

Supplementary Materials for Sampling to Distill: Knowledge Transfer from Open-World Data

Anonymous Authors

In this supplementary material, we provide more details of our method, organized as follows:

- In Section 1, we further supplement the sampling-based methods so that readers can better understand and choose among multiple DFKD methods.
- In Section 2, we provide more training details of the used data augmentations method, corresponding to Section 4.1 of the main body. Besides, we set additional experiments about the loss trade-off parameters in Equation (8) of the main body and the prototype number per class in Section 3.2 of the main body. We compare some limited DFKD methods on the MNIST dataset and two data-based KD methods.
- In Section 3, we compare the update steps of various DFKD methods about the distillation computational complexity, according to the original papers and open-source code.
- In Section 4, we visualize the classification results of the CIFAR-100 and ImageNet datasets to further and more fairly prove the effectiveness of the proposed sampling method.

1 SUPPLEMENT TO THE SAMPLING-BASED METHOD

To better understand the proposed sampling-based method, we provide more explanations, including: (a) the similarities and differences between the sampling-based and the generation-based methods, (b) the motivation and purpose of the sampling-based methods, and (c) the impact on students' performance with a variety of sampled data from different datasets & the comparison of ours and other sampling-based methods. These contents may be helpful for readers to choose among different DFKD methods.

1.1 Sampling-based & Generation-based Methods

Similarities. Both sampling-based and generation-based methods require a necessary assumption that the pre-trained teacher model has learned the distribution of the original dataset, *i.e.*, the teacher model can make directional predictions on the original data rather than near random predictions like an initialized model. Under this assumption, the generation-based methods constrain the generator from synthesizing data that satisfies the teacher's predictive distribution. Like the above process, the sampling-based methods try to find data that more satisfies the teacher's prediction in the open-world dataset to fit the distribution.

Differences. The substitution data comes from different sources. For the generation-based methods, one or more learnable generators are introduced to synthesize the substitution data. The generators update together with the student model. The additional generators bring additional memory and gradient calculation costs (*e.g.*, mainstream methods introduce a thousand generators for the ImageNet dataset). For the sampling-based methods, since there are many semantically rich unlabeled open-world data available in reality,

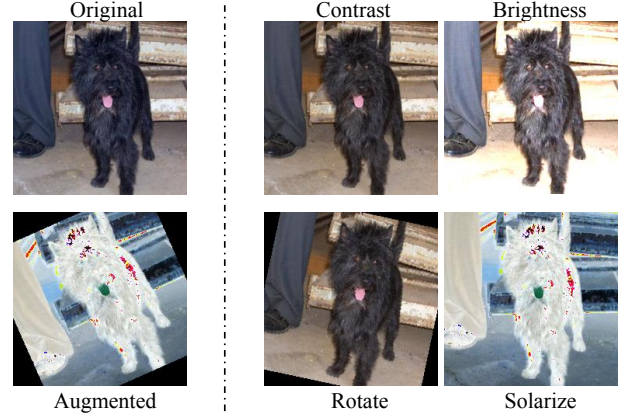


Figure 1: Visualization of RandAugment. The left is a schematic diagram of the original and augmented. The right is the results of each transform separately.

the appropriate unlabeled data is screened as substitution data. The above process overcomes the additional model update cost and thus has a faster speed.

1.2 The Motivation and Purpose of the Sampling-based Methods

In addition to avoiding unnecessary generation costs, the effectiveness of sampled data is the pursuit of different sampling methods, *i.e.*, how to sample less training data or train fewer epochs to get higher performance when the data range of the unlabeled sampling set is determined. The above goals are consistent and independent of how similar the unlabeled substitute data is to the test data. Even if there is a significant domain shift between the unlabeled substitute data and the original data, sampling-based methods still try to get the most helpful data from it.

In Table 2 of the main body, we show the impact of different numbers of sampled data (*i.e.*, 150k or 600k) on the student performance, which may be able to help readers weigh model performance and sampling cost.

1.3 Sampling from Different Datasets

Since different open-world datasets have different distributions with different semantic information, the teacher model has different familiarity with these datasets. For the current sampling-based methods, the sampled data entirely depends on the teacher's preference for unlabeled data. Therefore, different substitute datasets obviously affect the performance of students.

We compare three sampling-based methods (Mosaick [4], DFND [1], and our ODS) on the CIFAR-100 dataset. 150k or 600k unlabeled data are sampled from ImageNet [3], Places365 [12], and

Table 1: Student performance of different sampling-based methods with different substitute datasets.

Method	Unlabeled substitute datasets & the number of sampled data					
	ImageNet-150k	ImageNet-600k	Places365-150k	Places365-600k	Flickr1M-150k	Flickr1M-600k
Mosaick	74.59	75.91	73.86	74.75	73.79	74.94
DFND	74.20	74.42	73.37	74.25	73.56	74.46
ODSD	77.90	78.45	75.47	76.35	75.52	76.57

Flickr1M [6]. The teacher uses the ResNet-34 as its backbone, and the student uses the ResNet-18. As shown in Table 1, our ODSD outperforms other sampling-based methods in various settings, which proves the effectiveness of ODSD.

In the main body, we choose the same substitute datasets with the default settings in existing sampling-based methods [1] for fair comparison (see Table 1 of the main body for details). In addition, we have clearly marked the type (Sampling or Generation) of various methods in each table, which helps readers choose their favorable DFKD method. In general, we recommend trying our ODSD method first when open-world unlabeled data is available. It is faster than generation-based methods and performs competitively.

2 ADDITIONAL EXPERIMENTS

2.1 Data Augmentation

For the data augmentation involved, we combine four data transform methods: contrast, brightness, rotate, and solarize as shown in Figure 1. On the one hand, such a setting can reduce the computational cost of data transformation. On the other hand, it can also bring specific prediction difficulties to the teacher and student. In the contrastive task, prediction tasks will become difficult after augmentations. Furthermore, the representation quality of the network will be greatly improved. To reflect the simple-difficult data pair, we sample simple and studious data with a similar distribution to the original data and then apply data augmentation to highlight the complex and easy comparison. Based on this, the student can learn a differentiated knowledge representation to gain a better understanding ability.

Table 2: Diagnostic experiments on the total losses.

λ_1	Accuracy (%)		λ_2	Accuracy (%)	
	50k	150k		50k	150k
0	74.39	77.27	0	74.39	77.27
0.5	74.41	77.28	0.1	74.85	77.43
1	74.41	77.27	0.3	74.81	77.67
5	74.45	77.50	0.5	74.89	77.71
10	74.73	77.58	0.7	74.55	77.67
15	74.21	77.16	1	74.42	77.49
(a) $\mathcal{L}_{KD} + \lambda_1 \cdot \mathcal{L}_n$			(b) $\mathcal{L}_{KD} + \lambda_2 \cdot \mathcal{L}_c$		

2.2 Loss Trade-off Parameters λ

To verify the effectiveness and optimization parameters of the training objectives, we test the influence of \mathcal{L}_n and \mathcal{L}_c on students'

Table 3: Further analysis about $\mathcal{L}_c = \mathcal{L}_{c1} + \mathcal{L}_{c2}$.

λ	0	0.1	0.3	0.5	0.7	1
$\mathcal{L}_{KD} + \lambda \cdot \mathcal{L}_{c1}$	77.27	77.45	77.40	77.62	77.51	76.76
$\mathcal{L}_{KD} + \lambda \cdot \mathcal{L}_{c2}$	77.27	77.33	77.57	77.54	77.26	77.15

Table 4: Prototype number per class K .

Prototype		1	5	10	20	30	50
Accuracy (%)	50k	74.24	75.26	74.85	74.84	74.24	73.34
	150k	77.25	77.90	77.47	77.64	77.62	77.54

performance, respectively. We conduct the experiments on the CIFAR-100 dataset and use ResNet-34 as the teacher's backbone and ResNet-18 as the student's backbone. The student trains 200 epochs, and different data sampling numbers are set.

As shown in Table 2, the best experimental results have been shown with **bold**. We separately set the combination of \mathcal{L}_{KD} and the two proposed losses. Finally, we choose λ_1 as 10 and λ_2 as 0.5. Such parameter combination is also the default setting of other experiments. Further, as shown in Table 3, we carefully disassemble \mathcal{L}_c to study the impact of two separate parts \mathcal{L}_{c1} and \mathcal{L}_{c2} on student performance. The student performs well when $\lambda = 0.5$. For simplicity, we set $\lambda_2 = 0.5$.

Table 5: Student accuracy (%) on the MNIST dataset.

Methods	Type	Performance
Teacher	-	99.34
Student		98.97
[7]	Generation	92.47
DAFL [2]		99.16
DFAD [5]		98.90
ZSKD [8]		98.77
Wang <i>et al.</i> [11]		99.08
FEDGEN [13]		95.52
ODSD	Sampling	99.29

2.3 Prototype Number in Per Class K

The size of K reflects the complexity of intra-class modeling. When $K = 1$, the expression of the intra-class relationship is relatively simple, but sometimes it can not accurately exclude data with abnormal predictions. When K is particularly large, the average number

Table 6: “e”: epoch, “i”: iterations, “b”: batch size, “g”: the number of generators, and “n”: the number of data. The results of distillation computational complexity in various methods on various datasets are counted. The above values are based on the results reported in the original papers and the open-source codes.

Datasets	Methods	Type	DCC		Magnitude
			Generation module costs	Student module costs	
CIFAR	DAFL	Generation	2000e×120i×1024b	2000e×120i×1024b	5×10 ^{^8}
	DeepInv		2000e×256b	2000e×120i×1024b	2×10 ^{^8}
	ZSKT		80000e×128b	80000e×10i×128b	1×10 ^{^8}
	DFND	Sampling	-	600000n×200e	1×10 ^{^8}
	ODSD		-	600000n×200e	1×10 ^{^8}
ImageNet	DeepInv	Generation	3000i×84b×1000g	140000n×90e	3×10 ^{^8}
	DFD		10000i×64b×1000g	300000	6×10 ^{^8}
	Fast		400e×50i×128b×1000g	400e×1000i×100b	3×10 ^{^8}
	ODSD	Sampling	-	600000n×200e	1×10 ^{^8}
NYUv2	DFAD	Generation	300e×50i×64b	300e×250i×64b	6×10 ^{^6}
	ODSD	Sampling	-	200000n×20e	4×10 ^{^6}

of data at each prototype decreases, which leads to the problem of the prototype itself shifting. Table 4 reports our approach’s performance concerning the number of prototypes per class. Student performance does not always rise with the increase in the number of prototypes. Comprehensively considering student performance and calculation cost, we choose $K = 5$ and achieve the best performance. In this situation, our sampling method introduces 256.78s for sample selection for CIFAR100, which is almost negligible compared to the additional training of a customized generator.

2.4 Additional Experiments on MNIST Dataset

Considering that some DFKD methods are only applicable to simple datasets such as MNIST due to some limitations, we conduct additional comparative experiments on the MNIST dataset, which is composed of 28×28 pixel images from 10 classes (from 0 to 9). To make a fair comparison, we use LeNet-5 as the teacher model and LeNet-5-HALF as the student model, which is the same as the previous methods. For ODSD, the unlabeled dataset is the 64×64 ImageNet. The experimental results are shown in Table 5. Our method still achieves the SOTA performance. From the comparative experiments of the MNIST and ImageNet datasets, we can find that our method is applicable to multiple baselines of different scales, which shows that ODSD has good universality and competitiveness.

2.5 Discussion with CRD [10] & RKD [9]

We discuss the differences with existing data-based knowledge distillation methods separately. 1) For CRD [10], they focus on collecting the correlation between samples and categories in the dataset to assist the knowledge distillation process. Firstly, a large memory bank is continuously updated to contain as much knowledge about the entire dataset as possible while at the same time introducing additional storage and computing costs. Besides, for the DFKD task, the training and original data contain the domain shifts, and the data of multiple batches is variable (the generator’s update or the difference among sampled data). These issues may

Table 7: The comparison with CRD & RKD on CIFAR datasets. The teachers use ResNet-34, and the students use ResNet-18 as the backbones. GPU time indicates the training time of one epoch on a single RTX 3090 GPU. Memory indicates the memory usage with the batch size of 64.

Methods	CIFAR-10			CIFAR-100		
	Accuracy (%)	GPU time	Memory	Accuracy (%)	GPU time	Memory
CRD	93.56	274.25s	3184M	73.25	286.62s	3231M
RKD	92.46	136.34s	2117M	73.19	144.17s	2129M
ODSD	95.70	152.25s	2002M	78.45	160.72s	2012M

cause a biased memory bank, reducing CRD’s effectiveness for the DFKD task. Our method trains students in the current mini-batch, getting rid of the dependence on the consistency of the sample batch distribution in each item and reducing the computing and storage costs. 2) For RKD [9], they explore relationships among samples, which depend on the difference of samples in a mini-batch. However, in the DFKD task, the synthetic or sampled data is guided by the teacher’s model preferences (the generator’s update relies on the teacher’s pseudo-label and prediction & the sampling process refers to the teacher’s confidence). The preferences reduce the variance of samples in a batch, reducing RKD’s effectiveness. Our method explores the relationships of samples from multiple views and encourages the student to learn from both the teacher and the student itself, alleviating the dependence on training sample distribution similarity and sample quality.

In addition, we comprehensively compare these methods. The outstanding results in Table 7 demonstrate that our design has achieved comprehensive improvements on the DFKD task and also demonstrate a significant difference with existing methods.

3 UPDATE STEPS

For a long time, there has been a lack of a unified comparison index of update calculation in the field of DFKD. Only referring

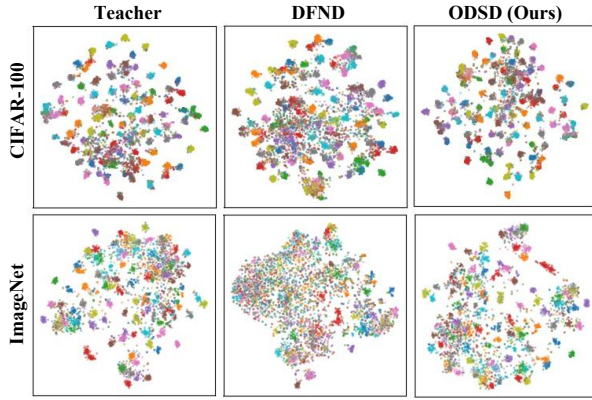


Figure 2: t-SNE visualization of the classification results on the CIFAR-100 and ImageNet datasets. Our proposed method obtains clearer clustering results, showing students have stronger learning abilities.

to the generation time of substitute data is not comparable to the method based on data sampling. Only considering the training time ignores the cost of generating substitute data. Comparing the total distillation time is computationally expensive, and the default batch size and the memory for generating data of some methods are enormous, making it difficult to reproduce accurately. For a more fair comparison, we introduce an evaluation metric of distillation computational complexity (DCC), which includes the cost of updating the generation module and the cost of updating the student module. DCC represents the total number of network updates referring to the unit of a single data, which means the algorithm's dependence on the number of data and iterations in the training process. It is a fair comparison to different algorithm types.

Table 6 shows the DCC metric of various algorithms on three datasets. For the CIFAR datasets, we sample the teacher-student backbone pairs with ResNet34-ResNet18, ResNet-50-ResNet-50 and ResNet-50-mobilenetv2 are selected on the ImageNet dataset and NYUv2 dataset, respectively. It is worth noting that the implementation details of some methods are missing or undetectable, so they do not appear in the table. We have tried our best to restore the distillation process of various methods. Although there is a lack of a unified evaluation method, fortunately, various methods seem to have reached a tacit understanding, making their DCC in the same order of magnitude gradient. At the same time, we also try to make our method consistent with previous work on this metric during the whole task.

4 ADDITIONAL VISUALIZATION

To further and more fairly prove the method's effectiveness, we compare the visualization results with the SOTA sampling-based method on the CIFAR-100 and ImageNet datasets. We sample 100 classes in the test set for the above two datasets. The aggregation degree of points reflects the learning degree of the network for the task. The results are shown in Figure 2. The students trained by our method have stronger learning abilities and can distinguish

different categories better. Our ODSD method can perform more effective knowledge mining and understanding than the DFND method [1].

REFERENCES

- [1] Hanting Chen, Tianyu Guo, Chang Xu, Wenshuo Li, Chunjing Xu, Chao Xu, and Yunhe Wang. 2021. Learning student networks in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6428–6437.
- [2] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. 2019. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3514–3522.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [4] Gongfan Fang, Yifan Bao, Jie Song, Xinchao Wang, Donglin Xie, Chengchao Shen, and Mingli Song. 2021. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems* 34 (2021), 11920–11932.
- [5] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. 2019. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006* (2019).
- [6] Mark J Huiskes and Michael S Lew. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. 39–43.
- [7] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. 2017. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535* (2017).
- [8] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. 2019. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*. PMLR, 4743–4751.
- [9] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3967–3976.
- [10] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.
- [11] Zi Wang. 2021. Data-free knowledge distillation with soft targeted transfer set synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10245–10253.
- [12] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.
- [13] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*. PMLR, 12878–12889.