \Psi_{\mathrm{D}}^*(y^+ \mid \boldsymbol{x})]\\ &-\mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}[\log \Psi_{\mathrm{D}}^*(y^- \mid \boldsymbol{x})]] \qquad (8)\\ &\Rightarrow \min_{\mathcal{N}_{\mathrm{G}}} \mathrm{KL}\left(Q_{X,Y}\|P_{X,Y}\right),\end{aligned}$$

Proof.

$$\max_{\mathcal{N}_{\mathrm{G}}} \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}\left[\log \Psi_{\mathrm{D}}^*(y^+|\boldsymbol{x})\right] \tag{9}$$

$$-\mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}\left[\log \Psi_{\mathrm{D}}^*(y^-|\boldsymbol{x})\right] \tag{10}$$

$$\Rightarrow \max_{\mathcal{N}_{\mathrm{G}}} \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}\left[\log \frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})}\right] \tag{11}$$

$$-\mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}\left[\log \frac{q(\boldsymbol{x},y)}{p(\boldsymbol{x})+q(\boldsymbol{x})}\right] \tag{12}$$

$$\Rightarrow \min_{\mathcal{N}_{\mathrm{G}}} \mathbb{E}_{\boldsymbol{x},y\sim Q_{X,Y}}\left[\log \frac{q(\boldsymbol{x},y)}{p(\boldsymbol{x},y)}\right] \tag{13}$$

$$\Rightarrow \min_{\mathcal{N}_{\mathrm{G}}} \mathrm{KL}(Q_{X,Y}\|P_{X,Y}) \tag{14}$$

□

Detailed Information of Benchmark Datasets

Comprehensive Dataset Overview

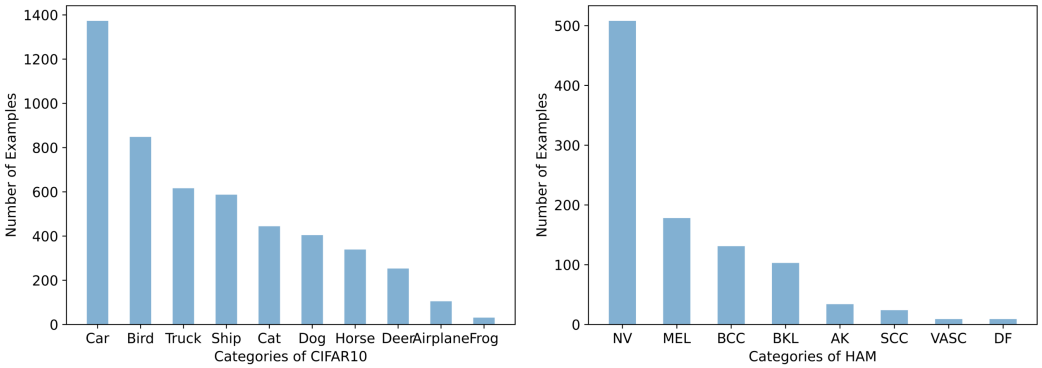
In Table A-1, we introduce the critical information of the benchmark datasets used in our experiment, including the image size, the number of categories, and the number of images in training and test sets. First, we can observe that the large-scale ImageNet contains a large number of examples, when using it as the original data, it is difficult to collect sufficient examples (more than 1 million) from the real-world. Therefore, it is essential to explore an efficient data-free distillation method that requires only a small amount of collected data. Second, a series of benchmark datasets contain 224×224 sized high-resolution images, the generation-based methods that without rely on real-world data are hard to generate high-quality examples to provide effective information for training the student network.

Moreover, Figure A-1 shows the number of collected examples in each category when the original data is the natural image datasets CIFAR10 and practical medical image

*Corresponding authors: Chen Gong, Shuo Chen.

Table A-1: Details of the benchmark datasets used in our experiment, the items with the prefix "#" denotes the number of the corresponding item.

Dataset	Type	Image size	#classes	#train	#test
CIFAR10	Original data	32×32	10	50,000	10,000
CIFAR100	Original data	32×32	100	50,000	10,000
CINIC	Original data	32×32	10	90,000	90,000
TinyImageNet	Original data	64×64	200	100,000	10,000
HAM	Original data	224×224	7	8,000	2,000
ISIC	Collected data	224×224	8	20,000	5,000
ImageNet	Original&collected data	224×224	1,000	1,281,167	50,000
WebVision	Collected data	224×224	1,000	980,449	N/A



(a) CIFAR10 **(b) HAM**

Figure A-1: The number of examples per category in the collected data ($\rho = 0.1$), where the original data is (a) CIFAR10 and (b) HAM, respectively.

dataset HAM, respectively. We can observe that the collected examples exhibit imbalanced class distribution, with several categories accounting for the majority of examples and other categories containing only a few examples. Therefore, our proposed data-free distillation method is very practical.

Visualization of Synthetic Examples

In Figure A-2, we show synthetic examples produced by the GAN trained on limited collected data using our teacher-guided generation module. Specifically, we generate synthetic instances for four original datasets, including CIFAR10, CINIC, TinyImageNet, and HAM. Despite these original datasets having significantly different image sizes (ranging from 32×32 to 224×224), the corresponding synthetic examples consistently exhibit high quality. These visual results demonstrate the effectiveness of our approach in leveraging the teacher network to address critical issues of GAN training on limited collected data. By doing so, our method can train reliable student networks on abundant high-quality synthetic examples.

Algorithm

The detailed training algorithm of our proposed HiDFD is summarized in Algorithm A-1. Our HiDFD contains two primary modules to train a reliable student network only using a small number of collected examples, *i.e.*, the teacher-guided generation and student distillation. In the teacher-guided generation, the GAN is trained on the limited collected data under the guidance of the teacher network, where the critical issues, *i.e.*, overfitting of the discriminator and imbalanced learning of the generator, are effectively resolved. In the student network, the collected examples are properly inflated via repeating and combined with sufficient high-quality synthetic examples to construct the hybrid data. Then, the reliable student network can naturally train on the hybrid data via the effective classifier-sharing-based feature alignment strategy.

Limitations and Broader Impacts

Limitations

The proposed method can train reliable student networks using very few collected examples. Compared with previous methods, we have reduced the data requirement by 99%, making our method suitable for practical applications. In general, the effectiveness of the proposed method depends on the quality and representativeness of collected data to some extent. If this collected data does not sufficiently represent the broader dataset or contains biases, the generated synthetic examples and the trained student network may inherit these flaws. In practice, collecting fewer representative examples in real-world applications is relatively easy. Therefore, we believe that these limitations can be overcome well.

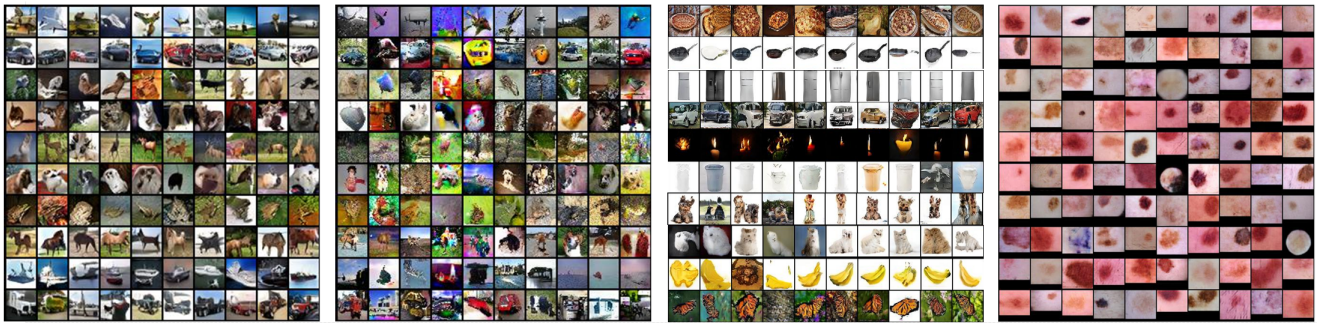


Figure A-2: Visualization of synthetic examples for the original tasks, including (a) CIFAR10, (b) CINIC, (c) TinyImageNet, and (d) HAM.

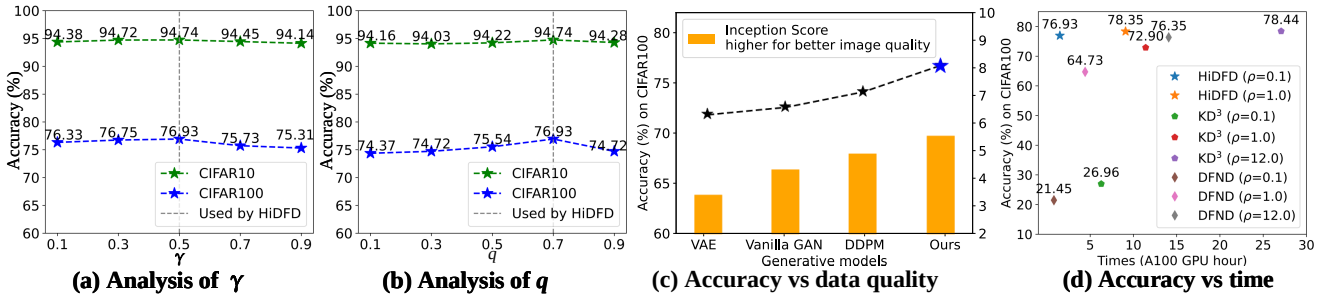


Figure A-3: Parametric sensitivities of (a) γ in Eq. (12) and (b) q in Eq. (9). Accuracies (in %) of the student networks trained with (c) synthetic examples generated by various generative models. (d) shows the accuracies and training times of various DFKD methods.

Broader Impacts

This paper proposed a novel data-free knowledge distillation method called HiDFD, which can train a compact and reliable student network using very few collected examples. In general, the proposed HiDFD could have the following positive impacts: 1) HiDFD eliminates the need for original training data required by traditional knowledge distillation methods, so it can help preserve data privacy for users; 2) HiDFD effectively compresses the large pre-trained models (*i.e.*, the teacher networks) into smaller and faster models (*i.e.*, the student networks) that are resource-efficient and suitable for deployment on devices with limited capabilities; 3) HiDFD focuses on the classification tasks, which underpin many practical downstream tasks like object detection and segmentation, suggesting its wide applicability; and 4) HiDFD is compatible with different DNNs (*e.g.*, ResNet and VGG).

Although HiDFD has few negative social impacts, when it compresses the large models of many Artificial Intelligence (AI) technologies and enables these compressed models in practical applications, the proposed HiDFD can be used for good and also for harm, depending on human intent. This actually falls into the general ethical debate on whether AI is good or not.

In conclusion, we believe our proposed work can be beneficial to society since many important real-world applications need compact and reliable models that stand to benefit from HiDFD when the available real-world data is limited.

Additional Experiments

Additional Parametric Sensitivities

There are also two tuning parameters q and γ in Eq. (9) and γ in Eq. (12), respectively. To analyze their sensitivities, we individually vary each parameter while keeping the others constant during training. The accuracies of the corresponding student networks are shown in Figure A-3. Despite the large fluctuations in these parameters, where q, $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, the accuracy curve of the student network remains relatively stable. These results indicate the robustness of our HiDFD against parameter variations. Additionally, the student network achieved the best performance when $q = 0.7$ and $\gamma = 0.5$, so we adopted such parameter configuration in our method.

Experiments on Various Collected Data

We compare the quality of synthetic examples produced by different generative models trained with the limited collected data and report the accuracies of the corresponding student networks in Figure A-3(c). Here, the higher-quality synthetic data consistently promotes a better student network, which demonstrates that improving the quality of synthetic examples can effectively improve the performance of

Algorithm A-1: Hybrid Data-Free Distillation

Require: Pre-trained teacher network $\mathcal{N}_{\mathrm{T}}$, limted collected data $\mathcal{D}_{\mathrm{c}}$.

Initialize the discriminator $\mathcal{N}_{\mathrm{D}}$ and generator $\mathcal{N}_{\mathrm{G}}$ in GAN, and the small student network $\mathcal{N}_{\mathrm{S}}$;
Module 1: Teacher-Guided Generation
repeat
    Sample the noise-label pair $(\boldsymbol{z}, y)$ to generate synthetic example $\mathcal{N}_{\mathrm{G}}(\boldsymbol{z}, y)$;
    Mitigate the overfitting of discriminator $\mathcal{N}_{\mathrm{D}}$ via blending $\mathcal{L}_{\mathrm{blend}}$ and transferring $\mathcal{L}_{\mathrm{trans}}$ operations in feature integration;
    Calculate the class frequency $\{\hat{n}_c\}_{c=1}^{C}$ of synthetic examples;
    Regularize the training of generator $\mathcal{N}_{\mathrm{G}}$ across all categories;
    Optimize the GAN via Adam;
until convergence
Module 1: Student Distillation
Generate abundant high-quality synthetic examples and inflate collected examples to construct the hybrid data $\mathcal{D}$;
repeat
    Sample the example $\boldsymbol{x} \in \mathcal{D}$ and input it into the teacher network $\mathcal{N}_{\mathrm{T}}$ and student network $\mathcal{N}_{\mathrm{S}}$ to obtain the features $\Phi_{\mathrm{T}}(\boldsymbol{x})$ and $\Phi_{\mathrm{S}}(\boldsymbol{x})$, respectively.
    Align the features $\Phi_{\mathrm{T}}(\boldsymbol{x})$ and $\Phi_{\mathrm{S}}(\boldsymbol{x})$ via $\mathcal{L}_{\mathrm{align}}$;
    Optimize the student network $\mathcal{N}_{\mathrm{S}}$ via SGD;
until convergence

Ensure: Lightweight student network $\mathcal{N}_{\mathrm{S}}$.

the student network.

Moreover, we compared the training time of our method with other collection-based DFKD methods, as depicted in Figure A-3(d). We can observe that our method can train a student network with satisfactory performance within a few hours on an A100 GPU. These results further demonstrate the effectiveness of our HiDFD in training reliable student networks leveraging limited collected data.