Sampling to Distill: Knowledge Transfer from Open-World Data

Yuzheng Wang[1], Zhaoyu Chen[1], Jie Zhang[2], Dingkang Yang[1], Zuhao Ge[1], Yang Liu[1], Siao Liu[1],
Yunquan Sun[1,3], Wenqiang Zhang[1,3*], Lizhe Qi[1,3*]
[1]Academy for Engineering & Technology, Fudan University, Shanghai, China [2]ETH Zurich
[3]Engineering Research Center of AI & Robotics, Ministry of Education, Academy for Engineering & Technology,
Fudan University, Shanghai, China

***Abstract*—Data-Free Knowledge Distillation (DFKD) is a novel task that aims to train high-performance student models using only the pre-trained teacher network without original training data. Most of the existing DFKD methods rely heavily on additional generation modules to synthesize the substitution data resulting in high computational costs and ignoring the massive amounts of easily accessible, low-cost, unlabeled open-world data. Meanwhile, existing methods ignore the domain shift issue between the substitution data and the original data, resulting in knowledge from teachers not always trustworthy and structured knowledge from data becoming a crucial supplement. To tackle the issue, we propose a novel Open-world Data Sampling Distillation (ODSD) method for the DFKD task without the redundant generation process. First, we try to sample open-world data close to the original data's distribution by an adaptive sampling module and introduce a low-noise representation to alleviate the domain shift issue. Then, we build structured relationships of multiple data examples to exploit data knowledge through the student model itself and the teacher's structured representation. Extensive experiments on CIFAR-10, CIFAR-100, NYUv2, and ImageNet show that our ODSD method achieves state-of-the-art performance with lower FLOPs and parameters. Especially, we improve 1.50%-9.59% accuracy on the ImageNet dataset and avoid training the separate generator for each class.**

***Index Terms*—Data-Free Knowledge Distillation, Open-World Unlabeled Data, Contrastive Learning, Relational Distillation**

I. Introduction

Deep learning has made refreshing progress in computer vision and multimedia fields [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. Despite the great success, large-scale models [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] and unavailable privacy data [27], [28], [29], [30], [31] often impede the application of advanced technology on mobile devices. Therefore, model compression and data-free technology have become the key to breaking the bottleneck. To this end, Lopes *et al.* [32] propose Data-Free Knowledge Distillation (DFKD). In this process, knowledge is transferred from the cumbersome model to a small model that is more suitable for deployment without using the original training dataset. As a result, this widely applicable technology has gained much attention.

To replace unavailable private data and effectively train small models, most existing data-free knowledge distillation methods rely on alternately training of the generator and the student, called the generation-based method. Despite not using the original training data, these generation-based methods have many issues. First, their trained generators are abandoned after the students' training [33], [34], [35], [36], [37], [38]. The training of generators brings additional computational costs, especially for large datasets. For instance, a thousand generators are trained for the ImageNet dataset [39], which introduces more computational waste [40], [41]. Then, a serious domain shift issue exists between the generated substitution data and the original training data. Because the substitute data are composed of random noise transformation without supervision information and are highly susceptible to teacher preferences [31]. As a result, the efficiency and effectiveness of the generation-based methods are constrained, affecting student performance [37], [42], [43].

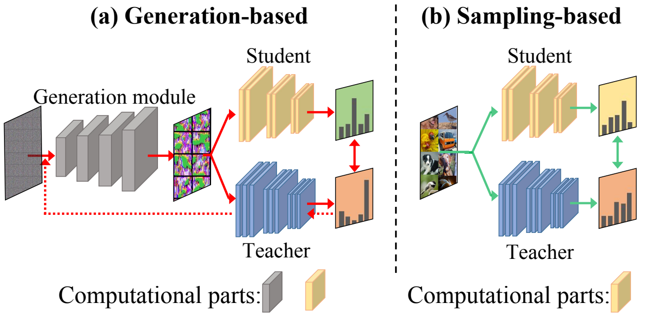

Fig. 1: Comparison of (a) generation-based and (b) sampling-based methods. The sampling-based process utilizes the open-world unlabeled data to distill the student network, so it does not need additional generation costs. At the same time, the extra knowledge in these unlabeled data enriches the knowledge representation from the teacher.

Rather than relying on additional generation modules, Chen *et al.* [44] propose a sampling-based method to train the student network via unlabeled data without the generation calculations. Compared with generation-based methods, sampling-based methods can avoid the training cost of generators, thus improving algorithm efficiency. The comparison of the two methods is shown in Figure 1. Meanwhile, they try to reduce label noise by updating the learnable noise matrix, but the noise matrix's computational costs are expensive. More restrictedly, their sampling method relies on the strict

*Corresponding authors.

confidence ranking and does not consider the data domain similarity issue (We discuss the distribution similarity between sampled data and original data in detail in Section 4.4). In addition, the existing generation-based and sampling-based methods can be summarized as simple imitation learning, *i.e.*, the student mimics the output of a particular data example represented by the teacher [45], [5], [29]. Therefore, these methods do not adequately utilize the implicit relationship among multiple data examples, which leads to the lack of effective knowledge expression in the distillation process.

Based on the above observations, we construct a sampling-based method to sample helpful data from easily accessible, low-cost, unlabeled open-world data, avoiding the unnecessary computational costs of generation modules. In addition, we propose two aspects of customized optimization. (**i**) To cope with the domain shift issue between the open-world and original data, we preferentially try to sample data with a similar distribution to the original data domain to reduce the shifts and design a low-noise knowledge representation learning module to suppress the interference of label noise from the teacher model. (**ii**) To explore the data knowledge adequately, we set up a structured representation of unlabeled data to enable the student to learn the implicit knowledge among multiple data examples. As a result, the student can learn from carefully sampled unlabeled data instead of relying on the teacher. At the same time, to explore an effective distillation process, we introduce a contrastive structured relationship between the teacher and student. The student can make better progress through the structured prediction of the teacher network.

In this paper, we consider a solution to the DFKD task that does not require additional generation costs. On the one hand, we look forward to the solution to data domain shifts from both data source and distillation methods. On the other hand, we try to explore an effectively structured knowledge representation method to deal with the missing supervision information and the training difficulties in the DFKD scenes. Therefore, we propose an Open-world Data Sampling Distillation (ODSD) method, which includes Adaptive Prototype Sampling (APS) and Denoising Contrastive Relational Distillation (DCRD) modules. Specifically, the primary contributions and experiments are summarized as follows:

- We propose a sampling-based method with the unlabeled open-world data. The method does not require additional training of one or more generator models, thus avoiding unnecessary computational costs and model parameters.
- During the sampling process, considering the domain shifts between the unlabeled data and the original data, we propose an Adaptive Prototype Sampling (APS) module to obtain data closer to the original data distribution.
- During the distillation process, we propose a Denoising Contrastive Relational Distillation (DCRD) module to suppress label noise and exploit knowledge from data and the teacher more adequately by building the structured relationships among multiple samples.
- The proposed method achieves state-of-the-art performance with lower FLOPs, improves the effectiveness of the sampling process, and alleviates the distribution shift between the unlabeled data and the original data.

II. Related Work

A. Data-Free Knowledge Distillation

Data-free knowledge distillation aims to train lightweight models when the original data are unavailable. Therefore, the substitute data are indispensable to help transfer knowledge from the cumbersome teacher to the flexible student. According to the source of these data, existing methods are divided into generation-based and sampling-based methods.

Generation-based Methods. The generation-based methods depend on the generation module to synthesize the substitute data. Lopes *et al.* [32] propose the first generation-based DFKD method, which uses the data means to fit the training data. Due to the weak generation ability, it can only be used on a simple dataset such as the MNIST dataset. The following methods combine the Generative Adversarial Networks (GANs) to generate more authentic and reliable data. Chen *et al.* [33] firstly put the idea into practice and define an information entropy loss to increase the diversity of data. However, this method relies on a long training time and a large batch size. Fang *et al.* [34] suggest forcing the generator to synthesize images that do not match between the two networks to enhance the training effect. Hao *et al.* [36] suggest using multiple pre-trained teachers to help the student, which leads to additional computational costs. Do *et al.* [37] propose a momentum adversarial distillation method to help the student recall past knowledge and prevent the student from adapting too quickly to new generator updates. The same domain typically shares some reusable patterns, so Fang *et al.* [41] introduce the sharing of local features of the generated graph, which speeds up the generation process. Since the generation quality is still not guaranteed, some methods spend extra computational costs on gradient inversion to synthesize more realistic data [46], [47]. In addition, Choi *et al.* [48] combine DFKD with other compression technologies and achieve encouraging performance. However, generation-based DFKD methods generate a large number of additional calculation costs in generation modules, while these modules will be discarded after students' training [44].

Sampling-based Methods. To train the student more exclusively, Chen *et al.* [44] propose to sample unlabeled data to replace the unavailable data without the generation module. Firstly, they use a strict confidence ranking to sample unlabeled data. Then, they propose a simple distillation method with a learnable adaptive matrix. Despite no additional generating costs and promoting encouraging results, their method ignores the intra-class relationships of multiple unlabeled data. Simultaneously, the simple strict confidence causes more data to be sampled for simple classes, leading to imbalanced data classes. In addition, their proposed distillation method is relatively simple and lacks structured relationship expression, which limits the student's performance.

B. Contrastive Learning

Contrastive learning makes the model's training efficient by learning the data differences [49]. The unsupervised training pipeline usually requires storing negative data by a memory bank [50], large dictionaries [51], or a large batch size [52].

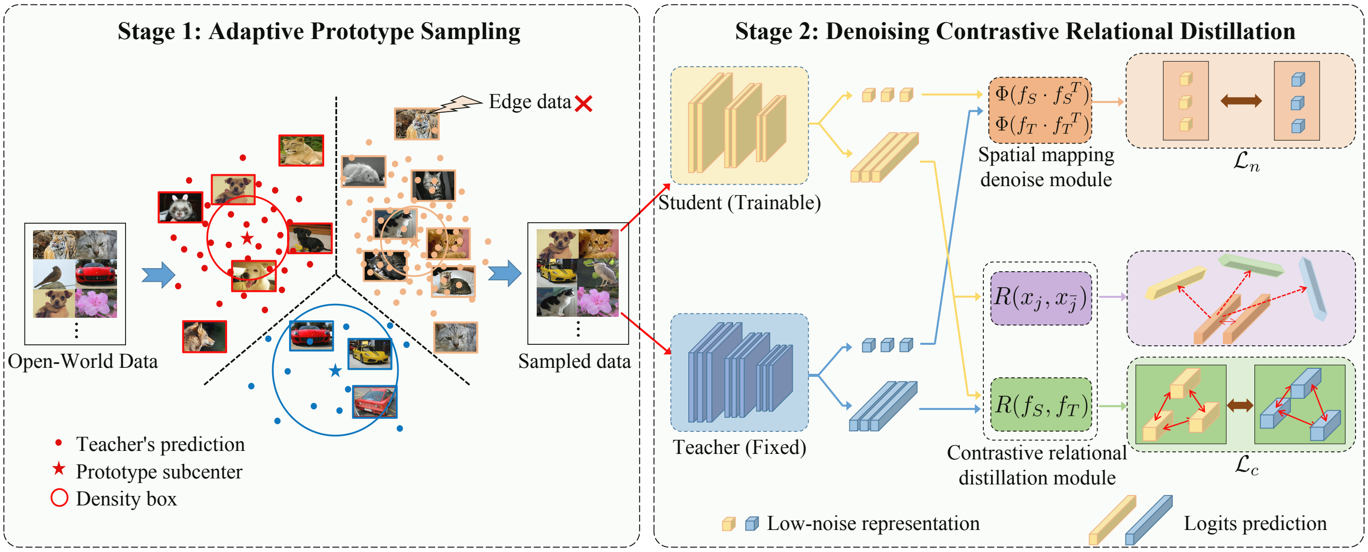

Fig. 2: The pipeline of our proposed ODSD. First, all open-world unlabeled data passes through adaptive prototype sampling so that the substitute dataset resembles the distribution of the original data. Then, based on these data, the student can make progress through low-noise information representation, data knowledge mining, and structured knowledge from the teacher.

Even it requires a lot of computation costs, *e.g.*, additional normalization [53], and network update operations [54]. The high storage and computing costs seriously reduce knowledge distillation efficiency. But at the same time, this idea of mining knowledge in unlabeled data may be helpful for the student's learning. Due to such technical conflicts, there are few methods to combine knowledge distillation and contrastive learning in the past perfectly. As a rare attempt, Tian *et al*. [55] propose a contrastive data-based distillation method by updating a large memory bank. However, data quality cannot be guaranteed for data-free knowledge distillation, and data domain shifts are intractable, making the above process challenging.

In this work, we attempt to explore additional knowledge from both the data and the teacher. Therefore, we further stimulate students' learning ability by using the internal relationship of unlabeled data and constructing a structured contrastive relationship. To our best knowledge, this is the first combination of data-free knowledge distillation and contrastive learning at a low cost during the distillation process, which achieves an unexpected effect.

III. Methodology

A. Overview

Our pipeline includes two stages: (**i**) unlabeled data sampling and (**ii**) distillation training, as shown in Figure 2. For the sampling stage, we sample unlabeled data by an adaptive sampling mechanism to obtain data closer to the original distribution. For the distillation stage, the student learns the knowledge representation after denoising through a spatial mapping denoise module. Further, we mine more profound knowledge of the unlabeled data and build the structured relational distillation to help the student gain better performance. The complete algorithm is shown in Algorithm 1.

B. Adaptive Prototype Sampling

The unlabeled data and the original data are distributed differently in many cases. To obtain the substitution data with a more similar distribution to the original data from the specific unlabeled dataset, we propose an Adaptive Prototype Sampling (APS) module, which considers the teacher's familiarity with the data, excludes misclassified offset noisy data, and focuses on the class balance of the sampled data. Based on these, we design three score indicators to evaluate the effectiveness of the unlabeled data for student training corresponding to the above three aspects, including the data confidence score, the data outlier score, and the class density score.

(a) Data Confidence Score. The teacher provides the prediction logits $P = [p_1, \dots, p_n] \in \mathbb{R}^{n \times C}$ on the unlabeled dataset $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_n\}$, where p_i denotes the prediction for the i-th sample satisfying $p_i \in \mathbb{R}^{1 \times C}$. n denotes the number of data, and C denotes the number of classes. Then the prediction is converted into the probability of the unified scale as $\tilde{p}_i = \sigma(p_i)$, where σ denotes the softmax layer and $\tilde{p}_i$ denotes the confidence probability corresponding to the predicted result class. Therefore, $\tilde{p} = [\tilde{p}_1, \dots, \tilde{p}_n]$ represents the confidence of each data in the unlabeled dataset. We choose the largest $\max\{\tilde{p}\}$ for normalization. The confidence score of i-th sample $\boldsymbol{x}_i$ can be calculated as: $sc_i = \frac{\tilde{p}_i}{|\max\{\tilde{p}\}|}$.

(b) Data Outlier Score. The data distribution of the substitution data and the original data is different. Therefore, the confusing edge data should be excluded, *i.e.*, the data with different distributions but also predicted as the same target class. For example, a tiger is predicted as the class of cat, as shown in the orange part of Stage 1 in Figure 2. We first separate the teacher predictions according to the predicted classes as $\rho_{i,c} = p_i \in c$. For each class, $\rho_{i,c}$ is clustered [56] to explore the intra-class relationships through k layering as $\mu_{c,k}$. Then the prediction features for the whole unlabeled dataset can be expressed as a group of CK prototypes as

Algorithm 1 The proposed ODSD algorithm.

```
Input: A frozen teacher network $f_T$, an unlabeled open-world dataset $X_U$, and the target number of sampled data $M$.
Module 1: Adaptive pototype sampling
for unlabeled data $\boldsymbol{x}_i$ in $X_U$ do
    Classify teacher predictions $p_i$ as $\rho_{i,c} = p_i \in c$;
    Calculate confidence probability: $\tilde{p}_i = \sigma(p_i)$
    Cluster the prediction vector as the prototypes $\mu_{c,k}$.
end for
for Prototypes $\mu_{c,k}$ in class $c$ do
    Obtain prototype similarity: $\tilde{o}_i = \cos(\rho_{i,c}, \mu_{c,k})_{k=1}^{K}$;
    Calculate intra-class outliers mean: $u_c = \frac{1}{n_c} \sum {}_{p_i \in c} \tilde{o}_i$;
    Calculate the density score $D_c = \frac{\sqrt{u_c}}{\log_e (n_c + C)}$.
end for
Calculate sampling score: $S = \frac{\tilde{p}_i}{|\max\{\tilde{p}\}|} - \frac{\tilde{o}_i}{|\max\{\tilde{o}\}|} + \frac{D_c}{|\max\{D\}|}$
Sample top-$M$ data with the highest score as $X_A$.
Module 2: Denoising contrastive relational distillation.
for $i$ in number of epochs do
    for training data $\boldsymbol{x}$ in $X_A$ do
        Calculate $\mathcal{L}_{total}$ as Eq.8 and update the student $f_S$.
    end for
end for
Output: The trained student $f_S$ and a reusable sampling list $L$ of the teacher $f_T$ on dataset $X_U$.
```

$\{\mu_{c,k} \in \mathbb{R}^{1 \times C}\}_{c,k=1}^{C,K}$, where c denotes the c-th class, and K denotes the hyperparameter representing the number of prototypes for each class. The prototype centers of the c-th class can be expressed as $\{\mu_{c,k}\}_{k=1}^{K}$. The outlier of each unlabeled data x_i can be calculated with the sum of the prototype centers of its class as $\tilde{o}_i = \sum_{k=1}^{K} \cos(\rho_{i,c}, \mu_{c,k})$, where $\cos$ denotes the cosine similarity metric. Similar to the confidence score, we select the maximum value $\max\{\tilde{o}\}$ for normalization. As a result, the outlier score can be calculated as: $so_i = \frac{\tilde{o}_i}{|\max\{\tilde{o}\}|}$.

(c) Class Density Score. To help the student learn various classes effectively, we calculate the class density for the class balance of the sampled data. As shown in Stage 1 of Figure 2, we increase the sampling range for classes with sparse data (the blue part) while we reduce the sampling range for classes with redundant data (the orange part). Based on this, we first separate the above intra-class outliers $\tilde{o}_i$ of all data by their predicted classes. The outliers mean value of each class can be calculated as $u_c = \frac{1}{n_c} \sum {}_{p_i \in c} \tilde{o}_i$, where n_c is the number of the data predicted as c-th class. Therefore, the Dcluster parameter D_c can be calculated as: $D_c = \frac{\sqrt{u_c}}{\log_e (n_c + C)}$, which reflects the data density predicted to be c-th class. The introduction of a constant C (the number of classes) helps the numerical stability when the available unlabeled data is small while having little effect on the results when the amount of data is sufficient (under normal conditions). After selecting the maximum value $\max\{D\}$ for normalization, the density score of each data can be calculated according to the predicted class as $sd_i = \frac{D_c}{|\max\{D\}|}$, when $\arg\max(p_i) = c$.

Finally, we calculate the total score as $S_{total} = sc_i - so_i + sd_i$. Based on this, the data closer to the distribution of the original data domain are sampled, which can help the student learn better. The quantitative analysis is shown in Table VII.

C. Denoising Contrastive Relational Distillation

After obtaining the high score data, the distillation process can be carried out. We denote f_T and f_S as the teacher and student networks and denote $\boldsymbol{x}$ as the data in sampled set X_A. According to the definition [57], the knowledge distillation loss is calculated as:

$$\mathcal{L}_{KD} = \sum_{\boldsymbol{x} \in X_A} D_{KL}(f_T(\boldsymbol{x})/\tau_{kd}, f_S(\boldsymbol{x})/\tau_{kd}), \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence, and τ_{kd} is the distillation temperature. $\mathcal{L}_{KD}$ allows the student to imitate the teacher's output. However, the main challenge is the distribution differences between the substitute and original data domains, leading to label noise interference. Simultaneously, the ground-truth labels are unavailable, so correct information supervision is missing. Therefore, we propose a Denoising Contrastive Relational Distillation (DCRD) module, which includes a spatial mapping denoise component and a contrastive relationship representation component to help the student get better performance and mitigate label noise.

Spatial Mapping Denoise. The distribution in the unlabeled data differs from the unavailable original data, which indicates the inevitable label noise. Inspired by manifold learning [58], low dimensional information representation contains purer knowledge with less noise interference [59]. Here, we utilize a low-dimensional spatial mapping denoise component to help the student learn low-noise knowledge representation. Based on this, we perform eigendecomposition Φ on the teacher's prediction and its transposed product matrix [60]. According to the distance invariance, the autocorrelation matrix d_{ij}^2 in a mini-batch as:

$$\sum_{i}^{N} \sum_{j}^{N} d_{ij}^2 = 2N \cdot tr(Z_t Z_t^T), \tag{2}$$

where N denotes the batch size, and $tr(\cdot)$ denotes the trace of a matrix. Z_t is the low-dimensional spatial vector representation from the teacher calculated as $\Phi(f_T(\boldsymbol{x}) \cdot {f_T}^T(\boldsymbol{x})) = Z_t = V_t \Lambda_t^{1/2}$, where V_t is the eigenvalue, and Λ_t is the eigenvector. Similarly, we can get the student predictions of low-dimensional representation as Z_s. Then, we set up a distillation loss to correct the impact of label noise by the spatial mapping of the two networks. The spatial mapping denoise distillation loss is calculated as:

$$\mathcal{L}_n = \ell_h(\Phi(f_T(\boldsymbol{x}) \cdot {f_T}^T(\boldsymbol{x})), \Phi(f_S(\boldsymbol{x}) \cdot {f_S}^T(\boldsymbol{x}))) = \ell_h(Z_t, Z_s), \tag{3}$$

where $\ell_h(\cdot, \cdot)$ denotes the Huber loss.

Contrastive Relational Distillation. The missing supervision information limits the student's performance. It is indispensable to adequately mine the knowledge in unlabeled data to compensate for the lack of information. To avoid a single imitation of a particular data example, we build two kinds of structured relationships to mine knowledge from the data and the teacher.

Firstly, the student can adequately explore the structured relation among multiple unlabeled data by learning the instance invariant. $\boldsymbol{x}_i, \boldsymbol{x}_j$ are the different data in a mini-batch. We calculate the prediction difference between data as follows:

$$\ell_s^{\boldsymbol{x}_i \boldsymbol{x}_j} = \frac{\cos(f_S(\boldsymbol{x}_i), f_S(\boldsymbol{x}_j))/\tau}{\sum_{k=1,k\neq i}^{2N} \cos(f_S(\boldsymbol{x}_i), f_S(\boldsymbol{x}_k))/\tau}, \tag{4}$$

where τ_1 denotes contrastive temperature. Next, we can calculate the consistency instance discrimination loss as:

$$\mathcal{L}_{c1} = -\frac{1}{N} \sum_{j=1}^{N} \log \ell_s^{\boldsymbol{x}_j \bar{\boldsymbol{x}}_j}, \tag{5}$$

where $\bar{\boldsymbol{x}}_j$ denotes the strong data augmentation of $\boldsymbol{x}_j$. This unsupervised method is especially effective when the teacher makes wrong results.

Secondly, we construct a structured contrastive relationship between the teacher and student, which promotes consistent learning between the teacher and student. The structured knowledge transfer process is calculated as:

$$\ell_{ts}^{\boldsymbol{x}'_i} = \frac{\cos(f_T(\boldsymbol{x}'_i), f_S(\boldsymbol{x}'_i))/\tau}{\sum_{k=1,k\neq i}^{4N} \cos(f_T(\boldsymbol{x}'_i), f_S(\boldsymbol{x}'_k))/\tau}, \tag{6}$$

where $\boldsymbol{x}' = \boldsymbol{x} \cup \bar{\boldsymbol{x}}$ denotes the set of the sampled data before and after strong data augmentation. And $\boldsymbol{x}'$ contains $2N$ samples for each batch. We calculate the teacher-student consistency loss as:

$$\mathcal{L}_{c2} = -\frac{1}{2N} \sum_{j=1}^{2N} \log \ell_{ts}^{\boldsymbol{x}'_j}. \tag{7}$$

The student can obtain better learning performance through the mixed structured and consistent relationship learning between the two networks. Then, the contrastive relational distillation loss is $\mathcal{L}_c = \mathcal{L}_{c1} + \mathcal{L}_{c2}$. Finally, we can get the total denoising contrastive relational distillation loss as:

$$\mathcal{L}_{total} = \mathcal{L}_{KD} + \lambda_1 \cdot \mathcal{L}_n + \lambda_2 \cdot \mathcal{L}_c, \tag{8}$$

where λ_1, λ_2 are the trade-off parameters for training losses.

IV. Experiments

A. Experimental Settings

Datasets. We evaluate the proposed ODSD method for the classification and semantic segmentation tasks. For classification, we evaluate it on widely used datasets: 32×32 CIFAR-10, CIFAR-100 [61], and 224 × 224 ImageNet [39]. For semantic segmentation, we evaluate the proposed method on 128 × 128 NYUv2 dataset [62]. Besides, the corresponding open-world datasets are shown in Table I, which is the same as DFND [44] for a fair comparison.

Implementation Details. The proposed model is implemented in PyTorch [63] and trained with RTX 3090 GPUs. For the CIFAR-10 and CIFAR-100 datasets, we conduct five sets of backbone combinations, set two groups of different numbers of sampled samples (150k or 600k), and train the students for 200 epochs. For the ImageNet dataset, we conduct three sets of backbone combinations and train the students for 200 epochs. The number of sampled samples is 600k. For the NYUv2 dataset, the DeeplabV3 [64] is used as the model architecture followed previous work. The teacher uses ResNet-50 [65] as the backbone, and the student uses mobilenetv2 [66]. We sample 200k unlabeled samples and train the student for 20 epochs. For the above datasets, we set τ_{kd} as 4 to be the same as other distillation methods and set τ as 0.5 to be the same as [52]. Besides, we set λ_1 as 10 and λ_2 as 0.5, use the SGD optimizer with momentum as 0.9, weight decay as 5×10^{-4}, the batch size N as 64, and cosine annealing learning rate with an initial value of 0.025.

TABLE I: Illustration of original private data and their corresponding substitute open-world datasets.

Original data	CIFAR	ImageNet	NYUv2
Unlabeled data	ImageNet	Flickr1M	ImageNet

Baselines. We compare two kinds of DFKD methods. One is to spend extra computing costs to obtain generation data by generation module, including DeepInv [46], CMI [47], DAFL [33], ZSKT [35], DFED [36], DFQ [48], Fast [41], MAD [37], DFD [40], KAKR [43], SpaceshipNet [67], and DFAD [34]. Another is to use unlabeled data from easily accessible open source datasets based on sampling, *i.e.*, DFND [44].

B. Performance Comparison

To evaluate the effectiveness of our ODSD, we comprehensively compare it with current SOTA DFKD methods regarding the student's performance, the effectiveness of the sampling method, and training costs.

Experiments on CIFAR-10 and CIFAR-100. We first verify the proposed method on the CIFAR-10 and CIFAR-100 [61]. The baseline "*Teacher*" and "*Student*" means to use the corresponding backbones of the teacher or student for direct training with the original training data, and "*KD*" represents distilling the student network with the original training data. Generation-based methods include training additional generators and calculating model gradient inversion. Sampling-based methods use the unlabeled ImageNet dataset. We reproduce the DFND using the unified teacher models, and the result is slightly higher than the original paper.

As shown in Table II, our ODSD has achieved the best results on each baseline. Under most baseline settings, ODSD brings gains of 1% or even higher than the SOTA methods, even though students' accuracy is very close to their teachers. In particular, the students of our ODSD outperform the teachers on some baselines. As far as we know, it is the first DFKD method to achieve such performance. The main reasons for its breakthrough in analyzing the algorithm's performance come from three aspects. First, our data sampling method comprehensively analyzes the intra-class relationships in the unlabeled data, excluding the difficult edge data and significant distribution differences data. At the same time, the number of data in each class is relatively more balanced, which is conducive to all kinds of balanced learning compared with other sampling methods. Second, our knowledge distillation method considers the representation of low-dimensional and low-noise information and expands the representation of knowledge

TABLE II: Student accuracy (%) on CIFAR datasets. **Bold** and underline numbers denote the best and the second best results.

Dataset	Method	Type	ResNet-34	VGG-11	WRN40-2	WRN40-2	WRN40-2
			ResNet-18	ResNet-18	WRN16-1	WRN40-1	WRN16-2
CIFAR-10	Teacher	-	95.70	92.25	94.87	94.87	94.87
	Student		95.20	95.20	91.12	93.94	93.95
	KD		95.58	94.96	92.23	94.45	94.52
	DeepInv [46]	Generation	93.26	90.36	83.04	86.85	89.72
	CMI [47]		94.84	91.13	90.01	92.78	92.52
	DAFL [33]		92.22	81.10	65.71	81.33	81.55
	ZSKT [35]		93.32	89.46	83.74	86.07	89.66
	DFED [36]		-	-	87.37	92.68	92.41
	DFQ [48]		94.61	90.84	86.14	91.69	92.01
	Fast [41]		94.05	90.53	89.29	92.51	92.45
	MAD [37]		94.90	-	-	-	92.64
	KAKR_MB [43]		93.73	-	-	-	-
	KAKR_GR [43]		94.02	-	-	-	-
	SpaceshipNet [67]		95.39	92.27	90.38	93.56	93.25
	DFND_150k [44]	Sampling	94.18	91.77	87.95	92.56	92.02
	DFND_600k [44]		95.36	91.86	90.26	93.33	93.11
	ODSD_150k		95.05	92.02	89.14	92.94	92.34
	ODSD_600k		**95.70**	**92.55**	**91.53**	**94.31**	**94.02**
CIFAR-100	Teacher	-	78.05	71.32	75.83	75.83	75.83
	Student		77.10	77.10	65.31	72.19	73.56
	KD		77.87	75.07	64.06	68.58	70.79
	DeepInv [46]	Generation	61.32	54.13	53.77	61.33	61.34
	CMI [47]		77.04	70.56	57.91	68.88	68.75
	DAFL [33]		74.47	54.16	20.88	42.83	43.70
	ZSKT [35]		67.74	54.31	36.66	53.60	54.59
	DFED [36]		-	-	41.06	60.96	60.79
	DFQ [48]		77.01	66.21	51.27	54.43	64.79
	Fast [41]		74.34	67.44	54.02	63.91	65.12
	MAD [37]		77.31	-	-	-	64.05
	KAKR_MB [43]		77.11	-	-	-	-
	KAKR_GR [43]		77.21	-	-	-	-
	SpaceshipNet [67]		77.41	71.41	58.06	68.78	69.95
	DFND_150k [44]	Sampling	74.20	69.31	58.55	68.54	69.26
	DFND_600k [44]		74.42	68.97	59.02	69.39	69.85
	ODSD_150k		77.90	72.24	60.55	71.66	72.42
	ODSD_600k		**78.45**	**72.71**	**60.57**	**72.71**	**73.20**

through data augmentation. The structured relationship distillation method helps the student effectively learn knowledge from both multiple data and its teacher. Finally, the knowledge of our ODSD does not entirely come from the teacher but also the consistency and differentiated representation learning of unlabeled data, which is helpful when the teacher makes mistakes. The previous methods ignore the in-depth mining of data knowledge, decreasing students' performance.

TABLE III: Student accuracy (%) on ImageNet dataset.

Method	Type	ResNet-50	ResNet-50	ResNet-50
		ResNet-18	ResNet-50	MobileNetv2
Teacher	-	75.59	75.59	75.59
Student		68.93	75.59	63.97
KD		68.10	74.76	61.67
DFD [40]	Generation	54.66	69.75	43.15
DeepInv$_{2k}$ [46]		-	68.00	-
Fast$_{50}$ [41]		53.45	68.61	43.02
DFND [44]	Sampling	42.82	59.03	16.03
ODSD		**58.24**	**71.25**	**52.74**

Experiments on ImageNet. We conduct experiments on a large-scale ImageNet dataset to further verify the effectiveness. Due to the larger image size, it is challenging to effectively synthesize training data for most generation-based methods. Most of them failed. A small number of methods train 1,000 generators (one generator for one class), resulting in a large amount of additional computational costs. In this case, our sampling method reduces the computational costs more significantly. We set up three baselines to compare the performance of our method with the SOTA methods. Table III reports the experimental results. Our ODSD still achieves several percentage points increase compared with other SOTA methods, especially in the cross-backbones situation (9.59%).

TABLE IV: Total FLOPs and params in DFKD methods.

Method	DeepInv	CMI	DAFL	ZSKT	DFQ	DFND	**ODSD**
FLOPs	4.36G	4.56G	0.67G	0.67G	0.79G	0.56G	0.56G
params	11.7M	12.8M	12.8M	12.8M	17.5M	11.7M	11.7M

TABLE V: APS compared with the SOTA sampling method.

Sampling methods		Method	
	KD	DFND	**ODSD**
Random	76.85	73.15	76.43
DFND	76.67	73.68	77.40
APS	**77.27**	**73.89**	**77.90**

TABLE VI: Segmentation results on NYUv2 dataset.

Algorithm	Teacher	Student	DAFL	DFAD	Fast	DFND	**ODSD**
mIoU	0.517	0.375	0.105	0.364	0.366	<u>0.378</u>	**0.397**

TABLE VII: Diagnostic studies of the proposed method.

Training objective $\mathcal{L}$				Data sampling scores S			
ID	Setting	Accuracy (%)		ID	Setting	Accuracy (%)	
		50k	150k			50k	150k
(1)	ours	**75.26**	**77.90**	(5)	ours	**75.26**	**77.90**
(2)	w/o $\mathcal{L}_n$	74.82	77.71	(6)	w/o sc_i	73.96	77.04
(3)	w/o $\mathcal{L}_c$	74.71	77.58	(7)	w/o so_i	68.07	76.67
(4)	w/o $\mathcal{L}_n, \mathcal{L}_c$	74.39	77.27	(8)	w/o sd_i	70.24	76.59

Due to the lack of structured knowledge representation, the DNFD algorithm performs poorly on the large-scale dataset. Comparing DFND and ODSD, our structured framework improves the overall understanding ability of the student.

Comparison of Training Costs. To verify that the generation-based methods add extra costs that we mentioned in the introduction section, we further calculate the total floating point operations (FLOPs) and parameters (params) required by various DFKD algorithms, as shown in Table IV. Our method only needs training costs and params of the student network without additional generation modules. Our sampling process introduces 256.78 seconds for sample selection ($K = 5$) on the CIFAR100 with a single RTX 3090 GPU (The teacher uses the *ResNet-34*) while the fastest generation-based method ZSKT also takes 1.54 hours to synthesize data. These generation modules will be discarded after student training, which causes a waste of computing power.

Comparison of Data Sampling Efficiency. To verify the sampling mechanism's effectiveness, we compare our APS method with the current SOTA unlabeled data sampling method DFND [44]. Three data sampling methods (random sampling, DFND sampling, and our proposed APS) are set on three different distillation algorithms, including: KD [57], DFND [44], and our ODSD method. Table V reports the results. For KD, we use the sampled data instead of the original generated data with $\mathcal{L}_{KD}$ distillation loss. From the result, this setting is competitive, even better than the distillation loss of DFND. For DFND, we reproduce it with open-source codes and keep the original training strategy unchanged. We find the performance of the DFND sampling method is unstable, which causes it to be lower than random sometimes. For ODSD, we use the distillation loss in Equation (8). Our proposed sampling method achieves the best performance in all three benchmarks and significantly improves performance. By comprehensively considering the data confidence, the data outliers, and the class density, our ODSD can more fully mine intra-class relationships of the unlabeled data. As a result, the sampled data are more helpful for subsequent student learning.

Experiments about Semantic Segmentation. In addition to image classification tasks, our algorithm can also effectively solve the problem of DFKD in image semantic segmentation on the NYUv2 dataset. Mean Intersection over Union (mIoU) is set as the segmentation evaluation metric. No generation module is defined for our method, and other settings are the same as DFAD [34]. Table VI shows segmentation results on the NYUv2 dataset. Our ODSD also achieves the best performance. Besides, we visualize the segmentation results of different networks to get more convincing results as shown in Figure 3. "*Input*" and "*Ground Truth*" represent the input test data and their corresponding real labels. Most data-free distillation algorithms hide the code of the segmentation part, so it is not easy to make a visual comparison. Here, we choose DFAD as the baseline algorithm of visualization. Our proposed ODSD algorithm achieves better segmentation results than DFAD, especially for object contour segmentation. The slight noise around the contour is effectively suppressed. Further, through in-depth mining the knowledge from the data and teacher, our student have gained better understanding ability.

C. Diagnostic Experiment

We conduct the diagnostic studies on the CIFAR-100 dataset. We use ResNet-34 as the teacher's backbone and ResNet-18 as the student's backbone. 50k and 150k data are sampled. Other settings are the same as the Table II.

Distillation Training Objective. We first investigate our overall training objective (cf. Equation (8)). Two different data sampling numbers are set in this experiment. As shown in the experiments (1-4) of Table VII, the model with $\mathcal{L}_{KD}$ alone achieves accuracy scores of 74.39% and 77.27% on 50k and 150k data sampling settings. Adding $\mathcal{L}_n$ or $\mathcal{L}_c$ individually brings gains (*i.e.*, **0.32%, 0.31%/ 0.43%, 0.44%**), indicating the effectiveness of our proposed distillation method. Our method performs better with **75.26%** and **77.90%**. With the above results, the proposed training objectives are effective and can help the student gain better performance.

Data Sampling Scores. To verify the effectiveness of the three sampling scores in section 3.2, we further conduct ablation experiments. Using all scores, the model can achieve the best performance with **75.26%** and **77.90%** accuracy shown in experiments (5-8) of Table VII. When the confidence score sc_i is abandoned, the familiarity of the teacher network with the sampled data decreases, reducing the amount of adequate information contained in the data. Without the outlier score so_i, the lack of modeling of the intra-class relationship of the data to be sampled leads to increased data distribution

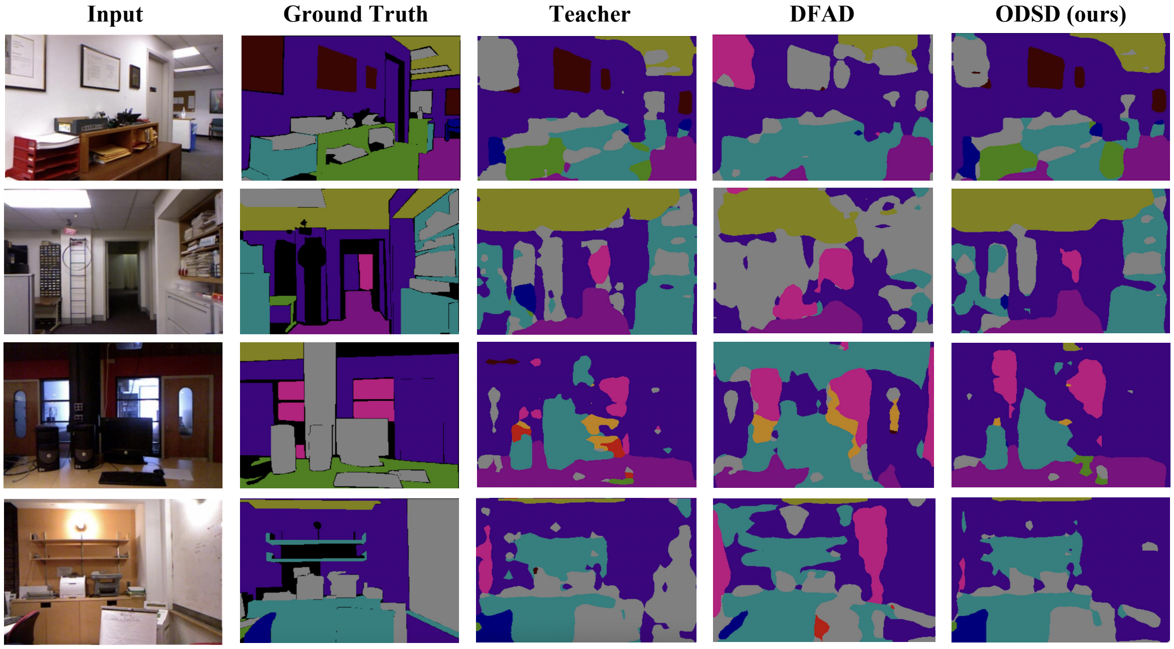

Fig. 3: Visualization segmentation results on the NYUv2 dataset.

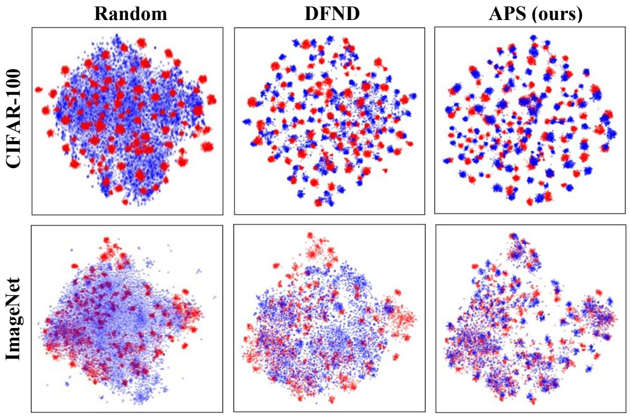

Fig. 4: t-SNE visualization of the data distributions on CIFAR-100 and ImageNet datasets. Red dots denote original domain data, while blue dots denote unlabeled sampling data. The distance between dot groups reflects the similarity between data domains. The data sampled by our APS method is more similar to that of the original domain, effectively reducing domain noise and improving learning performance.

difference between the substitute data domain and the original data domain. Further, the class density score sd_i can measure the number of data in each class and maintain the balance of the sampled data. In summary, all three score indicators can help students perform better.

D. Visualization

To verify the distribution similarity between the sampled data and the original data of our APS sampling method and the DFND sampling method, we use t-SNE [68] to visualize the data feature distribution. Teacher uses ResNet-34 as the backbone on the CIFAR-100 and ResNet-50 as the backbone on the ImageNet. For both datasets, we reserve 100 classes from validation data. In addition, we also visualize the distribution of data obtained by random sampling as a baseline reference. Figure 4 shows the data distribution results of different sampling methods. Our clustering results are closer to the extracted features of the original data. For the more complex ImageNet, this advantage is further amplified. Reducing the distribution difference between sampled and original data helps reduce data label noise, which is the key for the student to perform well.

V. Conclusion

Most existing data-free knowledge distillation methods rely heavily on additional generation modules, bringing additional computational costs. Meanwhile, these methods disregard the domain shifts issue between the substitute and original data and only consider the teacher's knowledge, ignoring the data knowledge. This paper proposes an Open-world Data Sampling Distillation method. We sample unlabeled data with a similar distribution to the original data and introduce low-noise knowledge representation learning to cope with domain shifts. To explore the data knowledge adequately, we design a structured knowledge representation. Comprehensive experiments illustrate the effectiveness of the proposed method, which achieves significant improvement and state-of-the-art performance on various benchmarks.

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] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[4] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[5] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[6] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3713–3722.

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[8] D. Yang, K. Yang, H. Kuang, Z. Chen, Y. Wang, and L. Zhang, "Towards context-aware emotion recognition debiasing from a causal demystification perspective via de-confounded training," *arXiv preprint arXiv:2407.04963*, 2024.

[9] Y. Liu, J. Liu, K. Yang, B. Ju, S. Liu, Y. Wang, D. Yang, P. Sun, and L. Song, "Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system," *IEEE Transactions on Industrial Informatics*, 2023.

[10] Y. Wang, Z. Chen, D. Yang, Y. Sun, and L. Qi, "Self-cooperation knowledge distillation for novel class discovery," *arXiv preprint arXiv:2407.01930*, 2024.

[11] D. Yang, Z. Chen, Y. Wang, S. Wang, M. Li, S. Liu, X. Zhao, S. Huang, Z. Dong, P. Zhai, and L. Zhang, "Context de-confounded emotion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19 005–19 015.

[12] D. Yang, S. Huang, Z. Xu, Z. Li, S. Wang, M. Li, Y. Wang, Y. Liu, K. Yang, Z. 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*, 2023, pp. 20 459–20 470.

[13] Y. Liu, D. Yang, G. Fang, Y. Wang, D. Wei, M. Zhao, K. Cheng, J. Liu, and L. Song, "Stochastic video normality network for abnormal event detection in surveillance videos," *Knowledge-Based Systems*, vol. 280, p. 110986, 2023.

[14] D. Yang, D. Xiao, K. Li, Y. Wang, Z. Chen, J. Wei, and L. Zhang, "Towards multimodal human intention understanding debiasing via subject-deconfounding," *arXiv preprint arXiv:2403.05025*, 2024.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[18] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[19] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Z. Chen, B. Li, J. Xu, S. Wu, S. Ding, and W. Zhang, "Towards practical certifiable patch defense with vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 148–15 158.

[21] D. Yang, Y. Liu, C. Huang, M. Li, X. Zhao, Y. Wang, K. Yang, Y. Wang, P. Zhai, and L. Zhang, "Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences," *Knowledge-Based Systems*, p. 110370, 2023.

[22] Y. Wang, Z. Chen, D. Yang, Y. Liu, S. Liu, W. Zhang, and L. Qi, "Adversarial contrastive distillation with adaptive denoising," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] D. Yang, K. Yang, Y. Wang, J. Liu, Z. Xu, R. Yin, P. Zhai, and L. Zhang, "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.

[24] S. Liu, Z. Chen, Y. Liu, Y. Wang, D. Yang, Z. Zhao, Z. Zhou, X. Yi, W. Li, W. 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)*, 2023, pp. 23 436–23 446.

[25] D. Yang, M. Li, D. Xiao, Y. Liu, K. Yang, Z. Chen, Y. Wang, P. Zhai, K. Li, and L. Zhang, "Towards multimodal sentiment analysis debiasing via bias purification," *arXiv preprint arXiv:2403.05023*, 2024.

[26] Y. Liu, Z. Xia, M. Zhao, D. Wei, Y. Wang, S. Liu, B. Ju, G. Fang, J. Liu, and L. Song, "Learning causality-inspired representation consistency for video anomaly detection," in *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 203–212.

[27] P. R. Burton, M. J. Murtagh, A. Boyd, J. B. Williams, E. S. Dove, S. E. Wallace, A.-M. Tasse, J. Little, R. L. Chisholm, A. Gaye *et al.*, "Data safe havens in health research and healthcare," *Bioinformatics*, vol. 31, no. 20, pp. 3241–3248, 2015.

[28] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[29] Y. Wang, Z. Ge, Z. Chen, X. Liu, C. Ma, Y. Sun, and L. Qi, "Explicit and implicit knowledge distillation via unlabeled data," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[30] Y. Wang, Z. Chen, D. Yang, P. Guo, K. Jiang, W. Zhang, and L. Qi, "Out of thin air: Exploring data-free adversarial robustness distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5776–5784.

[31] Y. Wang, D. Yang, Z. Chen, Y. Liu, S. Liu, W. Zhang, L. Zhang, and L. Qi, "De-confounded data-free knowledge distillation for handling distribution shifts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 615–12 625.

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

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

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

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

[36] Z. Hao, Y. Luo, H. Hu, J. An, and Y. Wen, "Data-free ensemble knowledge distillation for privacy-conscious multimedia model compression," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1803–1811.

[37] K. Do, H. Le, D. Nguyen, D. Nguyen, H. Harikumar, T. Tran, S. Rana, and S. Venkatesh, "Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation," *Advances in neural information processing systems*, 2022.

[38] J. Zhang, C. Chen, B. Li, L. Lyu, S. Wu, S. Ding, C. Shen, and C. Wu, "Dense: Data-free one-shot federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 21 414–21 428, 2022.

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

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

[41] G. Fang, K. Mo, X. Wang, J. Song, S. Bei, H. Zhang, and M. Song, "Up to 100x faster data-free knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6597–6604.

[42] K. Binici, N. T. Pham, T. Mitra, and K. Leman, "Preventing catastrophic forgetting and distribution mismatch in knowledge distillation via synthetic data," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 663–671.

[43] G. Patel, K. R. Mopuri, and Q. Qiu, "Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7786–7794.

[44] H. Chen, T. Guo, C. Xu, W. Li, C. Xu, C. Xu, and Y. Wang, "Learning student networks in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6428–6437.

[45] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

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

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

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

[49] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6210–6219.

[50] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[51] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[53] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[54] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[55] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *International Conference on Learning Representations*, 2020.

[56] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[57] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[58] H. S. Seung and D. D. Lee, "The manifold ways of perception," *science*, vol. 290, no. 5500, pp. 2268–2269, 2000.

[59] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowledge-Based Systems*, vol. 215, p. 106771, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121000344

[60] R. Horn and C. Johnson, "Matrix analysis cambridge university press, 1985," *Citation on*, p. 92.

[61] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[62] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[63] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[64] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[66] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[67] S. Yu, J. Chen, H. Han, and S. Jiang, "Data-free knowledge distillation via feature exchange and activation region constraint," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 266–24 275.

[68] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.