Small for Small: Exploring Optimal Teacher in Knowledge Distillation with Limited Data

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Knowledge distillation (KD) is a widely used framework for transferring knowledge from a teacher model to a student model. While prior studies have mainly attributed teacher quality to accuracy or the teacher–student capacity gap, it remains unclear how the optimal teacher changes under limited distillation data, such as in *data pruning* scenarios. To address this gap, we systematically study teacher selection in low-data KD by varying teacher width, training stage, and output structure under data pruning. Our experiments on CIFAR-100 and ImageNet show that the optimal teacher strongly depends on the amount of available data: Smaller, less-confident, or early-epoch teachers outperform larger or fully trained teachers in low-data regimes. We further show that effective teachers in such cases exhibit similar properties in terms of their output distribution, particularly in non-target class predictions. Finally, we show that modifying non-target logits can improve KD performance without retraining the teacher.

1. Introduction

Knowledge distillation (KD) [5] is a standard framework for transferring knowledge from a high-capacity teacher model to a compact student model. Prior studies have shown that KD performance depends not only on teacher accuracy, but also on *capacity gap* between teacher and student [3, 9]. When teacher is excessively larger than student, student may fail to effectively imitate teacher distribution, motivating conventional view that teacher with capacity close to student is preferable.

However, this capacity-centric view becomes incomplete when distillation is performed with limited data. Recent works have shown that larger teachers can degrade student performance in such regimes, and that even teachers smaller than the student can be more effective [1, 6]. These observations suggest that teacher usefulness in low-data KD cannot be explained by accuracy or capacity alone. This issue is practically important because low-data distillation naturally arises when training data is limited by annotation cost, storage constraints, privacy restrictions, or task-specific data collection. In such settings, the student can access teacher supervision only on a small number of examples, so KD becomes highly sensitive to which teacher is used. Thus, identifying effective teachers under limited data is crucial for making KD reliable in realistic deployment scenarios.

Motivated by these observations, this paper presents an empirical analysis of teacher selection in low-data KD. Unlike prior works that only indirectly suggest the benefit of smaller teachers under limited data, we explicitly study how the optimal teacher changes across data regimes. We show that the optimal teacher shifts toward smaller or early-epoch teachers in low-data regimes, and connect this shift to teacher output distributions, especially non-target class predictions. The contributions of this paper are as follows:

- In low-data regimes, we observe that, teachers with *extremely* smaller than the student can outperform larger teachers, including the self-distillation.
- We find that early-epoch teachers, which are less confident on the training set, can outperform fully trained teachers under the low-data regimes.
- We analyze teacher output distributions and find that small teachers and early-epoch teachers exhibit similar predictive behavior. Teacher-output adjustment (e.g. temperature scaling) improves low-data KD, but remains less effective than directly using small or early-epoch teachers.

2. Preliminaries

Knowledge Distillation. We start by briefly reviewing the formulation of KD. For the task, we consider multiclass classification with a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where x_i is an input and y_i is its corresponding label. Let $f_s(\cdot)$ and $f_t(\cdot)$ denote the student and teacher networks, respectively. Then, given an input x_i, each network outputs a logit vector, denoted by $f_s(x_i)$ and $f_t(x_i)$. Using the softmax function with temperature parameter $\tau(> 0)$, the teacher and student output distributions are defined as

$$p_t^\tau(x_i) = \mathrm{softmax}(f_t(x_i)/\tau), \quad p_s^\tau(x_i) = \mathrm{softmax}(f_s(x_i)/\tau). \tag{1}$$

During training, the student is optimized by minimizing the following KD objective, while the teacher is frozen.

$$\mathcal{L}_{\mathrm{KD}} := \frac{1}{n}\sum_{i=1}^{n}\left[(1-\lambda)\ell_{\mathrm{CE}}(f_s(x_i), y_i) + \lambda\tau^2 \mathrm{KL}(p_t^\tau(x_i)\|p_s^\tau(x_i))\right]. \tag{2}$$

Data Pruning. We now introduce the *data pruning* scenario considered in this work. Let $\rho \in [0, 1)$ denote the data pruning rate, meaning that a fraction ρ of the training samples from $\mathcal{D}$ is "pruned" before distillation. The remaining dataset is denoted as $\mathcal{D}_\rho$, whose cardinality is

$$|\mathcal{D}_\rho| = (1-\rho) \times n. \tag{3}$$

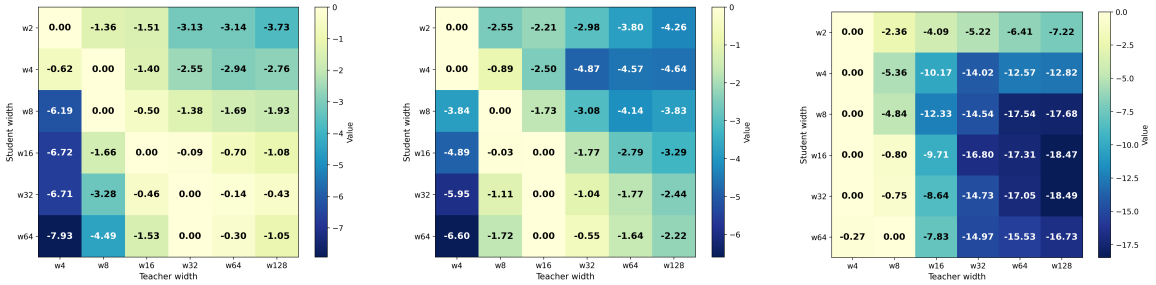
In this work, we study how the optimal teacher changes as the pruning rate ρ increases.

3. Experiments

In this section, we conduct experiments by differing by teacher scale, diverse pruning ratio, and adjusting teacher outputs for investigating the impact of teacher at diverse data regime.

Experimental Setup. To ensure a fair comparison with prior KD studies, we follow the standard hyperparameter recipes of widely used KD frameworks: RepDistiller [1, 10] for CIFAR-100 and MDistiller [11, 12] for ImageNet. We mainly conduct experiments using ResNet architectures with varying depths (10, 18, 34, and 50) and widths (2, 4, 8, 16, 32, 64 as the default width, and 128).

For CIFAR-100, student models are trained for 240 epochs with a batch size of 64. Here, we use SGD with momentum with an initial learning rate of 0.05, momentum of 0.9, and weight decay



(a) $\rho = 0$ (b) $\rho = 0.5$ (c) $\rho = 0.9$

Figure 1: Heatmaps of relative KD performance across student and teacher widths under different data-pruning rates at CIFAR-100. Rows and columns indicate student and teacher widths, respectively. All models use a ResNet-34 backbone with varying channel widths. Each cell reports the test-accuracy gap, in percentage points, comparing the best-performing teacher for the same student width and pruning rate. Results are averaged over 3 random seeds. Dataset pruning is performed by random selection while preserving class balance.

ρ	Student	Teacher			
		ResNet10	ResNet18	ResNet34	ResNet50
0.1	ResNet10	65.47 ±0.13	65.73 ±0.09	**66.00** ±0.09	65.92 ±0.06
	ResNet18	70.38 ±0.04	71.00 ±0.12	**71.47** ±0.08	71.08 ±0.03
0.4	ResNet10	64.46 ±0.05	64.59 ±0.17	**64.62** ±0.07	64.28 ±0.16
	ResNet18	68.90 ±0.12	69.41 ±0.06	**69.68** ±0.04	68.89 ±0.09
0.7	ResNet10	**60.91** ±0.23	60.85 ±0.09	60.68 ±0.21	59.68 ±0.14
	ResNet18	64.85 ±0.07	**64.99** ±0.04	64.43 ±0.09	63.17 ±0.03
0.9	ResNet10	**51.35** ±0.07	50.32 ±0.11	49.17 ±0.10	47.41 ±0.09
	ResNet18	**54.04** ±0.24	52.66 ±0.07	51.13 ±0.06	49.03 ±0.15

Table 1: Top-1 test accuracy (%) of students on ImageNet under different data-pruning rates. We vary both teacher and student depth using ResNet10, ResNet18, ResNet34, and ResNet50. **Bold** and underline indicate the best and second-best teacher for each student and pruning rate, respectively.

of 5×10^{-4}. The learning rate is decayed by a factor of 10 at epochs 150, 180, and 210. For the KL-divergence loss, we use a temperature of 4 and set the loss weight λ to 0.5.

For ImageNet, student models are trained for 100 epochs with a batch size of 512. We use SGD with an initial learning rate of 0.2, momentum of 0.9, and weight decay of 1×10^{-4}. The learning rate is decayed by a factor of 10 at epochs 30, 60, and 90. For the KL-divergence loss, we use a temperature of 1 and set λ to 0.5.

3.1. Teacher Capacity under Low-Data Regime

We observe that the optimal teacher changes depending on the data-pruning rate. In the full-data setting, Fig. 1(a) shows that teachers with capacities similar to the student tend to achieve better accuracy than substantially larger or smaller teachers. In contrast, Fig. 1(b) and Fig. 1(c) show

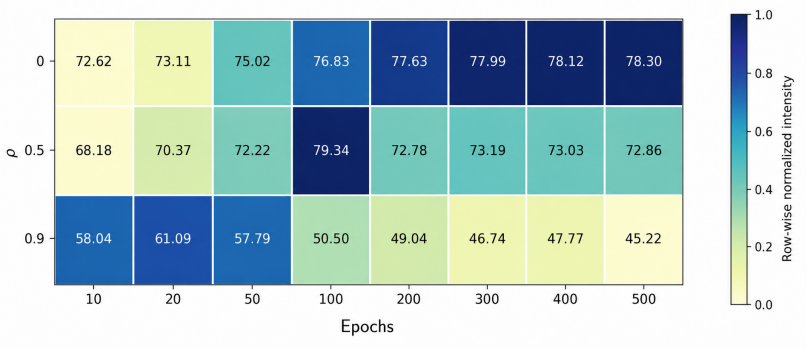


Figure 2: Heatmap of KD performance using teachers saved at different training epochs from the same training trajectory. Rows and columns correspond to drop rate ρ and teacher training epochs, respectively. Each cell reports the CIFAR-100 test accuracy of student, and colors are normalized within each row. KD is performed using ResNet-34 with channel width 64 in both teacher and student.

that the optimal teacher shifts toward smaller widths as ρ increases. Our results further show that this effect becomes more pronounced under severe data pruning: Even extremely small teachers (approximately 64$\times$ smaller than student) can outperform much larger teachers, with a performance gap of up to 18.49%p compared to larger teachers. This gap is much significant than what we observe in the full-data setting (i.e., $\rho = 0$).

Table 1 confirms the same tendency on ImageNet with depth-scaled ResNets. Under lower ρ, ResNet34 is generally the best teacher, whereas under higher ρ, the optimal teacher shifts to ResNet10 or self-distillation (i.e., the teacher and student share the same architecture). This indicates that the preferred teacher capacity decreases as the available distillation data becomes smaller, and that overly large teachers can hurt student performance in low-data regimes.

Takeaway 1. In low-data regimes, the optimal teacher for KD is not necessarily the largest or closest-capacity model; instead, smaller teachers can be more effective as the available distillation data becomes limited.

3.2. Teacher Confidence under Low-Data Regime

We next study the effect of teacher training stage by distilling from checkpoints saved at different epochs along the same teacher training trajectory. Fig. 2 shows that, in low-data regimes, the best teacher is not necessarily the fully trained teacher. When $\rho = 0.9$, the student achieves the highest accuracy with a teacher saved at an early training epoch, and the gap between the best and worst teacher reaches 15.87%p points. This suggests that, under severe data pruning, an early-epoch teacher with less confident predictions can provide more effective distillation signals than a fully trained teacher. In contrast, as more training data becomes available, the optimal teacher shifts toward later training epochs: the best teacher appears at epoch 100 when $\rho = 0.5$, and near the end of training when $\rho = 0$. These results indicate that the optimal level of teacher confidence depends strongly on the amount of available distillation data.

Takeaway 2. The optimal teacher confidence depends on the amount of available distillation data: early-stage, less-confident teachers are preferable under severe data pruning.

3.3. Output Adjustment Toward Low-Data-Friendly Teacher Distributions

Results in Sections 3.1 and 3.2 show that small teachers and early-epoch teachers are more effective than the fully trained large teacher in low-data regimes. We first examine whether these two types of teachers share similar output behavior. As shown in Table 2, the non-target outputs of the small teacher and the early-epoch teacher are more similar to each other than either is to the fully trained base teacher. Moreover in Appendix B, although the small teacher and the early-epoch teacher differ in model capacity and training trajectory, their output histograms exhibit similar distributional patterns, especially over non-target classes. This suggests that small and early-epoch teachers may provide a common form of non-target supervision that is more suitable when distillation data is limited. These results motivate a practical question: when only a fully trained large teacher is available, can we recover benefit of small or early-epoch teachers by directly adjusting its output distribution?

Model Pair	Cosine Similarity
Base $\leftrightarrow$ Small	0.721
Base $\leftrightarrow$ Early-epoch	0.684
Small $\leftrightarrow$ Early-epoch	0.828

Table 2: Non-target cosine similarity between teacher output distributions. Here, Base denotes ResNet34 with width 64 trained for 300 epochs, Small denotes ResNet34 with width 8, and Early-epoch denotes ResNet34 with width 64 trained for 20 epochs.

To answer this question, we compare two post-hoc output adjustment methods: global temperature scaling and non-target logit scaling. Following prior observations on temperature scaling [1], we find that global temperature adjustment yields only limited gains and fails to close the gap to small or early-epoch teachers, as shown in Table 3. In contrast, in Table 4, scaling only the non-target logits [7] substantially improves low-data KD without retraining the teacher, although it still remains less effective than directly using low-data-friendly teachers.

Remark. While non-target logit scaling improves KD performance, it does not fully match the gains obtained from small or early-epoch teachers in Sections 3.1 and 3.2. This indicates that low-data-friendly teachers are not characterized by entropy or sharpness alone. Since scaling the non-target logits with a positive coefficient preserves their ranking, it can reshape the output distribution but cannot recover a different class-confusion structure. Indeed, Table 6 shows that the top-2 non-target class changes for more than 70% of training samples during teacher training. Thus, output scaling can mimic the distributional shape of early-epoch teachers, but not their non-target ranking structure, which may explain the remaining performance gap.

4. Conclusion

In this work, we study which teachers are effective for knowledge distillation when the amount of distillation data is limited. Our results show that the conventional capacity-centric view of KD becomes insufficient in low-data regimes: smaller teachers and early-epoch, less-confident teachers can outperform larger or fully trained teachers when data is severely pruned. By analyzing teacher output distributions, we find that these effective teachers share similar predictive behaviors, especially in their non-target class distributions. Motivated by this observation, we show that scaling non-target logits can improve low-data KD without retraining the teacher, although it cannot fully reproduce the benefits of small or early-epoch teachers. Future work should provide a more complete characterization of low-data-friendly teachers beyond entropy or sharpness, especially by understanding how their non-target class rankings and class-confusion structures affect distillation.

References

[1] Emanuel Ben-Baruch, Adam Botach, Igor Kviatkovsky, Manoj Aggarwal, and Gérard Medioni. Distilling the knowledge in data pruning. In *International Conference on Machine Learning*, 2025.

[2] Yudong Chen, Xuwei Xu, Frank de Hoog, Jiajun Liu, and Sen Wang. Medium-difficulty samples constitute smoothed decision boundary for knowledge distillation on pruned datasets. In *The Thirteenth International Conference on Learning Representations*, 2025.

[3] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[4] Yeseul Cho, Baekrok Shin, Changmin Kang, and Chulhee Yun. Lightweight dataset pruning without full training via example difficulty and prediction uncertainty. In *International Conference on Machine Learning*, 2025.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[6] Giulia Lanzillotta, Felix Sarnthein, Gil Kur, Thomas Hofmann, and Bobby He. Revisiting knowledge distillation: The hidden role of dataset size. *arXiv preprint arXiv:2510.15516*, 2025.

[7] Xin-Chun Li, Wen-Shu Fan, Shaoming Song, Yinchuan Li, Shao Yunfeng, De-Chuan Zhan, et al. Asymmetric temperature scaling makes larger networks teach well again. *Advances in neural information processing systems*, 35:3830–3842, 2022.

[8] Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity difficulty in data pruning. In *International Conference on Learning Representations*, 2024.

[9] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[10] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

[11] 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*, pages 11953–11962, 2022.

[12] Borui Zhao, Quan Cui, Renjie Song, and Jiajun Liang. Dot: A distillation-oriented trainer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6189–6198, 2023.

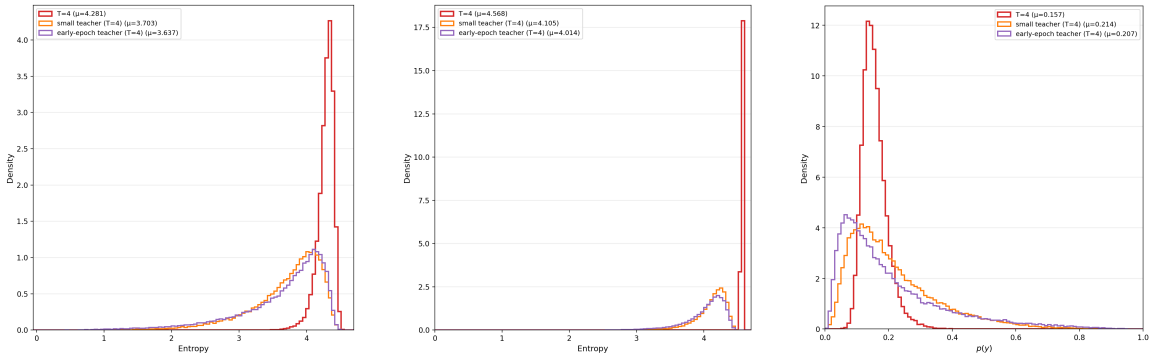
Appendix A. Related Work

Knowledge distillation and teacher capacity. Knowledge distillation (KD) transfers knowledge from a high-capacity teacher model to a smaller student by using the teacher's soft predictive distribution as supervision [5]. Beyond teacher accuracy, prior work has shown that the effectiveness of KD depends strongly on the capacity gap between the teacher and the student. For example, Cho and Hariharan [3] observed that an overly large teacher can be difficult for a small student to imitate, and Mirzadeh et al. [9] addressed this issue by introducing teacher assistants that bridge the capacity gap. These works motivate the common view that a teacher with capacity close to the student is often preferable to an excessively large teacher. In contrast, our work shows that this capacity-centric view is incomplete in low-data regimes: as the available distillation data becomes limited, even teachers much smaller than the student can become more effective than larger or closest-capacity teachers.

Knowledge distillation under data pruning and low-data regimes. Recent studies have investigated KD when only a subset of the original training data is available. Ben-Baruch et al. [1] showed that KD can be particularly useful in data pruning settings, where the student is trained on a reduced dataset rather than the full training set. Similarly, Chen et al. [2] studied the role of medium-difficulty samples and argued that carefully selected pruned data can provide effective supervision for distillation. These works mainly focus on how to select useful training samples for KD under data pruning. Our work takes a complementary direction: instead of only improving the pruned dataset, we study how the choice of teacher changes across data regimes. We show that, under severe data pruning, selecting an appropriate teacher can be as important as, or even more important than, improving the pruning strategy itself.

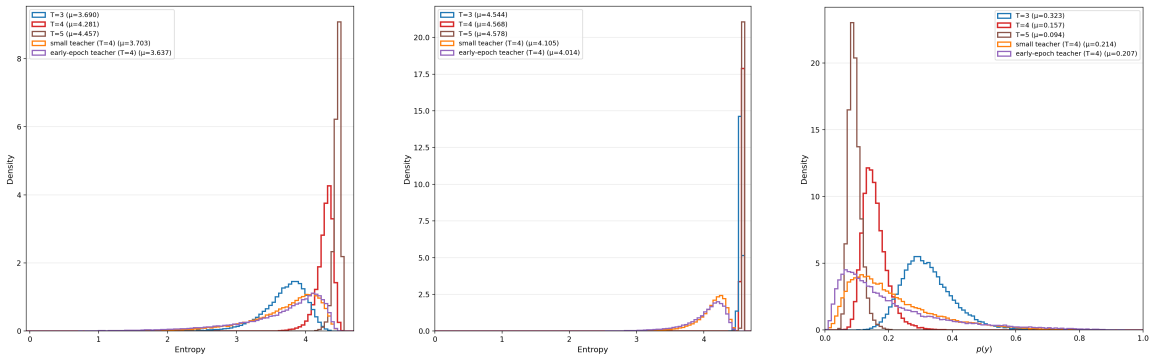
Modifying teacher outputs for distillation. Another line of work improves KD by modifying the teacher's output distribution. Temperature scaling is widely used to soften teacher predictions, allowing the student to learn from inter-class similarities rather than only from hard labels [5]. More recently, Li et al. [7] proposed asymmetric temperature scaling, which adjusts target and non-target logits differently to make large teachers more suitable for distillation. Related methods such as decoupled KD also emphasize that target and non-target knowledge play different roles in the distillation objective [11]. Our work revisits this perspective in low-data regimes. We find that simply changing the global teacher temperature is insufficient to explain the advantage of small or early-epoch teachers. Instead, modifying the non-target logit structure provides a more effective way to induce low-data-friendly supervision, although it still cannot fully recover the benefits of naturally small or early-epoch teachers.

Appendix B. Analysis of Diverse Teacher



(a) Softmax entropy (b) Non-target class softmax entropy (c) Target class probability

Figure 3: Distributions of teacher output statistics on the CIFAR-100 training set The thick orange curve corresponds to the best small teacher from Section 3.1, ResNet34 with width 8, and the thick purple curve corresponds to the best early-epoch teacher from Section 3.2, ResNet34 with width 64 trained for 20 epochs. We report (a) softmax entropy, (b) non-target class entropy after removing the target-class probability and re-normalizing the remaining probabilities, and (c) target class probability.

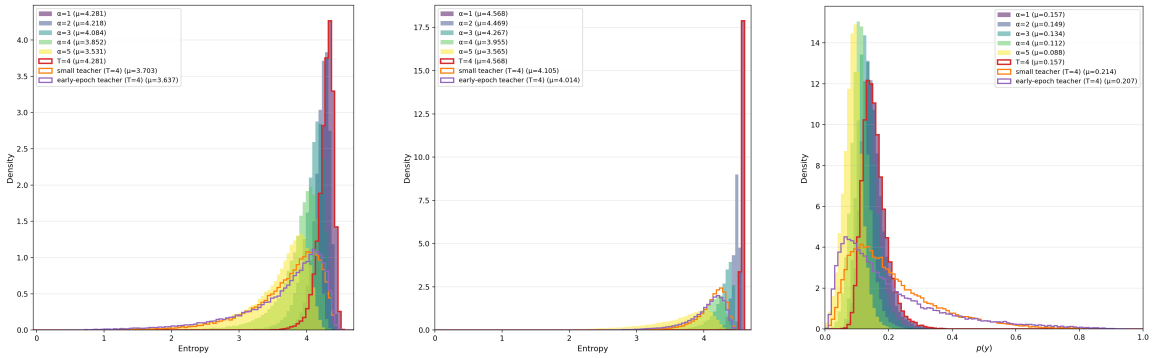


(a) Softmax entropy (b) Non-target class softmax entropy (c) Target class probability

Figure 4: Output-statistic distributions under temperature scaling. We report the same statistics as in Fig. 3, while sweeping the scaling coefficient τ applied to teacher output logits during KD.

Fig. 3 visualizes the entropy of teacher output probabilities and further decomposes it into non-target class entropy. We observe that small teachers and early-epoch teachers exhibit highly similar output distributions, even though they differ in model width and training trajectory. This suggests that there may exist a common teacher-output property that makes a teacher more effective in low-data regimes.

To examine whether such output statistics can be reproduced without changing the teacher itself, we modify the teacher logits before computing the soft target distribution. For simplicity, let z denote the teacher logit vector $f_t(\cdot)$. For temperature scaling, we apply the standard transformation



(a) Softmax entropy (b) Non-target class softmax entropy (c) Target class probability

Figure 5: Output-statistic distributions under non-target logit scaling. We report the same statistics as in Fig. 3, while sweeping the scaling coefficient α applied to the non-target logits of the fully trained teacher.

$q = \mathrm{softmax}(z/T)$. For non-target logit scaling, we keep the target logit unchanged except for temperature scaling, while additionally scaling only the non-target logits:

$$\tilde{z}_y = z_y/T, \qquad \tilde{z}_k = \alpha z_k/T \quad \text{for } k \neq y,$$

and compute $q = \mathrm{softmax}(\tilde{z})$. Thus, α directly controls the relative structure among non-target classes, whereas T globally scales the entire teacher logit vector.

Fig. 4 and Fig. 5 show how the teacher-output distribution changes when we modify the teacher logits by temperature scaling and non-target logit scaling, respectively. A notable observation is that simply scaling the entire logit vector with temperature does not reproduce the output statistics of the small or early-epoch teachers that perform well in low-data KD. Although temperature scaling changes the overall softmax entropy, its effect on the non-target class entropy is relatively limited; in particular, the non-target distribution does not shift toward the pattern observed in the best-performing small and early-epoch teachers.

In contrast, selectively scaling only the non-target logits produces a more targeted distributional shift. With an appropriate value of α, non-target logit scaling makes the teacher outputs more closely resemble those of the small and early-epoch teachers, especially in terms of non-target class entropy. This suggests that the effectiveness of these teachers in low-data regimes may not be explained solely by their overall confidence, but is closely related to how they structure probability mass among the non-target classes.

Appendix C. Adjusting Teacher Output during Training

ρ	τ_t								
	0.5	1	2	4 (base)	8	16	32	64	128
0	78.02	77.83	79.24	**80.04**	79.85	79.59	79.25	79.2	79.45
0.5	69.60	69.20	72.42	**73.67**	73.16	72.89	72.60	72.00	71.98
0.9	33.53	32.88	37.47	**43.85**	43.06	42.84	40.73	40.52	41.55

Table 3: Student accuracy (%) under different teacher temperatures τ_t and data drop rates ρ on CIFAR-100. For each drop rate, we vary only the teacher temperature while keeping the teacher model fixed. Bold entries indicate the best accuracy for each drop rate.

	Non-target logit scale α					
ρ	0.5	1 (base)	2	3	4	5
0	79.80	80.04	**80.46**	79.93	78.28	74.86
0.5	73.25	73.67	75.32	**75.82**	74.92	72.07
0.9	39.35	43.85	46.89	**50.84**	49.70	49.82

Table 4: Top-1 accuracy of the student when scaling the teacher's non-target-class logits with α during training.

Table 3 and Table 4 report the KD performance obtained by adjusting the teacher outputs, motivated by the observations in Appendix B. Table 3 shows that varying only the teacher temperature does not improve over the standard setting: the best accuracy is consistently achieved at $\tau_t = 4$, which is the commonly used temperature for CIFAR-100 KD, across all drop rates ρ.

In contrast, Table 4 shows that scaling the non-target logits can substantially improve KD performance, especially in the low-data regime. While the base setting $\alpha = 1$ is already competitive when the full dataset is used, larger values of α become more effective as the drop rate increases. In particular, $\alpha = 3$ yields the best performance at $\rho = 0.5$ and $\rho = 0.9$, improving the student accuracy from 73.67 to 75.82 and from 43.85 to 50.84, respectively. These results suggest that directly reshaping the non-target logit distribution provides more effective supervision than global temperature scaling. This is consistent with Appendix B, where non-target logit scaling better matches the output statistics of small and early-epoch teachers by simplifying the non-target class distribution.

Appendix D. Comparing with Data Pruning Algorithms

Teacher	Pruning Method	ρ					
		0.1	0.3	0.5	0.7	0.8	0.9
T:ResNet34w64	Random	78.73	76.62	73.20	67.13	61.83	46.04
	DUAL [4]	**79.26**	**78.35**	**75.63**	70.29	64.54	51.22
	D2 [8]	<u>79.04</u>	<u>77.60</u>	74.82	69.95	66.15	53.96
	MDSLR [2]	78.50	76.73	73.52	67.14	60.52	42.54
T:ResNet34w128	Random	78.49	75.87	72.50	66.20	60.82	44.81
T:ResNet34w32		78.80	77.16	<u>75.06</u>	68.88	63.81	47.84
T:ResNet34w16		78.12	77.17	75.08	**71.31**	<u>67.49</u>	54.10
T:ResNet34w8		75.33	74.68	73.61	<u>71.30</u>	**69.16**	**62.62**
T:ResNet34w4		71.70	70.68	68.53	65.79	63.78	<u>59.33</u>

Table 5: Student accuracy (%) under different data pruning ratios ρ on CIFAR-100. The upper block compares pruning methods using the same teacher, ResNet34 with width 64. The lower block reports the effect of teacher width when random pruning is used. Bold and underlined values indicate the best and second-best performance for each pruning ratio, respectively. Student is ResNet18 with width 64.

Table 5 further highlights that improving the pruning strategy alone is not sufficient to fully address the performance degradation in low-data regimes. Although DUAL [4] and D^2 [8] consistently improve over random pruning when using the same teacher, their gains become limited as the drop rate increases. In contrast, the lower block shows that changing the teacher can lead to a much larger improvement under severe data pruning. For example, at $\rho = 0.9$, using a smaller teacher substantially outperforms the large ResNet34 with width 64 teacher, even when advanced pruning methods are applied. This suggests that, in low-data distillation, the effectiveness of KD depends critically on selecting an appropriate teacher, rather than relying only on better data subset selection. Therefore, identifying and using an optimal teacher becomes especially important for improving KD efficiency when only limited training data is available.

Appendix E. Sample-wise Observation of Teacher Prediction

Epochs	20	50	100	200	300	400	500
20	0.000	0.683	0.727	0.726	0.719	0.722	0.734
50	-	0.000	0.679	0.718	0.707	0.711	0.733
100	-	-	0.000	0.712	0.724	0.716	0.750
200	-	-	-	0.000	0.697	0.700	0.726
300	-	-	-	-	0.000	0.682	0.702
400	-	-	-	-	-	0.000	0.708
500	-	-	-	-	-	-	0.000

Table 6: Pairwise ratio of samples whose top-2 predicted class changes across teacher training checkpoints. Rows and columns indicate teacher epochs. Experiments are conducted on CIFAR-100 using a ResNet-34 teacher with width 64.

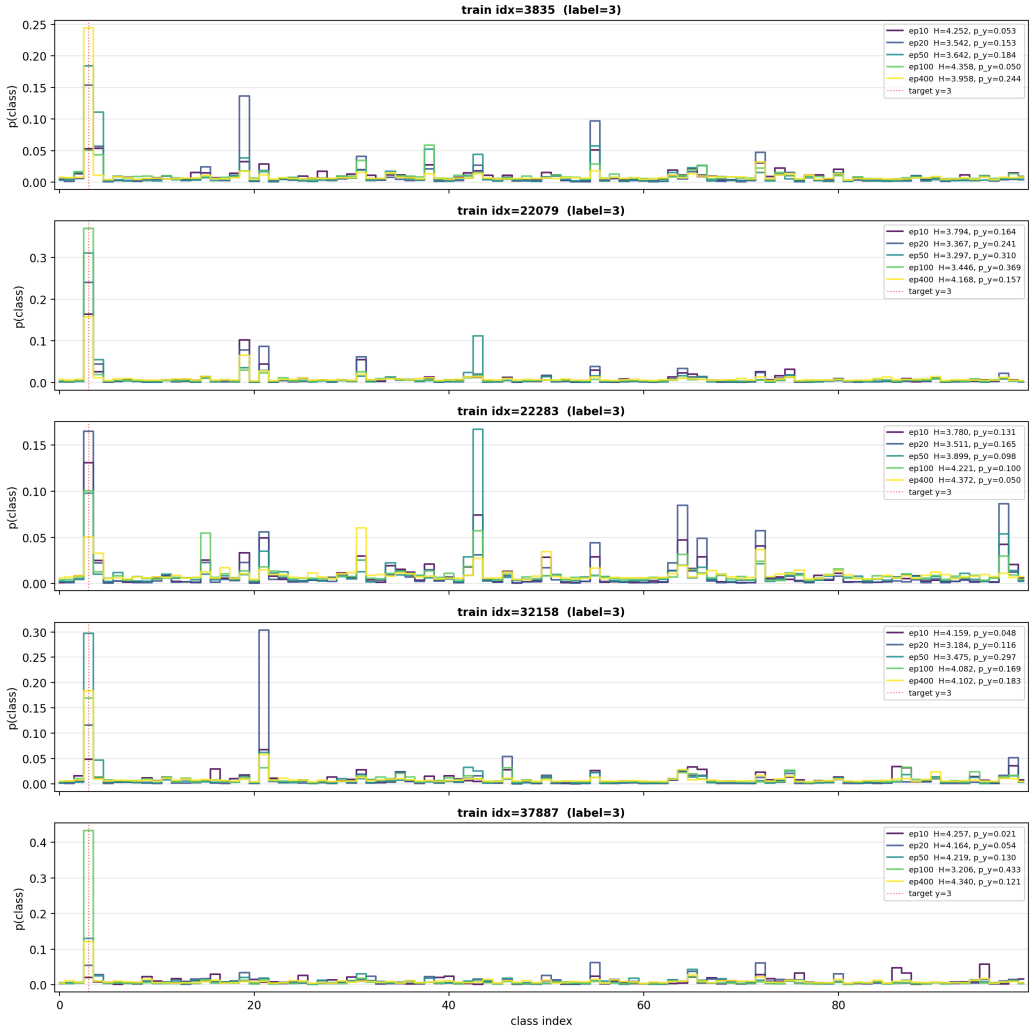


Figure 6: Sample-wise visualization of teacher output distributions across training epochs. Each subplot shows the predicted class probabilities for a single training sample, where different colors correspond to teacher checkpoints at different epochs. The dashed vertical red line indicates the ground-truth class. Although the target class remains the same, the probability mass assigned to non-target classes changes substantially during training, indicating that the non-target class ranking is not preserved along the teacher learning trajectory.