



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

Yuzheng Wang¹, Dingkang Yang¹, Zhaoyu Chen¹, Yang Liu¹, Siao Liu¹,
Wenqiang Zhang², Lihua Zhang¹, Lizhe Qi¹

¹Academy for Engineering & Technology, Fudan University

² Engineering Research Center of AI & Robotics, Ministry of Education, Fudan University

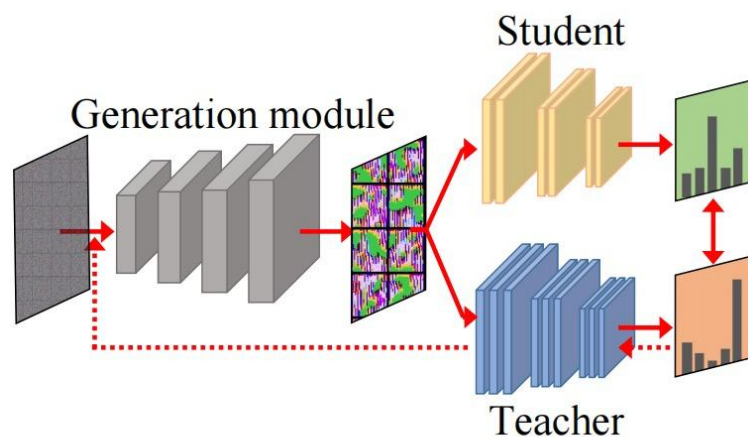
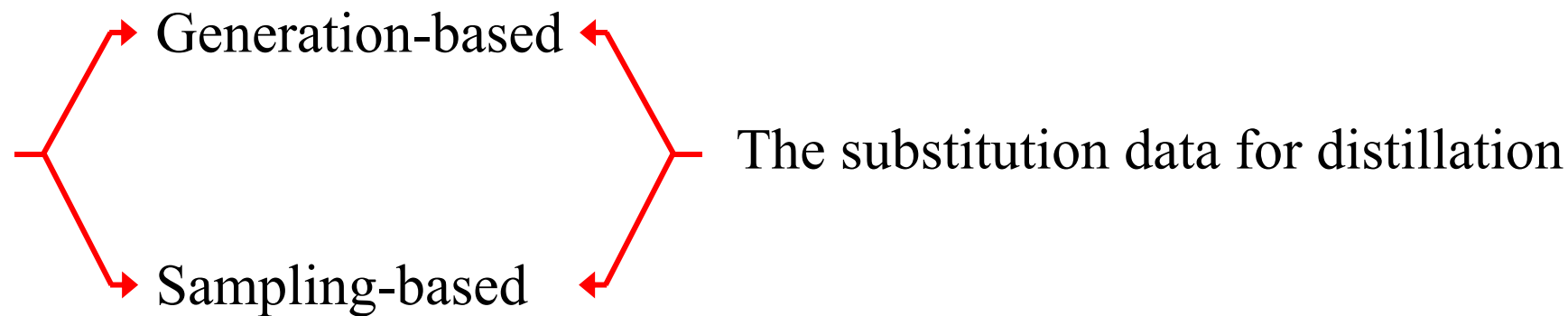
³Green Ecological Smart Technology School-Enterprise Joint Research Center

Outline



- 1. Background
- 2. Motivations & Toy Experiment
- 3. Proposed Method
- 4. Experimental Results
- 5. Conclusion

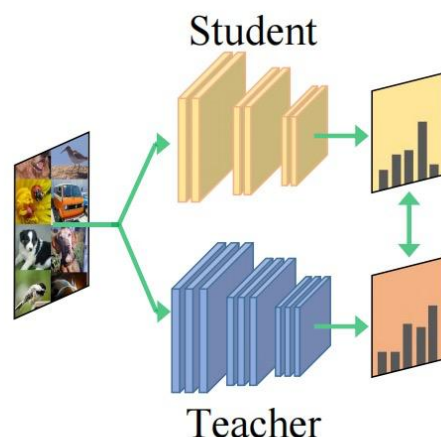
1. Background

Data-Free Knowledge Distillation (DFKD) task




Generation-based

Computational parts:  



Sampling-based

Computational parts: 

Existing Data-Free Knowledge Distillation (DFKD) task is based on the assumption that the original training data is not available due to privacy issues.

Existing methods synthesize data by generating modules or sample unlabeled data from the open-world. Based on this, they can be divided into: 1) generation-based; 2) sampling-based methods.

2. Motivations & Toy Experiment

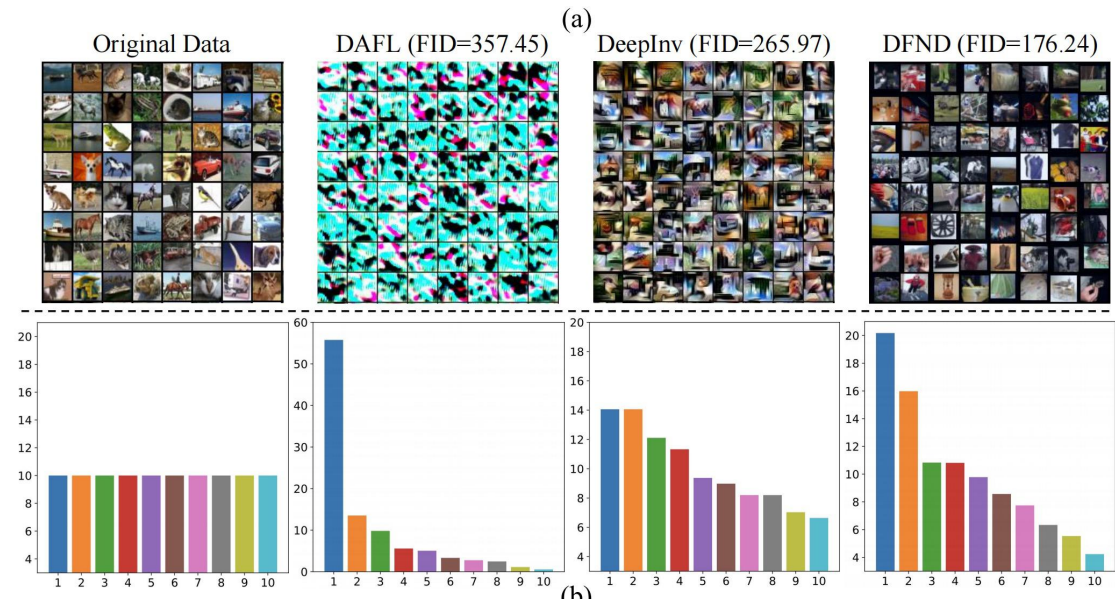
Motivations : For existing DFKD methods, the severe distribution shifts between their substitution and original data causes performance bottlenecks.

Analysis:

- For generation-based methods, the synthetic data 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.

Experiment:

By evaluating the substitution data and original training data of various DFKD methods, we found that there is a serious bias in image quality and category proportion.



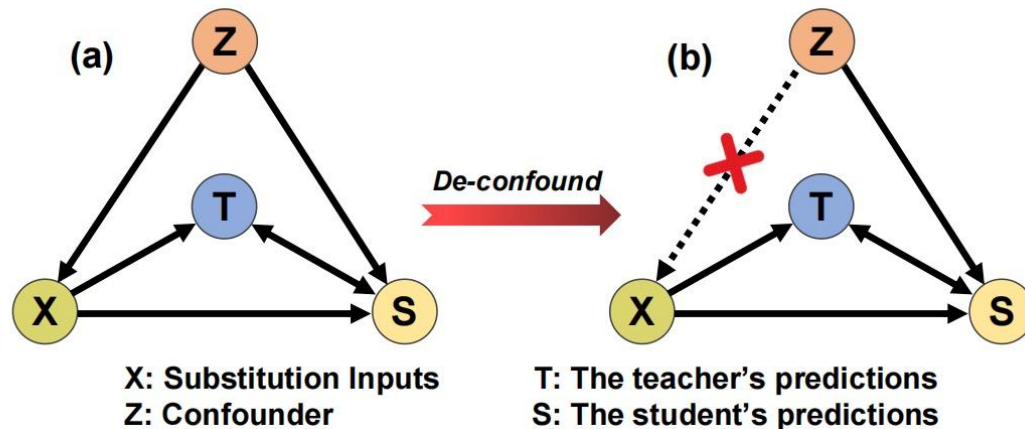
Causal inference is dedicated to dealing with bias issues. Can we use this technology to address the challenges of DFKD?

3. Proposed Method

The causal graph in DFKD task

Firstly, we customize the causal graph according to the properties of the variables in the DFKD task.

During the distillation process, the teacher and student are fed the same substitution data. Our causal graph is applicable to almost all existing DFKD methods so that it can be used as a general framework.



$Z \rightarrow X$:

The confounder Z causes the substitution data X to be biased compared to the original data.

$Z \rightarrow S$:

The detrimental confounder Z confounds and affects the student's training via the causal link.

$X \rightarrow T/S$ & $T \leftrightarrow S$:

The links reflect the interaction causal effect between these two predictions during knowledge distillation. Through these paths, the student can learn consistent knowledge from its teacher.

The backdoor causal path as $X \leftarrow Z \rightarrow S$.

3. Proposed Method

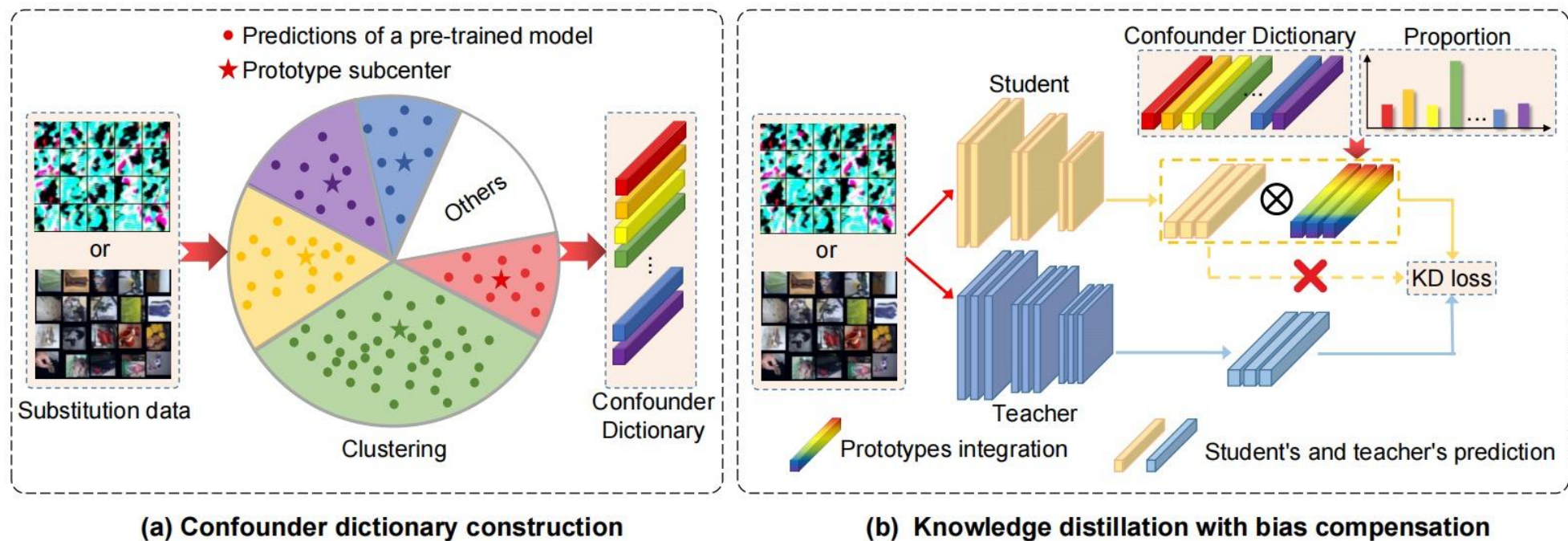
Theory and practice of backdoor adjustments

From theory:

According to the causal graph and the causal theory, cutting off the backdoor causal path $X \leftarrow Z \rightarrow S$ can suppress the interference of the impure knowledge.

In practice,

We design a two-stage framework including:
(i) cofounder dictionary construction
(ii) knowledge distillation with bias compensation.



3. Proposed Method

Backdoor Adjustment & do operator:

The original likelihood estimate:

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

Likelihood after adding the do operator:

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

Confounder Dictionary Construction:

Define dictionary:

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N]$$

Pre-trained model prediction and clustering stratification:

$$M = \{m_j \in \mathbb{R}^d\}_{j=1}^{N_m} \quad \mathbf{z}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} m_k^i$$

Knowledge Distillation with Bias Compensation:

Normalized Weighted Geometric Mean (NWGM):

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

The prior information:

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

The integration of the biased predictions and the prior information:

$$P(\mathcal{S}|do(\mathbf{X})) = \phi(f_S(\mathbf{X}), F(\mathbf{z}))$$

Predictions compensation:

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

4. Experimental Results

Performance Comparison

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**. {†, ‡, †, ‡} denote the provenance mentioned in the analysis.

Dataset	CIFAR-10					CIFAR-100				
T.backbone	resnet-34	vgg-11	wrn-40-2	wrn-40-2	wrn-40-2	resnet-34	vgg-11	wrn-40-2	wrn-40-2	wrn-40-2
S.backbone	resnet-18	resnet-18	wrn-16-1	wrn-40-1	wrn-16-2	resnet-18	resnet-18	wrn-16-1	wrn-40-1	wrn-16-2
Teacher	95.70	92.25	94.87	94.87	94.87	78.05	71.32	75.83	75.83	75.83
Student	95.20	95.20	91.12	93.94	93.95	77.10	77.10	65.31	72.19	73.56
DAFL	92.22	81.10	65.71 [†]	81.33	81.55	74.47	54.16	20.88 [‡]	42.83	43.70
DAFL+KDCI	92.62	81.31	74.56[†]	82.91	82.65	74.51	58.79	31.75[‡]	46.16	48.48
Fast	94.05	90.53	89.29	92.51	92.45	74.34	67.44	54.02	63.91	65.12
Fast+KDCI	94.56	91.16	89.62	93.09	92.85	75.10	68.97	54.69	67.09	68.12
CMI	94.24	91.24	89.16	91.93	92.00	74.64	66.68	55.28	63.44	64.22
CMI+KDCI	94.43	91.28	89.52	92.84	92.73	75.07	69.07	57.19	67.47	67.68
DeepInv	93.26	90.36	83.04	86.85	89.72	61.32 [‡]	54.13 [‡]	53.77	61.33	61.34
DeepInv+KDCI	93.67	91.42	83.47	89.32	91.06	74.59[‡]	69.67[‡]	55.22	62.13	65.90
Mosaick	95.27	91.69	90.03	93.28	92.94	75.91	71.58	59.32	66.61	67.36
Mosaick+KDCI	95.43	92.36	92.25	94.45	94.20	77.06	71.86	62.03	72.19	72.39
DFND	95.36	91.86	90.26	93.33	93.11	74.42	68.97	59.02	69.39	69.85
DFND+KDCI	95.44	92.54	92.47	94.43	94.43	77.09	72.12	66.37	74.20	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	28.79	101.67s	5745M
Fast+KDCI	38.23 (+9.44)	104.43s (+2.71%)	5748M (+0.05%)
DeepInv	20.68	255.26s	3312M
DeepInv+KDCI	34.84 (+14.16)	258.51s (+1.27%)	3316M (+0.12%)
DFND	42.64	129.16s	4196M
DFND+KDCI	49.54 (+6.90)	133.42s (+3.30%)	4198M (+0.05%)

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	53.45	43.02
Fast+KDCI	58.24 (+4.79)	50.12 (+7.10)
DeepInv	51.36	40.25
DeepInv+KDCI	55.27 (+3.91)	46.24 (+5.99)
DFND	42.82	16.03
DFND+KDCI	51.26 (+8.44)	34.32 (+18.29)

4. Experimental Results

Ablation study about prior information

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

Settings	(1) Original $F(\mathbf{z})$			(2) Random λ_i			(3) Random \mathbf{z}_i			(4) w/o $P(\mathbf{z}_i)$		
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

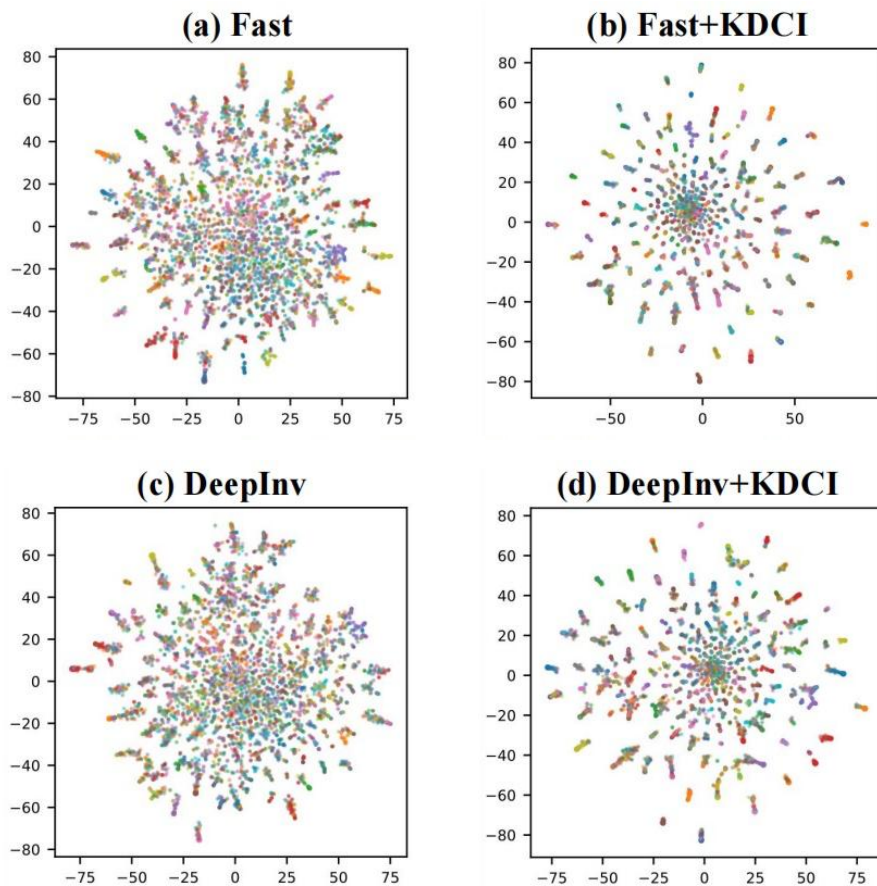
Ablation study about confounder dictionary

Table 5. Ablation studies about the confounder dictionary \mathbf{Z} . “w/o \mathbf{Z} ” denotes the vanilla version of DFKD methods. “*original \mathbf{Z}* ” denotes the original confounder from the teacher itself. “*other \mathbf{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		
\mathbf{Z}	w/o \mathbf{Z}	<i>original \mathbf{Z}</i>	<i>other \mathbf{Z}</i>	w/o \mathbf{Z}	<i>original \mathbf{Z}</i>	<i>other \mathbf{Z}</i>	w/o \mathbf{Z}	<i>original \mathbf{Z}</i>	<i>other \mathbf{Z}</i>	w/o \mathbf{Z}	<i>original \mathbf{Z}</i>	<i>other \mathbf{Z}</i>
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

4. Experimental Results

Visualization results



Case study of causal intervention


	Ground Truth	Vanilla Fast	w/ KDCI	
ImageNet		albatross	missile	albatross
		manhole_cover	petri_dish	manhole_cover
Tiny-ImageNet		coral_reef	lawn_mower	coral_reef
		lakeside	alp	lakeside
		seashore	lampshade	seashore

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

5. Conclusion

- 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 by up to 15.54% on the CIFAR-100 dataset.

Thanks !