De-confounded Data-free Knowledge Distillation for Handling Distribution Shifts

Yuzheng Wang[1*] Dingkang Yang[1*] Zhaoyu Chen[1] Yang Liu[1] Siao Liu[1]
Wenqiang Zhang[2] Lihua Zhang[1†] Lizhe Qi[1,2,3†]
[1]Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology, Fudan University
[2]Engineering Research Center of AI & Robotics, Ministry of Education, Academy for Engineering & Technology, Fudan University
[3]Green Ecological Smart Technology School-Enterprise Joint Research Center
{yzwang20, dkyang20}@fudan.edu.cn

Abstract

Data-Free Knowledge Distillation (DFKD) is a promising task to train high-performance small models to enhance actual deployment without relying on the original training data. Existing methods commonly avoid relying on private data by utilizing synthetic or sampled data. However, a long-overlooked issue is that the severe distribution shifts between their substitution and original data, which manifests as huge differences in the quality of images and class proportions. The harmful shifts are essentially the confounder that significantly causes performance bottlenecks. To tackle the issue, this paper proposes a novel perspective with causal inference to disentangle the student models from the impact of such shifts. By designing a customized causal graph, we first reveal the causalities among the variables in the DFKD task. Subsequently, we propose a Knowledge Distillation Causal Intervention (KDCI) framework based on the backdoor adjustment to de-confound the confounder. KDCI can be flexibly combined with most existing state-of-the-art baselines. Experiments in combination with six representative DFKD methods demonstrate the effectiveness of our KDCI, which can obviously help existing methods under almost all settings, e.g., improving the baseline by up to 15.54% accuracy on the CIFAR-100 dataset.

1. Introduction

Deep Neural Networks (DNNs), as a powerful and reliable tool, are increasingly expected to be applied to practical artificial intelligence scenes [1–7]. Despite significant progress, good performance of deep learning models is often inseparable from large-scale models [8–13] and high-quality original training data [14–20]. The dependencies hinder the deployment of this technology on mobile devices and data privacy scenes. Therefore, model compression and data-free technology have become the key to breaking through the bottleneck. To this end, Lopes *et al.* [21] propose the Data-Free Knowledge Distillation (DFKD) task. In this process, knowledge is transferred from the cumbersome model to a small model that is more suitable for deployment [22–24] without relying on the original training data. As a result, DFKD has received more attention due to its convenience and wide application.

Since the original training data is not available for privacy or other reasons [25], the key is how to supplement the new training data, *i.e.*, the substitution data. Based on the source of the substitution data, almost all existing DFKD methods can be divided into generation-based and sampling-based methods. Despite the impressive improvements achieved by these DFKD methods through complex loss stacking [26, 27] and knowledge distillation strategies [28, 29], the trained students still suffer from distribution shifts between the substitution and original data, which has long been overlooked. First, the quality of the synthetic or sampled images significantly differs from the original. Besides, for generation-based methods, the synthetic data

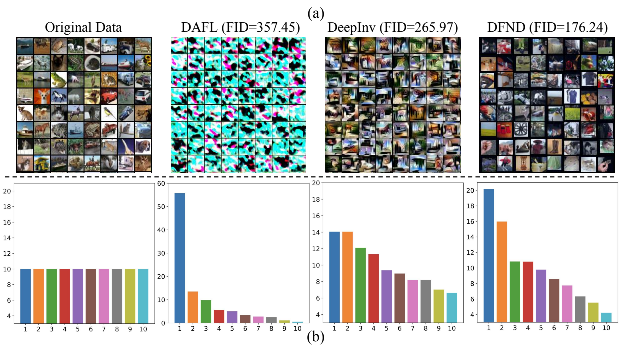

Figure 1. Diagrams of the distribution shifts between the original and substitute data for existing DFKD methods on CIFAR-10. (a) represents the random visualization and FID score of the synthetic data by DAFL, DeepInv, and sampled by DFND. (b) indicates the proportion of sample numbers in various classes (%) of the original and substitute data.

---
*Equal contribution. †Corresponding authors.

relies on the teacher's guidance, and it is easier to synthesize the class familiar to the generator. For sampling-based methods, the sampled data entirely depends on the teacher's preference for various classes. These protocols make the preference of the teacher model inevitably affect class proportions and also lead to distribution shifts. Such shifts confound the student learning process. For example, if a pre-trained teacher model is not familiar with a specific class $\mathcal{A}$, *i.e.*, it is difficult to obtain high confidence, resulting in fewer synthetic or sampled data belonging to $\mathcal{A}$. For the class balance, the teacher tends to classify ambiguous and indistinguishable data into $\mathcal{A}$, leading to the distribution shifts [30]. Relying on these data, the student is inevitably confused with the original testing data with the different distributions.

More intrigued, we select three DFKD methods (DAFL [31], DeepInv [26], and DFND [32]) and perform a toy experiment on the CIFAR-10 [33]. This toy experiment aims to show the distribution shifts between the substitution and original data. These methods include generation with generators (DAFL), generation through teacher model inversion (DeepInv), and sampling based on teacher preferences (DFND). We use the original data as a comparison benchmark and compare them from two aspects: the quality of images and class proportions. The results are shown in Figure 1. In Figure 1a, we randomly visualize the original data, the substitution data of DAFL, DeepInv, and DFND, and calculate the Fréchet Inception Distance (FID, lower is better) [34], a metric widely used to evaluate the quality of images. The substitution and original data are different for the data distribution domain. In Figure 1b, we test the class proportions (the substitution data is based on teacher pseudo-labels). A prominent result is that the classes of the substitution data are unbalanced due to teacher preferences, which greatly differ from the original data. These observations confirm the distribution shifts between the substitution and original data, confounding the student model.

Based on these observations, we attempt to introduce a new perspective with causal inference to handle the distribution shifts. During the application of theoretical causal inference [35] to the DFKD task, the challenges lie in describing and designing plausible causal effects and identifying and compensating for biased student learning on the substitution data with shifts. To this end, this paper attempts to address the challenges by drawing on instinctive human causalities [36] to find causal relationships among the variables in the DFKD task and optimize the biased student training process. We first disentangle the causalities and customize the causal graph according to the properties of the variables in the DFKD task. Based on this, we explore the causal paths from the substitution inputs $\boldsymbol{X}$ to the student predictions $\boldsymbol{S}$. Then, we propose a simple yet effective Knowledge Distillation Causal Intervention (KDCI) framework to achieve de-confounded DFKD and use the do-calculus $P(\boldsymbol{S}|do(\boldsymbol{X}))$ to calculate the actual causal effect, instead of classic likelihood $P(\boldsymbol{S}|\boldsymbol{X})$ without considering the shifts. KDCI can be easily combined with existing methods and use the backdoor adjustment [37] to de-confound and alleviate the impact of the shifts. Experiments on KDCI combined with six representative DFKD methods demonstrate its strong positive effect on the existing DFKD pipeline. Specifically, the primary contributions and experiments are summarized below:

- To our best knowledge, we are the first to alleviate the dilemma of the distribution shifts in the DFKD task from a causality-based perspective. Such shifts are regarded as the harmful confounder, which leads the student to learn misleading knowledge.
- We propose a KDCI framework to restrain the detrimental effect caused by the confounder and attempt to achieve the de-confounded distillation process. Besides, KDCI can be easily and flexibly combined with existing generation-based or sampling-based DFKD paradigms.
- Extensive experiments on the combination with six DFKD methods show that our KDCI can bring consistent and significant improvements to existing state-of-the-art models. Particularly, it improves the accuracy of the DeepInv [26] by up to 15.54% on the CIFAR-100 dataset.

2. Related Work

Data-Free Knowledge Distillation. Data-free knowledge distillation is a promising task to train small models while avoiding leakage of original training data [21, 38]. The critical point is how to supplement substitution data [39–42]. The existing methods are mainly divided into three types: Generative Adversarial Networks (GANs) generation [29, 31], teacher-based model inversion generation [26, 27], and unlabeled data sampling [28, 32, 43]. Chen *et al.* [31] introduce the generator into the DFKD task and improve teachers' familiarity with generating data. Fang *et al.* [29] propose feature sharing to simplify the generation process. To better generation quality, Yin *et al.* [26] explore the prior knowledge of the data. Fang *et al.* [27] introduce contrastive learning to enhance student performance. Chen *et al.* [32] and Fang *et al.* [28] select wild data and out-of-domain (OOD) data to reduce generation costs. Despite the promising performance, a long-overlooked issue is data distribution shifts, *i.e.*, the distribution bias of the student's training data and the original data is a confounder that significantly causes performance bottlenecks.

Causal Inference. Causal inference is a theory-oriented tool that seeks actual effects in a specific phenomenon [35], which has been studied and followed by diverse fields such as economics [44] and psychology [45] communities. The mainstream causal inference studies applied to neural information processing consist of two aspects: intervention

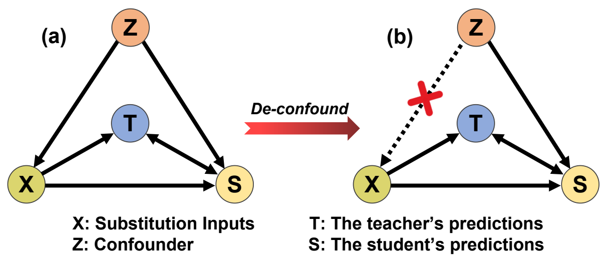

Figure 2. The causal graph. (a) The existing methods ignore distribution shifts. (b) The shifts are alleviated by causal inference.

[46–51] and counterfactuals [52–55]. Intervention is a technique for manipulating the original data distribution to reveal causal effects [37]. Counterfactual describes the imagined results generated by factual variables when treated differently [56]. Benefiting from the strong potential of causal inference to decouple spurious correlations among variables, it is gradually adopted to improve the performance of models for different downstream tasks, such as visual question answering [57], emotion recognition [58], and scene graph generation [52]. In contrast, to our best knowledge, this is the first work to identify the distribution shifts in the DFKD task through the causal intervention and alleviate the confounding effect caused by the shifts.

3. Methodology

3.1. Causal Graph of DFKD Task

First, we customize the causal graph according to the properties of the variables in the DFKD task. Specifically, the teacher is pre-trained with original training data, which is not disturbed by distribution shifts. For the student, it uses the substitution data to train while testing on the original data. The data distribution shifts indicate that it will be disturbed by the biased data [30]. During the distillation process, the teacher and student are fed the same substitution data. In this case, the student's predictions are constrained to learn the teacher's predictions. Following the same graphical notation as [59] for clarity and interpretability, we denote the variables with the notes $\mathcal{N}$ and construct the direct causal effects with the links $\mathcal{E}$. From Figure 2, there are four variables involved in the DFKD causal graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, which includes the substitution inputs $\boldsymbol{X}$, the confounder $\boldsymbol{Z}$, the teacher's predictions $\boldsymbol{T}$, and the student's predictions $\boldsymbol{S}$. In particular, our causal graph is applicable to almost all existing DFKD methods so that it can be used as a general framework. The details of the causal relationships are described as follows.

$\boldsymbol{Z} \rightarrow \boldsymbol{X}$. Existing DFKD methods rely on teacher predictions to supplement substitution data. For the generation-based methods, the generator is guided by the teacher and more inclined to synthesize data that is easier to synthesize [26, 27, 29, 31]. For the sampling-based methods, the data that the teacher is most [32] or least [28] familiar with is sampled. On the one hand, these synthetic or sampled data are always class-imbalanced. On the other hand, these sources of substitution data rely heavily on the teacher, so they are highly volatile and vulnerable to teacher preferences. These issues cause the distribution shifts between the original and substitution data. The shifts are treated as the harmful confounder $\boldsymbol{Z}$ [56]. On this basis, the confounder $\boldsymbol{Z}$ causes the substitution data $\boldsymbol{X}$ to be biased compared to the original data, *i.e.*, $\boldsymbol{Z} \rightarrow \boldsymbol{X}$.

$\boldsymbol{Z} \rightarrow \boldsymbol{S}$. Due to the distribution shifts between the substitution and original data, the student trained on the substitution data tends to produce and exhibit biased predictions during the testing stage. The detrimental confounder $\boldsymbol{Z}$ confounds and affects the student's training via the causal link $\boldsymbol{Z} \rightarrow \boldsymbol{S}$, which causes the performance bottleneck.

$\boldsymbol{X} \rightarrow \boldsymbol{T}/\boldsymbol{S}$ & $\boldsymbol{T} \leftrightarrow \boldsymbol{S}$. As with existing DFKD methods, both teacher and student make predictions on the substitution data $\boldsymbol{X}$ simultaneously. By constraining their prediction distributions, the student's parameters are updated for optimization. In our DFKD causal graph, the prediction processes of the teacher and student are represented as $\boldsymbol{X} \rightarrow \boldsymbol{T}$ and $\boldsymbol{X} \rightarrow \boldsymbol{S}$. The link $\boldsymbol{T} \leftrightarrow \boldsymbol{S}$ reflects the interaction causal effect between these two predictions during knowledge distillation. Through these paths, the student can learn consistent knowledge from its teacher.

According to the causal theory [35], the confounder $\boldsymbol{Z}$ as a common cause directly or indirectly impacts the substitution inputs $\boldsymbol{X}$ and the student's predictions $\boldsymbol{S}$ simultaneously. The knowledge transfer process from $\boldsymbol{T}$ to $\boldsymbol{S}$ increases the student's familiarity with these substitution data. However, the confounder $\boldsymbol{Z}$ causes $\boldsymbol{X}$ to shift the original data distribution, leading to impure knowledge, which adversely affects student performance. The detrimental effects follow the backdoor causal path as $\boldsymbol{X} \leftarrow \boldsymbol{Z} \rightarrow \boldsymbol{S}$.

3.2. Causal Intervention via Backdoor Adjustment

In the existing DFKD task, the pre-trained teacher model is fixed while the student model is learnable. As shown in Figure 2a, existing methods rely on the likelihood estimation of the student model as $P(\boldsymbol{S}|\boldsymbol{X})$. The knowledge transfer process is expressed as:

$$P(\boldsymbol{S}|\boldsymbol{X}) = \sum_{\boldsymbol{z}} P(\boldsymbol{S}|\boldsymbol{X}, KD\langle \boldsymbol{T} = f_T(\boldsymbol{X}), \boldsymbol{S} = f_S(\boldsymbol{X}, \boldsymbol{z})\rangle) P(\boldsymbol{z}|\boldsymbol{X}), \tag{1}$$

where $KD\langle , \rangle$ represents the knowledge distillation process between $\boldsymbol{T}$ and $\boldsymbol{S}$. $f_T(\cdot)$ and $f_S(\cdot)$ represent the teacher model and the student model. The confounder $\boldsymbol{Z}$ introduces the data distribution shifts via $P(\boldsymbol{z}|\boldsymbol{X})$, which makes the knowledge learned by the student impure. To get rid of the confounding effect caused by $\boldsymbol{Z}$, an intuitive idea is changing inputs $\boldsymbol{X}$ to overcome the data distribution shifts and make $\boldsymbol{X}$ unaffected by $\boldsymbol{Z}$, *i.e.*, we have to use the data from the same distribution with the original training set as the student's training data. However, it is not pos-

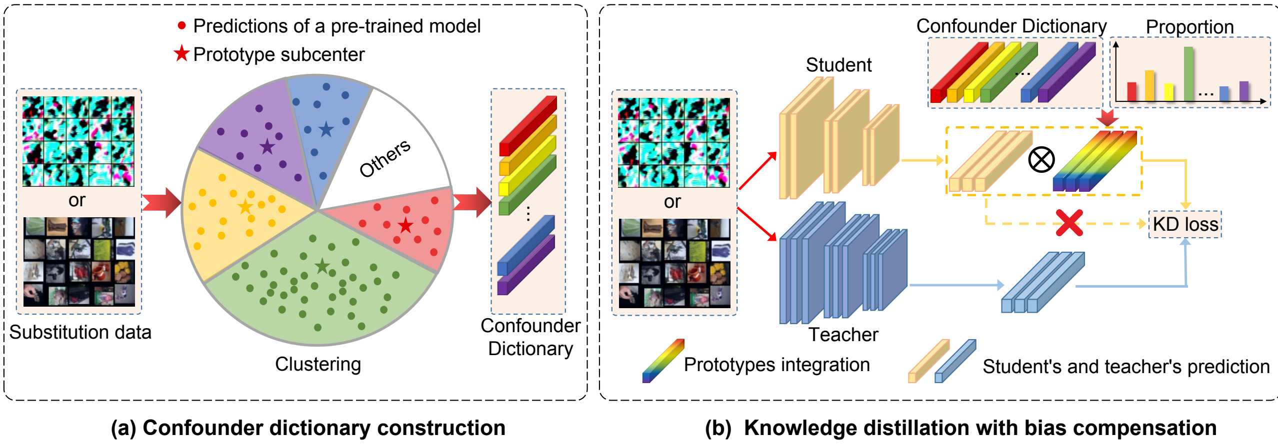

Figure 3. The overview of our KDCI. In stage (a), all substitution data is fed a pre-trained model to explore the prior knowledge and construct the confounder dictionary. In stage (b), the prototype integration is built by the confounder dictionary and is used to compensate for biased student predictions. The distillation loss is calculated between the teacher's prediction and the student's compensated prediction.

sible under the setting of the DFKD task. To tackle this issue, we introduce the backdoor adjustment [35] to construct causal intervention $P(\boldsymbol{S}|do(\boldsymbol{X}))$ and block the backdoor path between $\boldsymbol{X}$ and $\boldsymbol{S}$ via $\boldsymbol{Z}$. As a theoretical operation, implementing backdoor adjustment can be viewed as measuring the distribution shifts by estimating the average causal effect based on the class proportions. By compensating for shifted student predictions, we alleviate the shift issue and suppress the disturbance of the confounder $\boldsymbol{Z}$. In this case, the causal path from $\boldsymbol{Z}$ to $\boldsymbol{X}$ is cut-off in Figure 2b. The student learns pure knowledge with causal intervention $P(\boldsymbol{S}|do(\boldsymbol{X}))$ rather than original biased likelihood $P(\boldsymbol{S}|\boldsymbol{X})$. This process can be expressed as:

$$P(\boldsymbol{S}|do(\boldsymbol{X}))=\sum_{\boldsymbol{z}}P(\boldsymbol{S}|\boldsymbol{X},KD\langle\boldsymbol{T}=f_T(\boldsymbol{X}),\boldsymbol{S}=f_S(\boldsymbol{X},\boldsymbol{z})\rangle)P(\boldsymbol{z}), \tag{2}$$

where $\boldsymbol{X}$ is no longer disturbed by $\boldsymbol{z}$ since causal intervention forces $\boldsymbol{X}$ to integrate each $\boldsymbol{z}$ fairly into the predictions of $\boldsymbol{S}$, according to the corresponding prior $P(\boldsymbol{z})$.

3.3. De-confounded DFKD with KDCI

To de-confound the DFKD task, we propose a Knowledge Distillation Causal Intervention (KDCI) framework to alleviate the distribution shift issue. The overview of KDCI is shown in Figure 3, which contains two stages: *confounder dictionary construction* and *knowledge distillation with bias compensation*. First, after obtaining the substitution data and before training the student, we model the prior knowledge of these substitution data through the prototype clustering algorithm to obtain an intervention-driven confounder dictionary. Then, the biased student predictions are compensated based on the subcenters and proportions. Notably, for a general DFKD pipeline, our framework can be easily combined with other methods. The implementation of KDCI is as follows.

Confounder Dictionary Construction. Since the substitution data has no ground-truth information and the actual classes are ambiguous, we define a confounder dictionary $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N]$ to explore the prior knowledge of these data. N is a hyperparameter representing the confounder size and $\boldsymbol{z}_i \in \mathbb{R}^d$ is a single prototype. The prior knowledge implies the potential shifts and the differentiation information of class proportions. From Figure 3a, all substitution data is fed to an experienced pre-trained model (*e.g.*, the teacher model itself) to obtain the prediction feature set $M = \{m_j \in \mathbb{R}^d\}_{j=1}^{N_m}$, where N_m is the number of the substitution data. We employ the K-Means++ with principle component analysis as the prototype clustering algorithm. After clustering, each $\boldsymbol{z}_i$ represents a prototype feature cluster, and the prototype subcenter is put into the confounder dictionary as a prototype representation. The feature cluster is denoted as $\sum_{k=1}^{N_i} m_k^i$ and the subcenter is denoted as $\boldsymbol{z}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} m_k^i$, where N_i is the number of the prediction features in i-th cluster. Therefore, the prototype proportion can be calculated as $P(\boldsymbol{z}_i) = N_i/N_m$.

Knowledge Distillation with Bias Compensation. After confounder dictionary construction, we approximate a theoretical causal inference by the confounder dictionary and prototype proportions to compensate for biased student predictions to learn pure knowledge, as shown in Figure 3b. In practice, the calculation of $P(\boldsymbol{S}|do(\boldsymbol{X}))$ requires multiple forward passes of all $\boldsymbol{z}$ resulting in expensive computational costs. To simplify the above process, we apply the Normalized Weighted Geometric Mean (NWGM) [60] and approximate the Eq. (2) as:

$$P(\boldsymbol{S}|do(\boldsymbol{X}))\approx P(\boldsymbol{S}|\boldsymbol{X}, KD\langle f_T(\boldsymbol{X}), \sum_{\boldsymbol{z}} f_S(\boldsymbol{X}, \boldsymbol{z})P(\boldsymbol{z})\rangle). \tag{3}$$

During the knowledge transfer process, the update of student model parameters depends on the difference in predictions between the teacher and student, *e.g.*, calculating the Kullback-Leibler (KL) divergence as $f_S \leftarrow \eta\nabla_s KL(\boldsymbol{T}, \boldsymbol{S})$, where η denotes learning rate, and ∇_s denotes the gradient.

Considering the distribution shift of training data, we introduce the prepared prior information of the cofounder dictionary to optimize the above process. Based on this, the student predictions after compensation are represented as the integration of the biased predictions and the prior information as: $P(\boldsymbol{S}|do(\boldsymbol{X})) = \phi(f_S(\boldsymbol{X}), F(\boldsymbol{z}))$, where $\phi(\cdot)$ is a practically simple yet empirically powerful addition fusion strategy. The prior information $F(\boldsymbol{z})$ is calculated as:

$$F(\boldsymbol{z}) = \sum_{i=1}^{N} \lambda_i \boldsymbol{z}_i P(\boldsymbol{z}_i), \tag{4}$$

where λ_i is a weight coefficient that measures the importance of each prototype subcenter $\boldsymbol{z}_i$. $P(\boldsymbol{z}_i)$ is the proportion of data in the *i*-th cluster. Here, we design an implementation of λ_i with the additive attention as:

$$\lambda_i = softmax(\boldsymbol{W}_t \cdot Tanh(\boldsymbol{W}_q f_S(\boldsymbol{X}) + \boldsymbol{W}_k \boldsymbol{z}_i)), \tag{5}$$

where $\boldsymbol{W}_t \in \mathbb{R}^{d_n \times 1}$, $\boldsymbol{W}_q \in \mathbb{R}^{d_n \times d_h}$, and $\boldsymbol{W}_k \in \mathbb{R}^{d_n \times d}$ are learnable mapping matrices.

4. Experiments

4.1. Datasets and Models

Datasets. We evaluate the proposed framework on widely used classification datasets: CIFAR-10 [33], CIFAR-100 [33], Tiny-ImageNet [61], and ImageNet [16]. CIFAR-10 and CIFAR-100 contain 50,000 training samples and 10,000 testing samples of 32×32 resolution. Tiny-ImageNet contains 100,000 training samples, 10,000 validating samples, and 10,000 testing samples of 64×64 resolution. ImageNet contains 1000 classes with 1.28 million training samples and 50,000 validating samples of 224×224 resolution.

Models. We test the performance of various DFKD methods on several network architectures, including resnet [1], vgg [62], and wide resnet [63]. For CIFAR-10 and CIFAR-100, we use the pre-trained teacher models from CMI [27], unify the teacher models among all methods, and set up five teacher-student backbone combinations following existing settings [27–29]. For Tiny-ImageNet, we train a renset-34 teacher model without the mixup data augmentation [64]. And the student utilizes the renset-18 as its backbone. For ImageNet, we choose the same pre-trained resnet-50 teacher model with [65] for all baseline methods.

4.2. Method Zoo

To comprehensively verify the effectiveness of KDCI, we select representative DFKD methods, including generation-based and sampling-based methods. The generation-based methods spend extra computing costs to obtain substitute data by generative adversarial networks and teacher inversion, including DAFL [31], Fast [29], CMI [27], and DeepInv [39]. The sampling-based methods use unlabeled data as the substitute data, including Mosaick [28] and DFND [32]. For DAFL, Fast, and DeepInv, we follow the same settings as their original papers. For CMI, due to the unpublished pre-inversion data, we choose the base version of CMI, which leads to the performance slightly lower than that reported in the original paper. For Mosaick and DFND, we sample 600k unlabeled data in ImageNet [16] for CIFAR and Tiny-ImageNet, and 600k unlabeled data in Flicker1M dataset for ImageNet. Due to the image quality, the reported performance of Mosaick is slightly better than the original paper. The implementation details and loss functions of all the above methods are shown in ***Supplementary Sec.7***.

4.3. Confounder Setup

We use a pre-trained model to obtain the prediction feature set M. By default, the pre-trained model is the teacher itself, which is trained on original data. Each prediction feature m is extracted from the logits output of the last layer, and the hidden dimension d is equal to the number of classes. By default, the number of clusters N is the same across different datasets. For the substitution data in a mini-batch of model inversion [27, 39], the number of clusters N is 32. For the synthetic mini-batch from GANs [29, 31], the number of clusters N is 8. For the unlabeled substitution data in sampling methods [28, 32], the number of clusters N is 128. Due to different training paradigms, the way KDCI is combined with these methods is different. For the generation-based process, the generator and student models are updated alternately. We use a mini-batch of synthetic training data to construct the cofounder dictionary, and the dictionary will be updated as the generator is updated. For the sampling-based process, unlabeled data only needs to be filtered once. We build the confounder dictionary once before distillation. Pseudocode for the above processes and other training settings are shown in ***Supplementary Sec.1***.

4.4. Performance Comparison

To verify the proposed KDCI framework, we compare the original version and their KDCI-based version.

Results on CIFAR-10 and CIFAR-100. The results in Table 1 show the following vital observations. (**i**) KDCI consistently improves the performance of existing methods on all baselines across two datasets. (**ii**) For CIFAR-10, although the original students' performance is already close to their teachers', KDCI still provides promising gains (mostly 1%-2% improvement) for students by eliminating the harmful impact of confounder. For some baselines with poor results, KDCI brings significant improvement, *e.g.*, up to 8.85%† for DAFL. (**iii**) For CIFAR-100, KDCI can significantly improve various SOTA methods (about 3%-5% improvement on average). Under some settings, KDCI improves the original methods with slightly lower performance to competitive performance, *e.g.*, 15.54%‡ and

Table 1. The accuracy (%) on CIFAR-10 and CIFAR-100 about baseline methods vs. their KDCI-based version. **T.backbone** and **S.backbone** represent the backbones of the teacher and student. **Teacher** and **Student** refer to scratch training on original data. The improved results are marked in **bold**. $\{\dagger, \ddagger, \natural, \sharp\}$ denote the provenance mentioned in the analysis.

Dataset	**CIFAR-10**					**CIFAR-100**				
T.backbone **S.backbone**	resnet-34 resnet-18	vgg-11 resnet-18	wrn-40-2 wrn-16-1	wrn-40-2 wrn-40-1	wrn-40-2 wrn-16-2	resnet-34 resnet-18	vgg-11 resnet-18	wrn-40-2 wrn-16-1	wrn-40-2 wrn-40-1	wrn-40-2 wrn-16-2
Teacher **Student**	95.70 95.20	92.25 95.20	94.87 91.12	94.87 93.94	94.87 93.95	78.05 77.10	71.32 77.10	75.83 65.31	75.83 72.19	75.83 73.56
DAFL DAFL+**KDCI**	92.22 **92.62**	81.10 **81.31**	65.71† **74.56**†	81.33 **82.91**	81.55 **82.65**	74.47 **74.51**	54.16 **58.79**	20.88$^\sharp$ **31.75**$^\sharp$	42.83 **46.16**	43.70 **48.48**
Fast Fast+**KDCI**	94.05 **94.56**	90.53 **91.16**	89.29 **89.62**	92.51 **93.09**	92.45 **92.85**	74.34 **75.10**	67.44 **68.97**	54.02 **54.69**	63.91 **67.09**	65.12 **68.12**
CMI CMI+**KDCI**	94.24 **94.43**	91.24 **91.28**	89.16 **89.52**	91.93 **92.84**	92.00 **92.73**	74.64 **75.07**	66.68 **69.07**	55.28 **57.19**	63.44 **67.47**	64.22 **67.68**
DeepInv DeepInv+**KDCI**	93.26 **93.67**	90.36 **91.42**	83.04 **83.47**	86.85 **89.32**	89.72 **91.06**	61.32$^\natural$ **74.59**$^\natural$	54.13‡ **69.67**‡	53.77 **55.22**	61.33 **62.13**	61.34 **65.90**
Mosaick Mosaick+**KDCI**	95.27 **95.43**	91.69 **92.36**	90.03 **92.25**	93.28 **94.45**	92.94 **94.20**	75.91 **77.06**	71.58 **71.86**	59.32 **62.03**	66.61 **72.19**	67.36 **72.39**
DFND DFND+**KDCI**	95.36 **95.44**	91.86 **92.54**	90.26 **92.47**	93.33 **94.43**	93.11 **94.43**	74.42 **77.09**	68.97 **72.12**	59.02 **66.37**	69.39 **74.20**	69.85 **74.52**

Table 2. The accuracy (%) on Tiny-ImageNet dataset. The teacher uses resnet-34, and the student uses resnet-18 as the backbones. The teacher achieves an accuracy of 52.74%. The GPU time indicates the training time of one epoch on a single RTX 3090 GPU.

Method	**Accuracy (%)**	**GPU time**	**Memory-Usage**
Fast Fast+**KDCI**	28.79 **38.23** (+9.44)	101.67s **104.43s** (+2.71%)	5745M **5748M** (+0.05%)
DeepInv DeepInv+**KDCI**	20.68 **34.84** (+14.16)	255.26s **258.51s** (+1.27%)	3312M **3316M** (+0.12%)
DFND DFND+**KDCI**	42.64 **49.54** (+6.90)	129.16s **133.42s** (+3.30%)	4196M **4198M** (+0.05%)

13.27%$^\natural$ for DeepInv & 10.87%$^\sharp$ for DAFL. These strong gains demonstrate that KDCI can compensate for biased student predictions to learn pure knowledge by constructing prior knowledge on the substitution data whose data distribution differs from the original data distribution. (**iv**) We notice a small increase for KDCI-based Fast & CMI. The reasonable explanation is that they extract prior knowledge about the substitution data by accessing the statistics in the teacher's Batch Normalization layers [66], which implicitly apply the likelihood estimation and weaken our causal intervention. (**v**) Besides, we are pleasantly surprised to find that the students trained by sampling-based methods (*e.g.*, Mosaick & DFND) can slightly outperform the teacher in some settings (*e.g.*, vgg-11→resnet-18), both the original and KDCI-based versions. Both Mosaick and DFND utilize the unlabeled data. With the additional rich semantic knowledge, more students outperform their teachers with the help of KDCI framework.

Results on Tiny-ImageNet. For the Tiny-ImageNet, we conduct experiments with Fast, DeepInv, and DFND. The results are shown in Table 2. With the help of KDCI, the accuracy of the three methods is increased by 9.44%, 14.16%, and 6.90%, respectively. The Tiny-ImageNet dataset contains richer semantic information, which helps construct more expressive confounders and facilitates KDCI to bring more sufficient gains. Besides, we test and show the additional calculation and memory overhead. The overhead introduced by KDCI mainly comes from the confounder matrix. The additional overhead can be almost negligible since only a simple clustering algorithm is used.

Table 3. The accuracy (%) on ImageNet dataset. "→" denotes the teacher's (left) and student's (right) backbone pair.

Settings	resnet-50 → resnet-18	resnet-50 → mobilenetv2
Fast Fast+**KDCI**	53.45 **58.24** (+4.79)	43.02 **50.12** (+7.10)
DeepInv DeepInv+**KDCI**	51.36 **55.27** (+3.91)	40.25 **46.24** (+5.99)
DFND DFND+**KDCI**	42.82 **51.26** (+8.44)	16.03 **34.32** (+18.29)

Results on ImageNet. For the ImageNet, we conduct two backbone combinations with three baseline methods. The results are shown in Table 3. The generation-based methods (Fast & DeepInv) have to train 1,000 generators (one generator for one class). We speculate that a possible reason why KDCI has smaller gains for these two generation-based methods is that 'one generator for one class' may alleviate the distribution shifts issue to a certain extent and thereby weaken the effect of causal intervention. In comparison, the gain of KDCI for DFND is higher. Overall, from the experimental results of ImageNet, the positive impact of KDCI on students is also consistent. These results further validate the effectiveness of our method.

Combining the performance on the above datasets, we conclude that KDCI can provide more significant help on more complex datasets (*e.g.*, ImageNet & Tiny-ImageNet

Table 4. Ablation studies about the prior information $F(\boldsymbol{z}) = \sum_{i=1}^{N} \lambda_i \boldsymbol{z}_i P(\boldsymbol{z}_i)$ in Eq. (4). The results include (1) original $F(\boldsymbol{z})$, (2) random weight coefficient λ_i, (3) random confounder dictionary $\boldsymbol{z}_i$, and (4) without (w/o) prototype proportion $P(\boldsymbol{z}_i)$.

Settings	(1) **Original** $F(\boldsymbol{z})$			(2) **Random** λ_i			(3) **Random** $\boldsymbol{z}_i$			(4) **w/o** $P(\boldsymbol{z}_i)$		
Methods	Fast	DeepInv	DFND	Fast	DeepInv	DFND	Fast	DeepInv	DFND	Fast	DeepInv	DFND
CIFAR-10	**94.56**	**93.67**	**95.38**	93.92	91.56	95.28	93.35	91.84	94.94	93.70	92.76	95.11
CIFAR-100	**75.10**	**74.59**	**77.09**	74.79	72.72	76.86	73.76	72.81	76.14	74.60	72.66	76.97

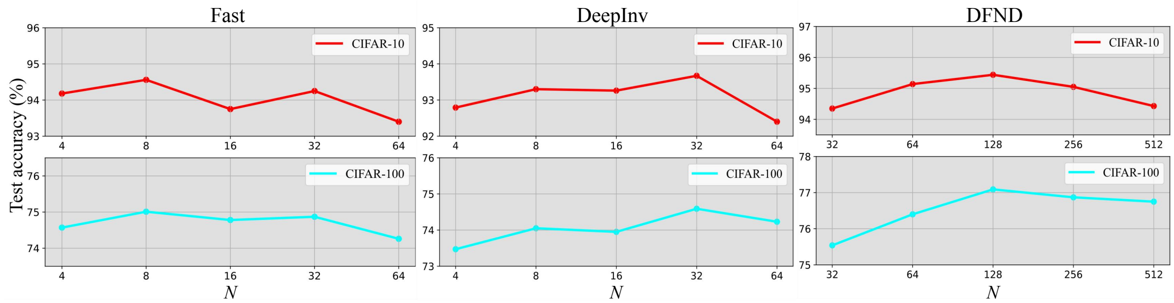

Figure 4. The test accuracy (%) on CIFAR-10 and CIFAR-100 datasets about different confounder dictionary size N. The teacher uses resnet-34, and the student uses resnet-18 as the backbones.

with more classes and various visual effects). More complex datasets are more susceptible to teacher preferences, leading to more severe distribution shifts. Further, the detrimental shifts inevitably lead to biased substitution data compared to the original data. Fortunately, KDCI favorably de-confound the biased student predictions, achieving significant performance improvements.

4.5. Analysis of Prior Information $F(\boldsymbol{z})$

We conduct ablation studies to validate the effectiveness of the components of prior information $F(\boldsymbol{z})$ in Eq. (4) used to compensate students for biased predictions in Table 4. We select three methods (Fast, DeepInv, and DFND) on both CIFAR-10 and CIFAR-100 datasets. The teacher and student use resnet-34 and resnet-18 as their backbones, respectively. Other settings are the same as Table 1.

Necessity of Weight Coefficient λ_i**.** The weight λ_i represents the degree of each confounder. Comparing (1) and (2), the random λ_i causes a decline in performance. Such results indicate that depicting the importance of each confounder is essential to achieve effective causal intervention.

Rationality of Confounder $\boldsymbol{z}_i$**.** The confounder $\boldsymbol{z}_i$ comes from the predicted feature representation of the pre-trained model, which directly implies prior knowledge about the substitution data. Comparing (1) and (3), students using our custom confounder significantly outperform the alternative confounder that are randomly initialized, which proves the validity of extracted prior knowledge.

Impact of Prototype Proportion $P(\boldsymbol{z}_i)$**.** The prototype proportion $P(\boldsymbol{z}_i)$ denotes the frequency of each confounder containing the knowledge of feature proportions. From (1) and (4), the proportion of each confounder plays a vital role in precise intervention implementation.

4.6. Analysis of Confounder Dictionary Z

The confounder dictionary $\boldsymbol{Z}$ is proposed to explore the prior knowledge of the substitution data. We investigate the effectiveness of $\boldsymbol{Z}$ in two perspectives: the confounder prototype size N and the selected pre-trained models. For the size, we select representative methods to test the effect of different N. For the selected pre-trained models, we use the models coming from other datasets with different numbers of classes. We swap the pre-training models on CIFAR-10 and CIFAR-100 to build the confounder and align the feature dimensions through a learnable mapping matrix.

Impact of Confounder Dictionary Size N**.** To justify the size N of the confounder $\boldsymbol{Z}$, we set five sets of N for each method. For Fast and DeepInv, $\boldsymbol{Z}$ comes from a mini-batch synthetic data. For DFND, $\boldsymbol{Z}$ comes from the sampled data. In Figure 4, designing the suitable N for methods that suffer from varying degrees of harmful shifts helps to perform deconfounded training better.

Impact of Confounder Dictionary Sources. Table 5 shows three settings with/without confounder dictionary $\boldsymbol{Z}$. We have two interesting discoveries. (**i**) First, an obvious conclusion is that using $\boldsymbol{Z}$ outperforms the original DFKD methods without $\boldsymbol{Z}$ in almost all settings. Such observations demonstrate the effectiveness of causal intervention. (**ii**) Second, swapping the confounders from CIFAR-10 and CIFAR-100 teacher models brings the performance decrease. For CIFAR-10, the distribution of the substitution data is simple. Simple distributions are over-separated when features are extracted using pre-trained models from complex distributions. We call this phenomenon *over-intervention*. The excessive causal intervention potentially causes the deviation of the confounder itself. For CIFAR-100, the distribution is more complex. The complex distributions are not well approximated when using pre-trained models with less discriminative ability. We call this phe-

Table 5. Ablation studies about the confounder dictionary $\boldsymbol{Z}$. "w/o $\boldsymbol{Z}$" denotes the vanilla version of DFKD methods. "*original* $\boldsymbol{Z}$" denotes the original confounder from the teacher itself. "*other* $\boldsymbol{Z}$" denotes the confounder from another pre-trained model, *i.e.*, swapping the confounder from the pre-training teacher models on CIFAR-10 and CIFAR-100 datasets.

Dataset	**CIFAR-10**						**CIFAR-100**					
Settings	resnet-34 → resnet-18			vgg-11 → resnet-18			resnet-34 → resnet-18			vgg-11 → resnet-18		
$\boldsymbol{Z}$	w/o $\boldsymbol{Z}$	*original* $\boldsymbol{Z}$	*other* $\boldsymbol{Z}$	w/o $\boldsymbol{Z}$	*original* $\boldsymbol{Z}$	*other* $\boldsymbol{Z}$	w/o $\boldsymbol{Z}$	*original* $\boldsymbol{Z}$	*other* $\boldsymbol{Z}$	w/o $\boldsymbol{Z}$	*original* $\boldsymbol{Z}$	*other* $\boldsymbol{Z}$
Fast	94.05	94.56	93.96	90.53	91.16	90.73	74.42	75.10	74.75	67.44	68.97	68.75
DeepInv	93.26	93.67	93.56	90.36	91.42	91.26	61.32	74.59	73.04	54.13	69.67	68.04
DFND	95.36	95.44	95.41	91.86	92.54	92.34	74.34	77.09	76.97	68.97	72.12	71.97

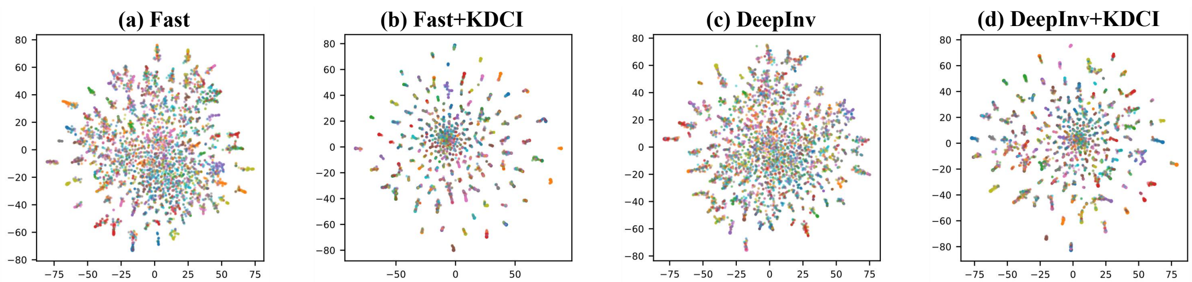

Figure 5. T-SNE results of vanilla and KDCI-based models performance on Tiny-ImageNet dataset. KDCI helps models obtain clearer clustering results, which show its strong positive impact.

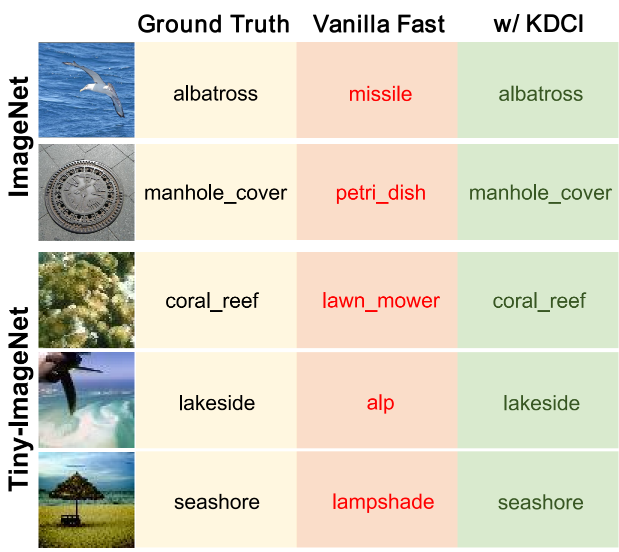

Figure 6. Qualitative results of the vanilla and KDCI-based version on ImageNet and Tiny-ImageNet.

nomenon *under-intervention*. The incomplete causal intervention would lead to gain reduction.

4.7. Qualitative Results

Further, we present qualitative results to further demonstrate the positive gains of KDCI over baseline methods.

Visualization Results. To intuitively show the help of KDCI to existing DFKD methods, we first visualize the student classification results with t-SNE [67] on the Tiny-ImageNet dataset. We reserve 100 classes of validating samples. From Figure 5, the KDCI-based versions (b)&(d) have fewer outliers and clearer clustering effects than the vanilla versions (a)&(c). These phenomena further confirm that our KDCI can well disentangle features from different classes, thus improving existing methods' performance.

Case Study of Causal Intervention. We select representative examples from ImageNet and Tiny-ImageNet datasets to show differences in student predictions before and after the intervention. As shown in Figure 6, KDCI can eliminate the prediction offset caused by some misleading features to a certain extent. For example, students from the vanilla Fast misclassify "albatross" as "missile" or "coral_reef" as "lawn_mower" due to large patches of similar background colour, and misclassify "manhole_cover" as "petri_dish" or "seashore" as "lampshade" due to similar shape. Fortunately, KDCI can repair prediction shifts in the above cases.

5. Conclusion

This paper proposes a novel perspective from causal inference to handle the distribution shifts in the Data-Free Knowledge Distillation (DFKD) task. By customizing the causal graph according to the properties of the variables in the DFKD, we propose a Knowledge Distillation Causal Intervention (KDCI) framework to de-confound the adverse effect caused by the shifts between the substitution and original data. KDCI can be flexibly combined with most existing methods. Numerous experiments prove that KDCI can consistently help existing methods and provide an alternative causal intervention insight.

Acknowledgements

This work is supported by the Shanghai Engineering Research Center of AI & Robotics, Fudan University, China, the Engineering Research Center of AI & Robotics, Ministry of Education, China, and the Green Ecological Smart Technology School-Enterprise Joint Research Center.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 5

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[3] Dingkang Yang, Yang Liu, Can Huang, Mingcheng Li, Xiao Zhao, Yuzheng Wang, Kun Yang, Yan Wang, Peng Zhai, and Lihua Zhang. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Systems*, page 110370, 2023.

[4] Yang Liu, Dingkang Yang, Gaoyun Fang, Yuzheng Wang, Donglai Wei, Mengyang Zhao, Kai Cheng, Jing Liu, and Liang Song. Stochastic video normality network for abnormal event detection in surveillance videos. *Knowledge-Based Systems*, 280:110986, 2023.

[5] Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. Robust emotion recognition in context debiasing. *arXiv preprint arXiv:2403.05963*, 2024.

[6] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1642–1651, 2022.

[7] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1708–1717, 2022. 1

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

[11] Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

[12] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15148–15158, 2022.

[13] Yang Liu, Jing Liu, Kun Yang, Bobo Ju, Siao Liu, Yuzheng Wang, Dingkang Yang, Peng Sun, and Liang Song. Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system. *IEEE Transactions on Industrial Informatics*, 2023. 1

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 1

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 5

[17] Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, et al. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20459–20470, 2023.

[18] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Yang Liu, Siao Liu, Wenqiang Zhang, and Lizhe Qi. Adversarial contrastive distillation with adaptive denoising. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[19] Zuhao Ge, Lizhe Qi, Yuzheng Wang, and Yunquan Sun. Zoom-and-reasoning: Joint foreground zoom and visual-semantic reasoning detection network for aerial images. *IEEE Signal Processing Letters*, 29:2572–2576, 2022.

[20] Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, et al. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23436–23446, 2023. 1

[21] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017. 1, 2

[22] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1

[23] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[24] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Pinxue Guo, Kaixun Jiang, Wenqiang Zhang, and Lizhe Qi. Out of thin

air: Exploring data-free adversarial robustness distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5776–5784, 2024. 1

[25] Paul R Burton, Madeleine J Murtagh, Andy Boyd, James B Williams, Edward S Dove, Susan E Wallace, Anne-Marie Tasse, Julian Little, Rex L Chisholm, Amadou Gaye, et al. Data safe havens in health research and healthcare. *Bioinformatics*, 31(20):3241–3248, 2015. 1

[26] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8715–8724, 2020. 1, 2, 3

[27] Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021. 1, 2, 3, 5

[28] Gongfan Fang, Yifan Bao, Jie Song, Xinchao Wang, Donglin Xie, Chengchao Shen, and Mingli Song. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:11920–11932, 2021. 1, 2, 3, 5

[29] Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 6597–6604, 2022. 1, 2, 3, 5

[30] Kien Do, Thai Hung Le, Dung Nguyen, Dang Nguyen, Haripriya Harikumar, Truyen Tran, Santu Rana, and Svetha Venkatesh. Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:10055–10067, 2022. 2, 3

[31] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3514–3522, 2019. 2, 3, 5

[32] Hanting Chen, Tianyu Guo, Chang Xu, Wenshuo Li, Chunjing Xu, Chao Xu, and Yunhe Wang. Learning student networks in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6428–6437, 2021. 2, 3, 5

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 2

[35] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. 2, 3, 4

[36] Linda J Van Hamme and Edward A Wasserman. Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, 25(2):127–151, 1994. 2

[37] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 3

[38] Paul Micaelli and Amos J Storkey. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 2

[39] Kartikeya Bhardwaj, Naveen Suda, and Radu Marculescu. Dream distillation: A data-independent model compression framework. *arXiv preprint arXiv:1905.07072*, 2019. 2, 5

[40] Liangchen Luo, Mark Sandler, Zi Lin, Andrey Zhmoginov, and Andrew Howard. Large-scale generative data-free distillation. *arXiv preprint arXiv:2012.05578*, 2020.

[41] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 710–711, 2020.

[42] Yuzheng Wang, Zuhao Ge, Zhaoyu Chen, Xian Liu, Chuangjia Ma, Yunquan Sun, and Lizhe Qi. Explicit and implicit knowledge distillation via unlabeled data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[43] Yuzheng Wang, Zhaoyu Chen, Jie Zhang, Dingkang Yang, Zuhao Ge, Yang Liu, Siao Liu, Yunquan Sun, Wenqiang Zhang, and Lizhe Qi. Sampling to distill: Knowledge transfer from open-world data. *arXiv preprint arXiv:2307.16601*, 2023. 2

[44] Hal R Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016. 2

[45] E Michael Foster. Causal inference and developmental psychology. *Developmental Psychology*, 46(6):1454, 2010. 2

[46] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10760–10770, 2020. 3

[47] Yingjie Chen, Diqi Chen, Tao Wang, Yizhou Wang, and Yun Liang. Causal intervention for subject-deconfounded facial action unit recognition. *arXiv preprint arXiv:2204.07935*, 2022.

[48] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9847–9857, 2021.

[49] Yang Liu, Zhaoyang Xia, Mengyang Zhao, Donglai Wei, Yuzheng Wang, Siao Liu, Bobo Ju, Gaoyun Fang, Jing Liu, and Liang Song. Learning causality-inspired representation consistency for video anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 203–212, 2023.

[50] Dingkang Yang, Dongling Xiao, Ke Li, Yuzheng Wang, Zhaoyu Chen, Jinjie Wei, and Lihua Zhang. Towards multimodal human intention understanding debiasing via subject-deconfounding. *arXiv preprint arXiv:2403.05025*, 2024.

[51] Xiang Deng and Zhongfei Zhang. Comprehensive knowledge distillation with causal intervention. *Advances in Neu-*

ral Information Processing Systems (NeurIPS), 34:22158–22170, 2021. 3

[52] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3716–3725, 2020. 3

[53] Teng Sun, Wenjie Wang, Liqiang Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. *arXiv preprint arXiv:2207.11652*, 2022.

[54] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, 2021.

[55] Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. Towards multimodal sentiment analysis debiasing via bias purification. *arXiv preprint arXiv:2403.05023*, 2024. 3

[56] Judea Pearl. *Causality*. Cambridge University Press, 2009. 3

[57] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12700–12710, 2021. 3

[58] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. Context deconfounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19005–19015, June 2023. 3

[59] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19:2, 2000. 3

[60] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015. 4

[61] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[63] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5

[64] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[65] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11962, 2022. 5

[66] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456. pmlr, 2015. 6

[67] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 8