

Equally Critical: Samples, Targets, and Their Mappings in Datasets

Runkang Yang¹ Peng Sun^{2,3} Xinyi Shang⁴ Yi Tang⁵ Tao Lin^{3,*}
¹ShanghaiTech University ²Zhejiang University ³Westlake University
⁴University College London ⁵South China University of Technology

Abstract

Data inherently possesses dual attributes: samples and targets. For targets, knowledge distillation has been widely employed to accelerate model convergence, primarily relying on teacher-generated soft target supervision. Conversely, recent advancements in data-efficient learning have emphasized sample optimization techniques, such as dataset distillation, while neglected the critical role of target. This dichotomy motivates our investigation into understanding how both sample and target collectively influence training dynamic. To address this gap, we first establish a taxonomy of existing paradigms through the lens of sample-target interactions, categorizing them into distinct sample-to-target mapping strategies. Building upon this foundation, we then propose a novel unified loss framework to assess their impact on training efficiency. Through extensive empirical studies on our proposed strategies, we comprehensively analyze *how variations in target and sample types, quantities, and qualities* influence model training, providing **six key insights** to enhance training efficacy.

1 Introduction

The remarkable progress of deep learning has been largely propelled by the availability of massive amounts of data [9, 22, 25]. Yet, the question of how to effectively utilize data to maximize model training efficiency still remains a fundamental challenge [1, 6]. As the cornerstone of model training, data inherently possesses two fundamental attributes: samples and targets². Among these, knowledge distillation [23] has been proposed to effectively accelerate model convergence using soft targets generated by pre-trained teachers, and has gained huge success to drive the unprecedented development of lightweight models like DistilBERT [37] and DeepSeek-R1 Series [16]. Sample optimization, on the other hand, has been a central focus of data-efficient learning [54, 34] driven by the increasing demand for training models under limited data. Recent studies [39, 12] have explored strategies that can leverage smaller datasets without compromising performance, like data pruning strategies [31] can achieve competitive performance

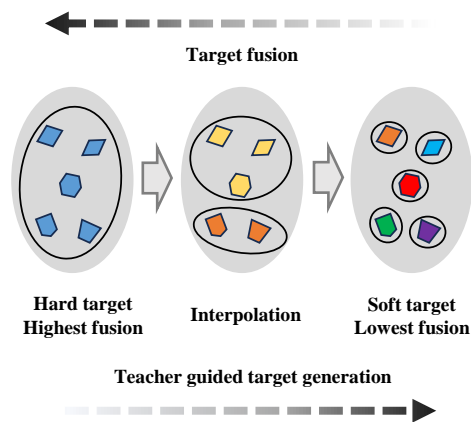


Figure 1: The key difference between hard and soft targets lies in **target fusion**: hard targets map all (*augmented*) samples of the same class to a shared target, while soft targets assign each one a unique target.

*Corresponding author.

²In this paper, we decouple dataset into two essential components: samples X and targets Y . For example, in a classification dataset, X represents the images, while Y consists of the one-hot encoded targets.

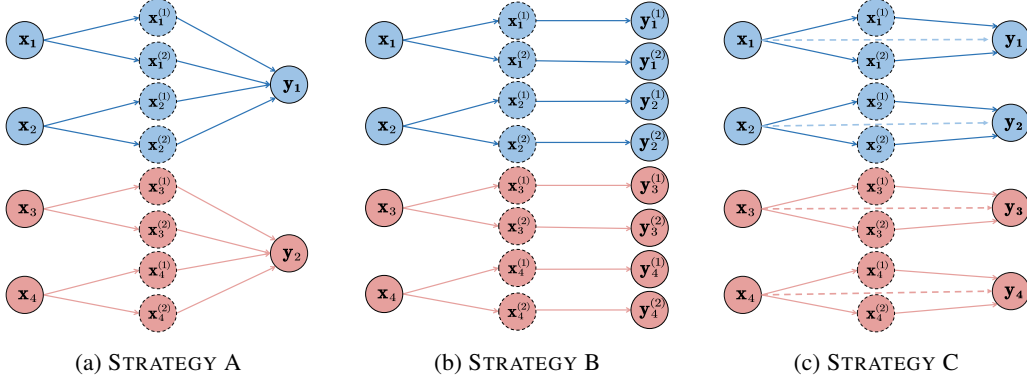


Figure 2: **Three different sample-to-target mapping strategies.** **STRATEGY A:** Multiple augmented samples within the same class are mapped to one same one-hot target. **STRATEGY B:** Each augmented sample is mapped to a unique soft target. **STRATEGY C:** Multiple augmented views of a sample are mapped to one same soft target. Blue and red colors denote class 1 and class 2, respectively. For each strategy, x_1 and x_2 means original samples. The middle column illustrates the augmented samples produced by various augmentation strategies. On the left, y_1 and y_2 represent the targets associated with each augmented sample.

with reduced training budgets by focusing on the most informative samples. Optimization on samples also include training using synthetic [12] and augmented [13] images.

Despite these achievements, *prior research have predominantly treated samples and targets as independent entities*. Advancements in data-efficient learning methods [41, 46, 17] have highlighted the importance of both samples and targets. For instance, efficient dataset distillation methods heavily rely on soft targets generated by pre-trained models [41]. Furthermore, [40] provide theoretical insights demonstrating that both targets and samples within a dataset significantly influence training dynamics and convergence. A detailed discussion of related work can be found in [Appendix A](#).

In this work, we demonstrate that in supervised learning, an implicit and inherent correspondence exists between samples and their associated targets, characterized by the degree of target fusion shared among different samples. Hard targets indicate that all samples within the same class share an identical target, while soft targets represent the opposite, as illustrated in [Figure 1](#). However, hard and soft targets merely represent two cases of this relationship, and how this correspondence affects model training still remain underexplored.

To enable a systematic examination of samples and targets for model training, we first review existing paradigms in the sample-target relationship, categorizing them into distinct sample-to-target mapping strategies: multiple-to-one mappings (typical in conventional supervised learning), multiple-to-multiple mappings (common in knowledge distillation [23] using teacher models), and the proposed multiple-to-few mappings, as detailed in [Section 2.1](#), we then propose a unified loss framework to evaluate their impact on model training in [Section 2.2](#). Building on this foundation, we take computer vision task as a case study and conduct a comprehensive analysis of *how variations in target and sample types, quantities, and qualities affect training efficiency*. Our analysis covers: (a) diverse target generation strategies in [Section 3.1](#) (target quantity), (b) teacher models with different performance in [Section 3.2](#) and various augmentation in [Section 3.3](#) (target quality), and (c) sample quantity and quality in [Section 4](#). We list our **six key findings** below:

- Under the same training budge, decoupling backbone and classifier for separate training can greatly benefit downstream tasks.
- Stronger teacher models enhance final performance, but weak ones aid early learning.
- Our proposed STRATEGY C can achieve higher final accuracy compared to traditional knowledge distillation, while retaining the ability to accelerate early-stage training.
- Teacher models trained with augmentations like MixUp do not consistently improve student performance, while RandomResizedCrop adversely affects student training.
- With limited data, soft targets provide advantages, but larger sample sizes favor one-hot targets.
- RandomResizedCrop is most effective for one-hot targets, while Mixed-based augmentation excels with soft targets.

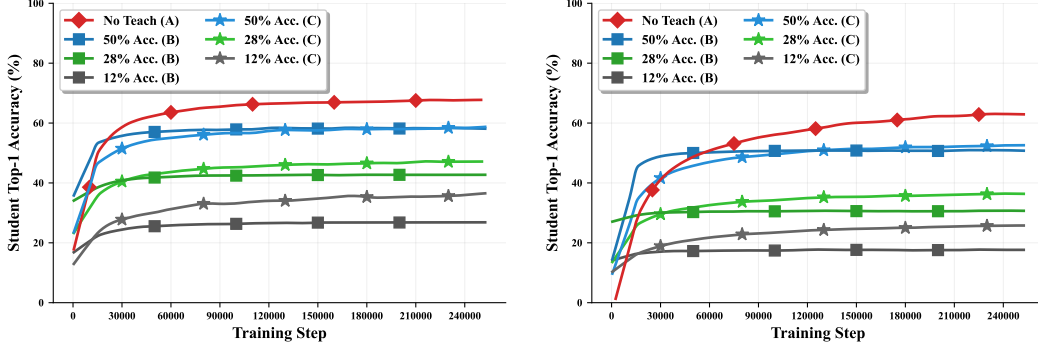


Figure 3: **Student training performance under three strategies using ResNet50 (Left) and ViT (Right) as backbone on ImageNet.** Under different backbones, STRATEGY A shows long-term advantages, while STRATEGY B exhibits short-term benefits. And the advantages of STRATEGY C become increasingly prominent when applied to weaker teacher models.

2 Preliminary

This section begins by formally defining three mapping strategies between samples X and targets Y . Subsequently, we critically analyze the limitations of existing loss functions in evaluating the impact of these mapping strategies on the representational capacity of models. We then introduce a novel unified loss function.

2.1 Definition of Mapping Strategies

Previous studies have demonstrated that data-efficient techniques, such as knowledge distillation and dataset distillation, which reconstruct the relationship between samples X and targets Y , can significantly accelerate model training. For instance, knowledge distillation leverages soft targets generated by teacher models to expedite the training process [45]. To thoroughly examine the relationship between samples X and targets Y , we define mapping strategy, denoted as $\psi : \mathcal{X} \rightarrow \mathcal{Y}$, where \mathcal{X} represents the data space and \mathcal{Y} the target space. Importantly, augmentations are crucial in deep learning, significantly boosting training effectiveness [50, 13]. Consequently, this paper emphasizes the use of augmented samples.

STRATEGY A: Mapping multiple augmented samples within the same class into a one-hot target.

In conventional representation learning, data pairs (X, Y) represent a non-injective mapping, denoted as STRATEGY A, as shown in Figure 2a. For example, for a classification task, one class c may include multiple samples X_c , indicating that multiple samples are mapped to a same class. The mapping strategy is formalized as $\psi_A(\mathbf{x}_i^{(j)}) = c$, where j denotes the j -th augmented sample of the original sample \mathbf{x}_i .

STRATEGY B: Mapping each augmented sample within the same class into a unique soft target.

For efficient knowledge distillation and dataset distillation, these approaches reconstruct the sample-to-target relationship using more informative soft targets. Each augmented sample corresponds to a unique soft target generated by a pre-trained model, establishing an injective mapping from augmented samples to soft labels, as shown in Figure 2b. Specifically, the mapping strategy is formalized as: $\psi_B(\mathbf{x}_i^{(j)}) = \mathbf{y}_i^{(j)}$, where $\mathbf{y}_i^{(j)}$ represents the soft target generated by a pre-trained model for the augmented sample $\mathbf{x}_i^{(j)}$. It is obvious that STRATEGY B introduces soft targets, which includes more information, thereby accelerating the training process of the model.

STRATEGY C: Mapping multiple augmented views of one sample into a same soft target.

While STRATEGY B provides valuable information through soft targets, it inevitably introduces noise by associating a single sample with multiple labels, potentially resulting in sub-optimal results. To mitigate this issue while preserving the advantages of soft targets, we introduce a novel mapping strategy, STRATEGY C. This approach associates all augmented samples of an original sample with a single soft target, as shown in Figure 2c. It establishes a non-injective mapping from augmented samples to targets while maintaining an injective mapping from original samples to targets.

Specifically, the mapping function is defined as $\psi_C(\mathbf{x}_i^{(j)}) = \mathbf{y}_i$, where \mathbf{y}_i is generated by a pre-trained model for the original sample \mathbf{x}_i . It preserves the rich information of soft targets while minimizing the noise introduced by multiple soft targets, which effectively balances the consistency of STRATEGY A with the variability of STRATEGY B.

Table 1: **Comparison between the deconstructed loss function designed in our paper and traditional loss function.** The result shows the student performance using STRATEGY B under different teachers. Combining the backbone and classifier into a unified framework with a standard loss function will cause the accuracy of the student model consistently lower than the corresponding teacher. While with our proposed decoupled training framework, the student model can outperform the teacher, especially for weak teachers. (*dataset: CIFAR-10, backbone: ResNet-18, training step: 50k*)

Loss	Teacher Top-1 Accuracy (%) $\pm 1\%$								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
Standard	9.97 \pm 0.3	19.41 \pm 0.6	28.68 \pm 0.4	39.92 \pm 0.3	50.01 \pm 0.3	59.79 \pm 0.4	69.90 \pm 0.4	79.96 \pm 0.5	89.52 \pm 0.2
Ours	35.96 \pm 0.4	39.70 \pm 0.4	41.61 \pm 0.5	46.11 \pm 0.3	57.18 \pm 0.4	63.90 \pm 0.5	72.58 \pm 0.2	82.93 \pm 0.4	90.35 \pm 0.3
Gain	3.61	2.04	1.45	1.16	1.14	1.07	1.04	1.04	1.01

2.2 Design A Unified Loss Function

Having defined three distinct sample-to-target mapping strategies, we subsequently investigate their impact on the representational capacity of models. Following previous benchmarks and research [19, 4, 15], the representation ability of a trained model can be evaluated using an offline linear probing strategy, which involves decoupling the model into a backbone and a classifier. However, two inherent challenges emerge when employing conventional loss functions: (1) They inherently integrate these mapping strategies during training, complicating the evaluation of individual strategies. (2) They obscure the backbone’s representational ability due to classifier influence. Therefore, in this subsection, we first review conventional loss functions and analysis and their limitations. Subsequently, we propose a novel unified loss function to mitigate these challenges.

2.2.1 Conventional Loss Functions

Cross-entropy (CE) loss is widely recognized as a standard objective for training deep neural networks on classification tasks. However, CE loss continues to propagate one-hot target information to the model’s backbone, implicitly introducing the multiple-to-one mapping to STRATEGY A. This characteristic complicates the independent analysis of each mapping strategy with respect to representation ability, which is unsuitable for our objectives. Furthermore, most knowledge distillation approaches [14, 44, 53] treat the backbone and classifier as a unified entity during training. This coupling potentially obscures the backbone’s feature extraction capabilities, as they may be inevitably affected by the classifier’s performance. In summary, evaluating the impact of mapping strategies on representation ability of backbone remains challenging.

2.2.2 Separation Loss Function Design

To address the challenges previously outlined, we propose a novel loss function designed to evaluate the influence of various strategies on representation ability. Specifically, we employ the Kullback-Leibler (KL) Divergence to leverage the information embedded in soft targets during the training of the backbone. For the classifier, we decouple it from the entire model and train it independently using cross-entropy (CE) loss. This approach prevents the transmission of one-hot target information from the classifier to the backbone, thus maintaining the separation between these mapping strategies. Furthermore, this decoupled training strategy ensures that the classifier performance does not adversely affect the representational capacity of backbone.

Backbone Training. For backbone training, denoted as ϕ_θ , we employ the Kullback-Leibler (KL) divergence to align its predictions with the soft targets generated by the teacher models, thus effectively incorporating the STRATEGY B and STRATEGY C schemes.

To decouple the effect of the classifier, we introduce a novel softmax layer \mathbf{g} , projecting the outputs of ϕ_θ into soft targets. This layer replaces the classifier, thereby ensuring the classifier’s isolation from the training process. The loss for backbone training is formulated as follows:

$$\mathcal{L}_B(\phi_\theta, \mathbf{g}) = \sum_{i=1}^N \sum_{j=1}^{n_i} D_{\text{KL}} \left(\psi(\mathbf{x}_i^{(j)}) \parallel \mathbf{g}(\phi_\theta(\mathbf{x}_i^{(j)})) \right), \quad (1)$$

where N is the number of data points, n_i is the number of augmented samples for \mathbf{x}_i , $\psi \in \{\text{STRATEGY A}, \text{STRATEGY B}, \text{STRATEGY C}\}$ represents the mapping strategy, and $\mathbf{x}_i^{(j)}$ denotes the j -th augmented sample of the original sample \mathbf{x}_i .

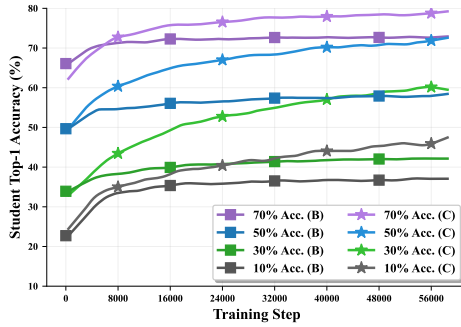


Figure 4: **Comparison of STRATEGY B and STRATEGY C.** **Left:** STRATEGY C demonstrates superior performance in long-term training. Even when using a randomly initialized teacher model with 10% accuracy, the student model trained with STRATEGY C outperforms the student model trained under STRATEGY B with a teacher model of 30% accuracy. **Right:** Results are consistent on different backbones. (*training step: 39k*)

Classifier Training. Following previous works [14, 44], we employ the Cross-Entropy (CE) loss to leverage one-hot targets in training the classifier \mathbf{h} . The loss for classifier training is defined as follows:

$$\mathcal{L}_H(\mathbf{h}) = - \sum_{i=1}^N \sum_{j=1}^{n_i} y_i \log(\mathbf{h}(\phi_{\theta}(\mathbf{x}_i^{(j)}))), \quad (2)$$

where y_i is the one-hot target of the original sample \mathbf{x}_i .

Finally, the total loss function for model training is the sum of the backbone and classifier losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_B(\phi_{\theta}, \mathbf{g}) + \mathcal{L}_H(\mathbf{h}). \quad (3)$$

2.2.3 Loss Function Evaluation

To further compare the differences between our designed loss function and traditional loss functions, we first train a series of teacher models with varying performance (in terms of their validation accuracy). Subsequently, these teacher models are employed to train student models to convergence, as shown in Table 1. Under the training framework of our decoupled loss function, the student models consistently outperform their corresponding teachers. And relatively weaker teacher models can significantly enhance the performance of student models.

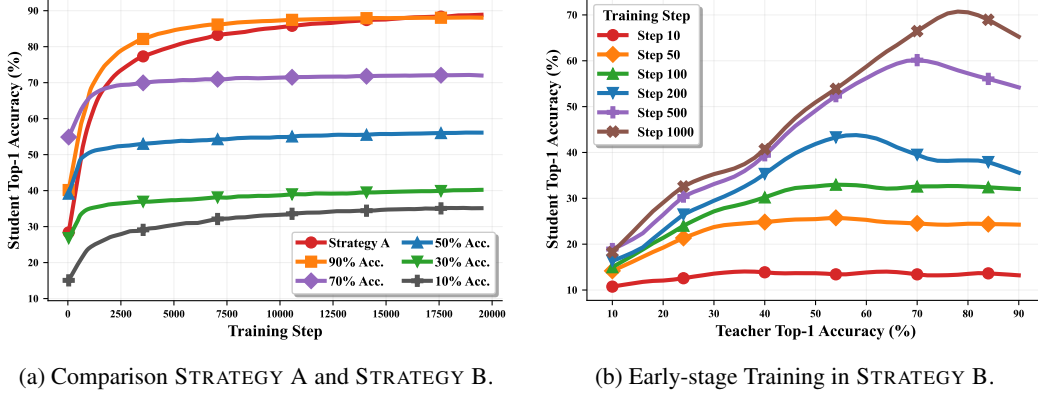
3 What role do targets play? A Comprehensive Investigation of Mapping Strategies

This section systematically investigates the impact of various mapping strategies $\psi : \mathcal{X} \rightarrow \mathcal{Y}$ and different target generation on the representational capacity of the backbone network ϕ_{θ} . For experimental setup, see Appendix B. We begin by comparing and analyzing how different target types, namely one-hot targets in STRATEGY A, soft targets in STRATEGY B, and our proposed STRATEGY C, affect the representational capacity of ϕ_{θ} , as detailed in Section 3.1. Furthermore, we assess the effect of soft targets generated by teachers with different performance, as discussed in Section 3.2. Finally, we explore the impact of soft targets produced by teacher models trained with diverse augmentation strategies, as described in Section 3.3.

3.1 Comparing analysis of Target mapping: What kinds of Targets Drives Better Performance?

STRATEGY B: more targets, faster convergence. We first use a series of teacher models with various accuracies to train the student. The results, as presented in Figure 5a, demonstrate that *models trained with STRATEGY B exhibit accelerated learning during early training stages*. This is due to the additional information provided by soft targets. Conversely, models trained with the traditional STRATEGY A, despite slower improvements in the early-stage training, continue to improve performance over time and ultimately surpass the models trained with STRATEGY B.

STRATEGY C: Fewer Targets, higher performance. Though both STRATEGY B and STRATEGY C using soft labels generated by a pre-trained teacher model. The key distinction lies in the number



(a) Comparison STRATEGY A and STRATEGY B. (b) Early-stage Training in STRATEGY B.
Figure 5: STRATEGY B accelerates early-stage training but achieves inferior final performance. Left: Comparison of STRATEGY A and STRATEGY B that employs teacher models of differing accuracies. **Right:** Weaker teacher models can facilitate early-stage learning. (dataset: CIFAR-10, backbone: ResNet-18.)

of targets used during model training: STRATEGY B assigns different soft targets for each augmented sample, whereas STRATEGY C utilizes the same soft target for all augmented samples derived from the same original sample. We conduct extensive experiments to assess these strategies with different teachers on different datasets and backbones. As shown in [Figure 3](#), [Figure 4](#) and numerical results in [Table 7](#), [Table 8](#), The results show that though in the early training phase, STRATEGY B exhibits an advantage when the teacher model’s accuracy is relatively high. As training progresses, the student model trained using STRATEGY C generally achieves higher gains. And compared to traditional STRATEGY A, STRATEGY C can also retain the ability to accelerate early stage training. For CIFAR10, We further conduct experiments employing a teacher model with different accuracy and compare the performance of the student model at different training stages under these two strategies, as illustrated in [Figure 9](#) and [Table 3](#). We find that as the teacher model’s accuracy increases from 10% to 90%, the improvement in student model performance initially rises and then declines. Notably, with a teacher model accuracy of 30%, the performance gain (a factor of 1.35 increase) achieved with STRATEGY C is substantially greater than that with a 90% accuracy teacher model (a factor of 1.01 increase). These results suggest that the benefit of STRATEGY C diminishes when the teacher model’s accuracy is sufficiently high.

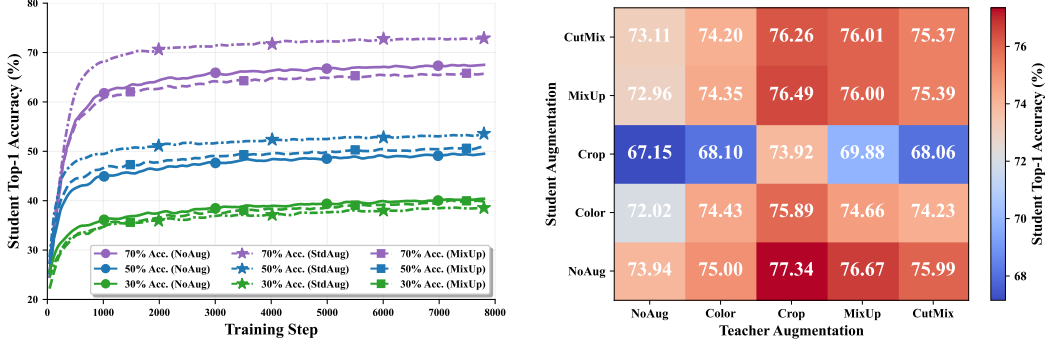
Claim 1 . STRATEGY B accelerates the early-stages training but ultimately limits accuracy in relation to teacher accuracy. Compared to STRATEGY B, STRATEGY C consistently achieves higher final converged accuracy while maintaining short-stage acceleration.

3.2 Early-Stage Learning: Stronger Teachers Are Not Always Optimal.

Previous experiments have demonstrated that the performance of student models is typically influenced by the targets generated by different teacher models, with strong teachers generally improving the final accuracy of the student model. However, it is crucial to note that *stronger teacher models are not universally advantageous*. In this experiment, we focus on the early stages of student training. We fix the number of training steps³ to examine whether stronger teachers lead to greater improvements. As illustrated in [Figure 5b](#), for a fixed number of training steps, as the teacher model’s accuracy increases, there is an initial improvement followed by a decline in the performance of student models. This suggests that strong teacher models do not consistently provide significant advantages and may even hinder student model training during early-stage learning.

Claim 2 . While stronger teacher models generally enhance student model performance over the long term, weaker teacher models can facilitate early-stage learning, particularly when the number of training steps is limited.

³We also experimented with more training steps, see [Table 2](#) in [Appendix C](#) for details.



(a) Data Augmentations for Teacher Model

(b) Data Augmentations for Teacher and Student Model

Figure 6: **Left: Teacher models trained with various augmentation strategies.** Teacher models with high accuracy, trained using MixUp, can not help student model training. **Right: Applying various augmentations to both teacher and student models.** RandomResizedCrop significantly reduces student model performance.

3.3 Impact of Data Augmentation on Soft Target Generation

Prior sections have concentrated on the influence of teacher model performance on student model training. In this section, we focus on the intrinsic properties of teacher models. A notable study by [50] highlights that advanced augmentation techniques like MixUp can enhance a model’s generalization and robustness by creating new training samples through linear interpolation between pairs of examples and their corresponding labels. Therefore, this section explores the influence of data augmentation techniques employed during teacher model training.

Teacher Models with Various Data Augmentations. In this experiment, we use teacher models with three identical performance levels (30%, 50%, and 70%) but trained with different augmentation strategies: (i) **NoAug**: no augmentation, (ii) **StdAug**: standard augmentation without MixUp, and (iii) **MixUp**: MixUp augmentation. Detailed descriptions of these strategies are provided in [Appendix B](#). Additional augmentation strategies applied to the teacher model are discussed in [Table 4](#) in [Appendix C](#).

The results, as depicted in [Figure 6a](#), demonstrate that *for high-accuracy teacher models, applying MixUp augmentation to the teacher does not benefit the training of the student model*. Interestingly, when teacher model accuracy is relatively low (e.g., 30% accuracy, as shown by the green line), the advantage of standard and MixUp strategies diminishes, and models trained with no augmentation show the best performance.

Both student and Teacher Models with Various Data Augmentations. To further assess the impact of augmentation techniques applied on student models themselves, we fix the teacher model’s accuracy⁴ at 70% and train the student model using various data augmentations. The data augmentation techniques include five strategies: NoAug, RandomResizedCrop⁵, StdAug, MixUp, and CutMix [47]. Detailed descriptions of these augmentation strategies are provided in [Appendix B](#). The results, shown in [Figure 6b](#), reveal that:

- (a) For the teacher model, employing RandomResizedCrop, MixUp, or CutMix significantly enhances student model accuracy compared to other methods.
- (b) For the student model, using RandomResizedCrop significantly reduces performance unless the teacher model was also trained with RandomResizedCrop.

Claim 3 . (1) Teacher models trained with augmentation strategies such as MixUp can not consistently enhance student model performance, and no augmentation for the weaker teacher better facilitates student training.. (2) Additionally, employing RandomResizedCrop for student model training significantly impairs performance unless consistently applied to both teacher and student model training.

⁴Results for teacher models with different accuracy levels are available in [Figure 10](#) in [Appendix C](#).

⁵RandomResizedCrop is notably regarded as a strong augmentation method, alongside mix-based strategies.

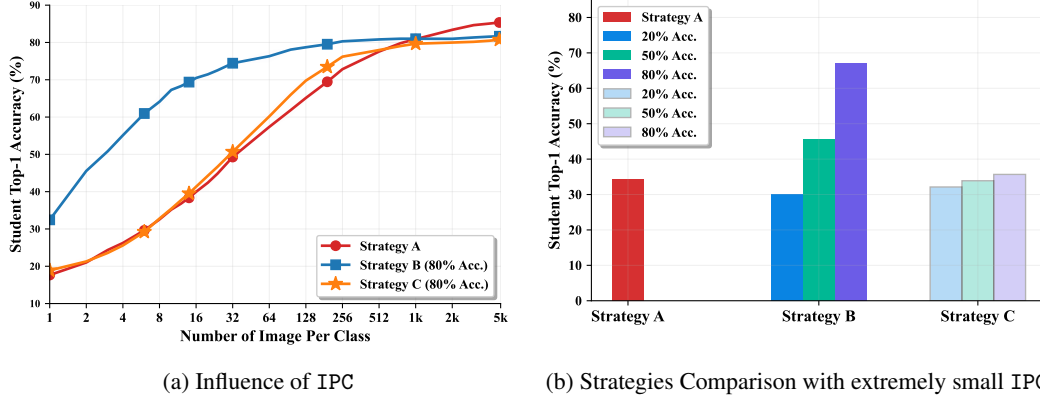


Figure 7: **Left: Student model accuracy scaling with increasing Images Per Class (IPC).** With limited IPC, STRATEGY B deviates from the typical power-law scaling compared with other strategies, whereas STRATEGY A exhibits notable advantages when IPC is sufficient. **Right: Strategy comparison when IPC = 10.** STRATEGY B demonstrates significant benefits, while STRATEGY C shows limited advantages.

4 How Does Sample Variation Affect Performance?

In preceding sections, we focus on how variations in target mappings impact the model training process. Notably, equally crucial to the performance of deep learning models is the quantity and quality of training samples. In this section, we first investigate how sample quantity affects model performance, followed by an analysis of sample quality.⁶

4.1 Less is More? Probing Data Quantity through IPC Variations

Influence of IPC on Three Mapping Strategies. We begin by investigating how sample quantity influences model performance across three strategies using Images Per Class (IPC) as a metric. For each IPC configuration, student models are trained using these three strategies for 10,000 steps, a point at which models nearly achieve convergence. The results are depicted in Figure 7a. Key observations include:

- (a) Both STRATEGY A and STRATEGY C exhibit typical *power-law scaling*. As the number of images per class increases, the rate of improvement significantly diminishes. Notably, STRATEGY C demonstrates a slight advantage over STRATEGY A when the sample size is limited, such as when IPC is below 500.
- (b) For STRATEGY B, *limited IPC results in substantial performance gains*, achieving approximately 30% higher margins than STRATEGY A and STRATEGY C. However, as IPC increases, the benefits diminish. For example, when IPC surpasses 1K, the student model’s accuracy approaches that of the teacher model and is eventually surpassed by STRATEGY A.

Extremely small IPC Across Three Strategies. We further conduct experiments with the three strategies under extremely limited samples (e.g., IPC = 10). For STRATEGY B and STRATEGY C, we employ teacher models with three different performance levels (20%, 50%, and 80%), all trained with identical augmentation. As illustrated in Figure 7b, the results indicate that (1) with STRATEGY B, the student model nearly achieves the teacher model’s accuracy, and teacher models with higher accuracy provide a greater advantage. (2) Conversely, with STRATEGY C, the student model performance exhibits negligible differences despite the varying accuracies of the teacher models.

Claim 4 . (1) Student models trained with STRATEGY B in low IPC achieve significantly higher accuracy compared to STRATEGY A and STRATEGY C. Moreover, teacher models with higher accuracy provide a greater advantage in STRATEGY B. (2) Yet, as sample size increases, the traditional STRATEGY A demonstrates superior performance, as both STRATEGY B and STRATEGY C become limited by the teacher model’s accuracy.

⁶The accuracy of the teacher models in Section 4.1 and Section 4.2 is 80% and 90%, respectively. Additional experiments with different teacher model accuracies are detailed in Figure 11 and Figure 12 in Appendix C.

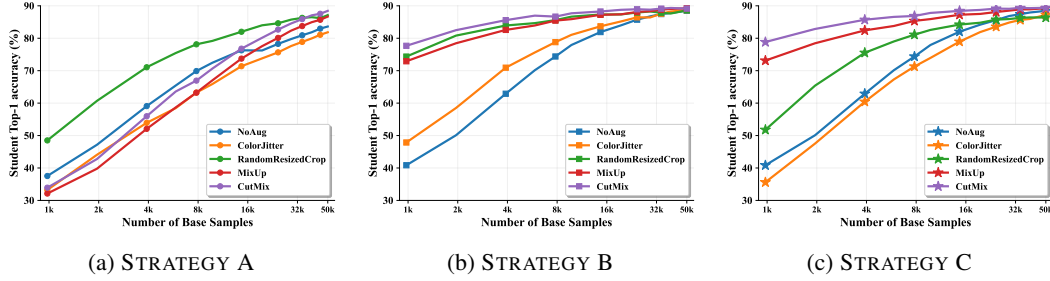


Figure 8: **Scaling behavior of models using various data augmentation methods across three strategies.** For STRATEGY A, RandomResizedCrop enhances model performance significantly. In STRATEGY B and STRATEGY C, augmentation strategies surpass scaling law limitations when samples are limited. However, with sufficient samples, models converge to similar accuracy as the teacher model under both strategies, regardless of the augmentation used.

4.2 Strategic Enhancements: Data Augmentation for Superior Training

Beyond sample quantity, sample quality substantially impacts model performance. Data augmentation techniques [32, 38], such as random cropping, rotation, flipping, and color jittering, introduce controlled transformations that increase the diversity of training data. [13] investigates the relationship between data augmentations and scaling laws and quantify the value of augmentations. And in this section, we evaluate the impact of various data augmentation strategies. For each strategy, we vary the number of samples per class (IPC) and apply different data augmentation methods to assess their influence on model training. The results, depicted in Figure 8, are as follows:

- STRATEGY A (Figure 8a): The influence of augmentation strategies varies with sample size. RandomResizedCrop significantly outperforms other augmentation strategies (ColorJitter, MixUp, CutMix) regardless of sample size. Notably, when the sample size is limited, no augmentation except RandomResizedCrop proves beneficial. However, when the sample size is sufficient, e.g., IPC exceeds 16K, CutMix and MixUp provide slight advantages over no augmentation.
- STRATEGY B (Figure 8b): All data augmentation strategies enhance model performance regardless of sample size. In particular, MixUp, CutMix, and RandomResizedCrop provide significant improvements with limited samples. Interestingly, with abundant samples, augmentation strategies converge to similar performance levels as that of the teacher model.
- STRATEGY C (Figure 8c): No Augmentation, ColorJitter, and RandomResizedCrop exhibit a performance pattern akin to STRATEGY A, while MixUp and CutMix show scaling behavior similar to STRATEGY B.

Claim 5 . (1) For STRATEGY A, RandomResizedCrop is the most effective augmentation method, particularly with limited sample sizes, whereas other strategies may not be beneficial. (2) Both STRATEGY B and STRATEGY C can surpass traditional STRATEGY A scaling law limitations through augmentation when data is scarce. With sufficient data, models utilizing STRATEGY B converge to similar performance levels regardless of the augmentation method employed.

5 Conclusion and Future Work

Conclusion. To the best of our knowledge, this study is the first to emphasize the critical role of samples, targets, and their mappings in efficient learning. We present three sample-to-target mapping strategies and propose a unified framework for evaluation. Our analysis systematically investigates the impact of target and sample types, along with their quantities and qualities on model training. Based on this analysis, we provide six key insights for improving training efficiency.

Future Work. The experiments presented in this paper are conducted *within the context of computer vision*. In future, we aim to perform a more comprehensive analysis of these phenomena *theoretically* and *experiment on other modality*, such as text data. By examining results across diverse datasets, networks and advanced algorithms, we will assess the robustness and generalizability of our findings.

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the International Conference on Computer Vision (ICCV), 2021.
- [3] Mingxing Chen, Vijay Badrinarayanan, Chung-Ching Lee, and Andrew Rabinovich. Semi-supervised knowledge distillation for model compression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.
- [5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9640–9649, 2021.
- [6] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282, 2017.
- [7] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368):829–836, 1979.
- [8] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. Advances in Neural Information Processing Systems, 35:810–822, 2022.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 248–255. IEEE, 2009.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations, 2021.
- [12] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7382–7392, 2024.
- [13] Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. arXiv preprint arXiv:2210.06441, 2022.
- [14] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. International Journal of Computer Vision, 129(6):1789–1819, 2021.
- [15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33:21271–21284, 2020.
- [16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

- [17] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In International Conference on Learning Representations, 2024.
- [18] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. Clip and complementary methods. Nature Reviews Methods Primers, 1(1):1–23, 2021.
- [19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778. IEEE, 2016.
- [21] Zeyi He, Huaxiong Li, Xin Liu, and Steven C Hoi. Knowledge adaptation: Teaching to adapt. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 384–393, 2020.
- [22] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409, 2017.
- [23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop, 2015.
- [24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.
- [25] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.
- [26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in neural information processing systems, 33:18661–18673, 2020.
- [27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [28] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. CS 231N, 7(7):3, 2015.
- [29] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology (TIST), 13(2):1–41, 2022.
- [30] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7):3523–3542, 2021.
- [31] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. Advances in neural information processing systems, 34:20596–20607, 2021.
- [32] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

- [34] Debasis Ray and Jakob Hanni. Data efficient learning of molecular slow modes from nonequilibrium metadynamics. ChemRxiv, 2025.
- [35] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, S Yu Philip, and Lifang He. Deep clustering: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems, 2024.
- [36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018.
- [37] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.
- [38] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. Journal of Big Data, 6(1):60, 2019.
- [39] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. Advances in Neural Information Processing Systems, 35:19523–19536, 2022.
- [40] Peng Sun, Yi Jiang, and Tao Lin. Efficiency for free: Ideal data are transportable representations. arXiv preprint arXiv:2405.14669, 2024.
- [41] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
- [42] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019.
- [43] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12196–12205, 2022.
- [44] Runjian Xu, Jiaming Zhang, Xiaoyang Hu, and Dahua Lin. Knowledge distillation meets self-supervision. arXiv preprint arXiv:2006.07114, 2020.
- [45] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo 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, pages 4133–4141, 2017.
- [46] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. arXiv preprint arXiv:2306.13092, 2023.
- [47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 6023–6032. IEEE, 2019.
- [48] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In International conference on machine learning, pages 12310–12320. PMLR, 2021.
- [49] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. Network representation learning: A survey. IEEE transactions on Big Data, 6(1):3–28, 2018.
- [50] Hongyi Zhang, Yu Zhong, Iwao Sato, and et al. Mixup: Beyond empirical risk minimization. In International Conference on Learning Representations, 2018.
- [51] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 6514–6523, 2023.

- [52] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. arXiv preprint arXiv:2006.05929, 2020.
- [53] 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.
- [54] Yuanlong Zhu et al. Data-efficient learning with active label cleaning and dynamic curriculum. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

A Related Work

To assess the impact of data—consisting of samples X and targets Y —on the learning, we revisit three distinct paradigms in deep learning and examine their respective data structures.

Conventional representation learning. Representation learning [49] enables models to automatically discover and extract meaningful features from raw data, facilitating tasks such as classification [29], clustering [35], and segmentation [30]. These tasks typically rely on datasets with explicit sample-target pairs, such as ImageNet-1K [9]. Recent studies [33, 18] have extended this framework by utilizing text-image pairs to train robust representation models. In these approaches, text is implicitly used as a target through contrastive learning [4, 26]. Moreover, self-supervised learning methods [15, 48, 5, 2] have been developed to extract representations from unlabeled data by generating supervision signals through mechanisms like momentum models.

Summary. *Current representation learning techniques construct learning targets Y from samples X , explicitly or implicitly, to guide models in learning the nonlinear mapping between them.*

Soft label learning via knowledge distillation. Knowledge distillation, introduced by [23], is a prominent technique for model compression in which a smaller student model learns from the outputs, or soft labels, of a larger teacher model. This approach effectively transfers knowledge with minimal performance degradation, enabling smaller, efficient models while maintaining capability [14]. Subsequent research has expanded its application to enhance model robustness [44], facilitate domain adaptation [21], and support semi-supervised learning [3]. Moreover, [45] demonstrates that knowledge distillation can accelerate the optimization process during training.

Summary. *Knowledge distillation generates unique **soft targets** Y for each sample via a pre-trained teacher model, embedding rich information that enables the student model to closely mimic the behavior of the teacher.*

Data-efficient learning through condensed data. Traditional dataset distillation aims to replicate the original dataset’s behavior within a distilled dataset by minimizing discrepancies between surrogate neural network models trained on both synthetic and original data. Key metrics in this process include matching gradients [52], features [43], distributions [51], and training trajectories [8]. However, these methods incur substantial computational overhead due to the ongoing calculation of discrepancies, requiring numerous iterations for optimization and convergence. This makes them challenging to scale to large datasets, such as ImageNet [9]. An effective strategy involves identifying metrics that capture critical dataset information [46, 41], thereby eliminating the need for exhaustive comparisons between original and distilled datasets. Consequently, these methods scale efficiently to large datasets, such as ImageNet-1K. For instance, SRe²L [46] compacts the entire dataset into a model, such as pre-trained neural networks like ResNet-18 [20], and subsequently extracts knowledge into images and targets, forming a distilled dataset. Recently, RDED [41] suggests that images accurately recognized by strong observers, such as humans and pre-trained models, are pivotal for effective learning.

Summary. *Dataset distillation methods extract key information from the original dataset, forming more informative samples X and targets Y that facilitate data-efficient learning.*

B Experimental Setup

Dataset and Network. We conduct experiments using both large-scale and small-scale datasets, specifically CIFAR-10 [27], CIFAR-100 [27], Tiny-ImageNet [28] and ImageNet-1k [10]. In the main body, our model utilizes ResNet-18 [20] as the backbone, selected due to its widespread use and its generalizable performance across different architectures [24]. We also evaluate MobileNetV2 [36], EfficientNet [42], ResNet-50 and Vision Transformer [11] as the backbone on the proposed STRATEGY C. The evaluation metric for all experiments is the Top-1 classification accuracy (%) on the test set. In the main body, we primarily present results on CIFAR-10 and ResNet-18, while others are provided in [Appendix D](#).

Batch size and training step settings. All experiments in the main body were conducted using the CIFAR-10 dataset. By default, the number of images per class in the training set is 5000, except for experiments in [Section 4.1](#), [Section 4.2](#). For all experiments throughout the paper, samples were

randomly shuffled at the beginning of each epoch, and the batch size was fixed at 128. The number of batches per epoch after applying drop last strategy is 390 for both CIFAR10 and CIFAR100, and 780 for TinyImageNet. For the experiments in Figure 5 and 14, training was run for 50 epochs (corresponding to 20,000 steps for CIFAR10 and CIFAR100, and 40,000 steps for TinyImageNet), with validation averaged 10 times per epoch. For the experiments in Figure 6a and 16, training was run for 20 epochs (7,800 steps for CIFAR10 and CIFAR100, and 15,600 steps for TinyImageNet), also with validation averaged 10 times per epoch. For the experiments in Figure 9, training was run for 150 epochs (i.e., 58,500 steps), with validation performed once at the end of each epoch (i.e., every 390 steps). And for experiments in Section 4.1, Section 4.2 and Figure 6b, 10, 11, 12, 17, 18, the total number of training steps is set to fixed 10,000 steps.

Model architecture. For the experiments on CIFAR10, CIFAR100, and TinyImageNet, both the teacher and student models were implemented using a modified ResNet-18 architecture (no pretrained weights) where the initial convolutional layer was replaced with a 3×3 convolutional layer, using stride 1 and padding 1 to better accommodate the input image sizes. The batch normalization layer was adjusted accordingly to match the new convolutional output, and the max-pooling layer was removed. To better evaluate the backbone of the model, we pruned the final fully connected layers, and a linear layer was attached from the output feature dimension to the number of classes in the corresponding dataset, the backbone of the model and the attached classifier layer were trained separately, as described earlier in Section 2.

Pre-trained Teacher Models. To thoroughly investigate the influence of teacher models with varying accuracies on the representation capability of the student model ϕ_θ , we first train a series of teacher models with differing accuracies, employing various augmentations on the training set. using the traditional STRATEGY A strategy (i.e., using one-hot targets) and trained them with the standard cross-entropy loss function. Validation was performed at the end of each step, and the teacher model was saved the first time it reached the preset accuracy, which was then used for subsequent student model training, and we ensure that the validation error for each teacher model with a specific accuracy is less than 1%. For STRATEGY B and STRATEGY C, the student model is trained using the soft targets by the teacher model, with the temperature of the soft labels set to 2.0.

Plot settings. To reduce noise and provide clearer visualization of long-term trends, we applied locally weighted scatterplot smoothing (LOWESS) [7] to the figures that examine long-term behavior, such as Figures 5, 4, 9, 14, 15, and 16, using 10% of the data points for each local regression. Considering that the curve fitting applied is not well-suited for observing early-stage data, particularly when assessing the impact of STRATEGY C in the initial phases, we provide additional numerical results in the early training phase in Table 3, 6, 7, 8. We conduct at least five trials for each experiment and report the standard deviations for the proposed three strategies,.

Hyper parameter settings. We employed the default implementation of the AdamW optimizer in the PyTorch Lightning Module, where the learning rate is set to 1×10^{-3} , and the weight decay is 0.01. The metric for evaluating the performance of all models is their Top-1 validation accuracy on the test set. Experiments were conducted on NVIDIA RTX 4090 GPUs and Intel Xeon processors.

Augmentation definition. We define the data augmentation techniques mentioned throughout the paper as follows:

No Augmentation (NoAug): Only normalization is applied based on the corresponding dataset.

Standard Augmentation (StdAug):

- **ColorJitter (or Color in short):** Apply random changes to the brightness, contrast, saturation with a probability of 0.8 to each image.
- **RandomGrayscale:** Converts the image to grayscale with a 0.2 probability.
- **RandomHorizontalFlip:** Flips the image horizontally with a 50% chance.
- **RandomResizedCrop (or Crop in short):** Resizes and crops the image to the input size with a scaling factor between 0.08 and 1.0.

Mix-based Augmentation (not include StdAug in our settings):

- **MixUp**: Perform a linear interpolation between two randomly selected images, where the interpolation factor, λ , is drawn from a Beta distribution with parameters $(\alpha, \beta) = (0.8, 0.8)$ which controls the degree of interpolation between the two images.
- **CutMix**: A bounding box is randomly generated within one image, and the pixel values inside the box are replaced by the corresponding region from a second randomly selected image. The area of the bounding box is determined by a parameter sampled from a Beta distribution with parameters $(\alpha, \beta) = (1.0, 1.0)$ which controls the relative size of the cutout region and the pasted-in region.

For all experiments except in [Section 3.3](#), [Section 4.2](#) and [Table 4, 9](#), [Figure 10, 16, 18](#), we employed the **Standard Augmentation (StdAug)** and normalization for both teacher models and student models. For the experiments in [Figure 6a, 16](#), and [Table 9](#), we applied **StdAug** to the student model. For the experiments in [Figure 8](#) and [18](#), **StdAug** was applied to the teacher model.

C Additional Experimental Results

We further conducted more experiments on the CIFAR10 dataset to supplement and verify the patterns observed in the main body. We first show the experimental patterns can be observed more clearly under the step setting in [Section 3.1](#) through [Table 2](#). And we compare STRATEGY B and STRATEGY C using more teachers during different training stages to observe under what conditions can STRATEGY C hold a greater advantage through [Table 3](#), and we find that STRATEGY C will not yield a significant advantage when the accuracy of the teacher model is high enough (see [Figure 9](#)).

In addition to the data augmentation strategies used in [Section 3.3](#), we further compared the effects of applying more data augmentation strategies to teacher models with different accuracies in [Table 4](#). We find that in most cases only when applying the RandomResizedCrop strategy to the teacher model does it significantly help the training of the student model, while most other individual augmentation strategies tend to have a negative impact.

Aside from the investigation of samples in [Section 4](#), we further conducted more experiments and found that the scaling behavior of student models trained by teacher models with different accuracies shows slight differences, as shown in [Figure 11](#) and [12](#).

In addition to the experiments above, we performed ablation studies on the effects of batch size and learning rate, as shown in [Figure 13](#) and [Table 5](#).

All values presented in the tables represent the Top-1 accuracy of the student model under the corresponding teacher model settings with the percentage symbol (%) omitted for simplicity.

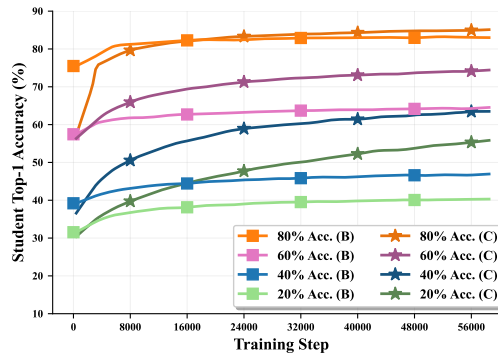


Figure 9: **Comparison of STRATEGY B and STRATEGY C under more different teacher models with different accuracies.** It further shows that as the accuracy of the teacher model increases, the advantage of using STRATEGY C over STRATEGY B becomes less apparent.

D Additional Dataset

We further conducted experiments on different datasets, including CIFAR100 and TinyImageNet to validate the findings presented in the main body. We first compared STRATEGY A and STRATEGY

Table 2: **The Top-1 accuracy of the student model under different training steps and different teachers.** The **bolded** data represent the maximum student Top-1 accuracy in each training step. Only under a limited number of training steps (i.e., 1000 steps setting in the main body), the teacher model with relatively medium accuracy will show its advantage.

Training Step	Teacher Top-1 Accuracy (%) $\pm 1\%$								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
10	10.59	12.92	13.39	14.08	13.86	15.53	14.25	15.14	14.38
20	11.85	18.57	18.79	18.47	19.51	19.02	19.11	17.94	19.37
50	13.46	20.94	23.54	24.97	27.69	26.64	27.46	26.42	25.52
100	14.76	23.19	26.71	31.15	35.24	32.55	32.32	33.00	32.22
150	15.63	25.59	27.95	33.38	37.77	37.81	36.57	36.52	34.66
200	18.52	26.23	31.11	36.10	43.29	43.31	41.32	39.44	37.93
300	19.16	27.66	32.00	36.47	46.73	51.51	50.69	46.67	42.99
500	19.49	29.26	33.43	38.06	48.93	55.72	61.06	57.68	54.79
700	21.99	29.84	33.59	38.73	48.83	56.50	63.71	65.16	62.27
1000	24.38	31.17	33.57	39.32	49.63	58.38	65.91	70.13	67.90
1200	24.50	32.35	34.98	40.64	51.56	58.81	67.45	73.09	70.23
1500	25.88	32.30	35.31	39.92	51.76	58.79	68.00	75.29	73.92

Table 3: **Comparing the student models trained with STRATEGY B and STRATEGY C at different training stages.** The Gain represents the performance improvement factor of the student model using STRATEGY C compared to STRATEGY B, with **bold** represent cases where the gain is greater than 1 (i.e., STRATEGY C is more advantageous). For a high-accuracy teacher model, STRATEGY C does not show advantages with fewer training steps. However, for a low-accuracy teacher model, training advantages are observed in both early and late training stages. And as the accuracy of the teacher model increases from 10% to 90%, the gain of using STRATEGY C compared to STRATEGY B first increases then gradually decreases.

Step	Strategy	Teacher Top-1 Accuracy (%) $\pm 1\%$								
		10%	20%	30%	40%	50%	60%	70%	80%	90%
1950	B	25.09 ± 0.5	33.31 ± 0.6	35.93 ± 0.4	40.29 ± 0.6	51.63 ± 0.3	59.12 ± 0.6	66.93 ± 0.4	73.31 ± 0.4	72.80 ± 0.6
	C	27.06 ± 0.7	32.54 ± 0.9	34.79 ± 0.7	41.28 ± 0.5	52.95 ± 0.4	57.55 ± 0.7	63.14 ± 0.6	67.99 ± 0.5	68.66 ± 0.5
	Gain	1.08	0.98	0.97	1.02	1.03	0.97	0.94	0.93	0.94
7800	B	32.02 ± 0.4	36.66 ± 0.5	38.35 ± 0.5	43.19 ± 0.4	54.44 ± 0.4	61.76 ± 0.7	71.24 ± 0.6	81.27 ± 0.6	87.33 ± 0.4
	C	34.07 ± 0.5	39.86 ± 0.8	42.96 ± 0.5	50.57 ± 0.6	60.68 ± 0.5	65.89 ± 0.6	72.47 ± 0.5	79.67 ± 0.7	84.00 ± 0.6
	Gain	1.06	1.09	1.12	1.17	1.12	1.07	1.02	0.98	0.96
39000	B	35.96 ± 0.4	39.70 ± 0.4	41.61 ± 0.5	46.11 ± 0.5	57.18 ± 0.3	63.90 ± 0.4	72.58 ± 0.5	82.93 ± 0.6	90.35 ± 0.4
	C	42.86 ± 0.5	51.98 ± 0.2	55.96 ± 0.4	61.34 ± 0.6	69.84 ± 0.4	73.02 ± 0.5	77.93 ± 0.6	84.29 ± 0.5	91.17 ± 0.5
	Gain	1.19	1.31	1.35	1.33	1.22	1.14	1.07	1.02	1.01

B on the two datasets in Figure 14 and 15, and further investigated the behavior of STRATEGY C through Table 6. We also compared the impact of applying different image augmentation strategies to the teacher models in Figure 16 and Table 9, and further explored the scaling behavior influenced by changing the sample size and applying different augmentation strategies on the student models, as shown in Figure 17 and 18. To have a more comprehensive evaluation on the differences between our proposed strategy STRATEGY C and traditional methods, we conducted experiments on the ImageNet dataset, which is larger and more representative of real-world scenarios. We selected ResNet50 and ViT as the backbone architectures for teacher and student models, and batch size was set to 512, and we validated every 500 steps, with other hyperparameters remain unchanged. The visual results are shown in Figure 3, while detailed numerical results are provided in Table 7, 8.

Table 4: **Comparison of different augmentation method applied for the teacher model using STRATEGY B (training step=7800).** The **bolded** data represent the maximum student Top-1 accuracy in each column. Besides the advantage of NoAug over MixUp in the teacher model shown in section 3.3, this table further indicates that applying RandomResizedCrop to the teacher model will benefit the training of the student model, regardless of the accuracy of the teacher model.

Augmentation (on teacher)	Teacher Top-1 Accuracy (%) $\pm 1\%$							
	10%	20%	30%	40%	50%	60%	70%	80%
NoAug	31.20	35.55	40.29	43.92	49.50	56.94	67.59	77.43
ColorJitter	32.18	35.74	38.44	41.61	50.06	58.83	65.88	77.37
Crop	31.14	37.43	41.27	42.93	54.26	66.11	75.37	82.28
MixUp	30.30	35.00	39.81	43.79	50.93	61.44	65.79	75.78
CutMix	30.74	35.86	37.78	41.23	48.41	55.72	68.65	73.24
StdAug	31.53	35.65	38.49	49.52	53.57	63.34	73.00	80.59

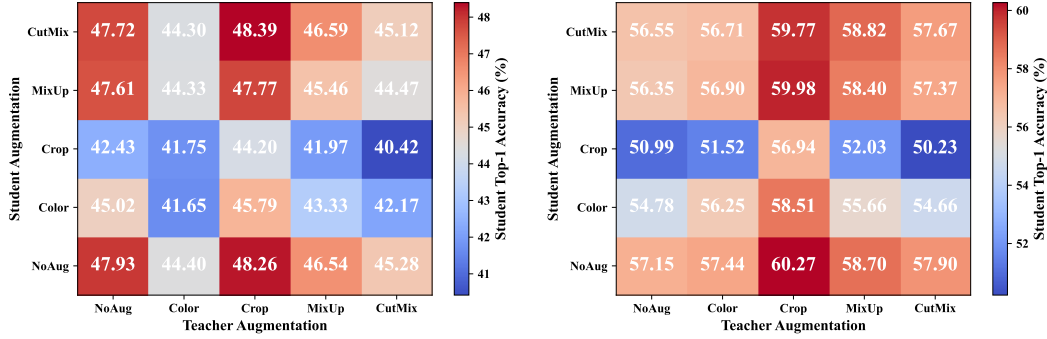


Figure 10: **Ablation experiments of training the student model with 30% (Left) and 50% (Right) teacher models in the main body.** It is evident that using the RandomResizedCrop strategy has certain disadvantages for the student model. Additionally, applying the MixUp strategy to teacher models with lower accuracy undermines the effectiveness of training the student model compared to NoAug.

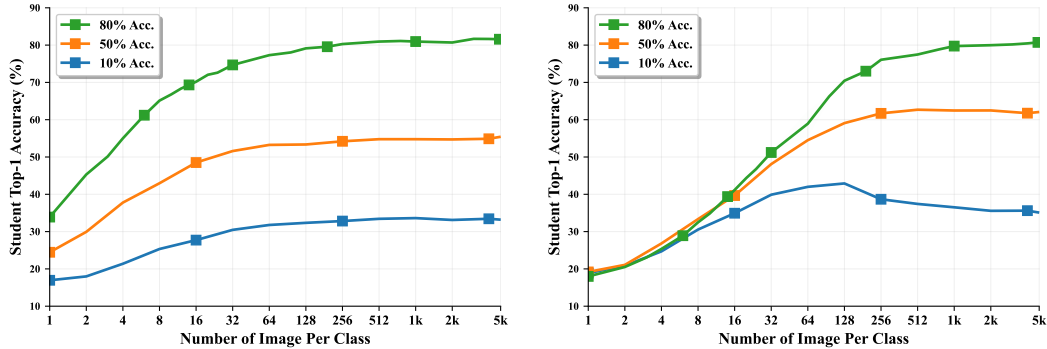


Figure 11: **The scaling behavior across different teachers (by changing the number of image per class) under STRATEGY B (left) and STRATEGY C (right).** In STRATEGY B, different teacher models exhibit similar patterns of scaling, whereas in STRATEGY C, the performance of the student model trained by a randomly initialized teacher model (i.e., 10% Acc.) initially improves but then declines as the IPC increases.

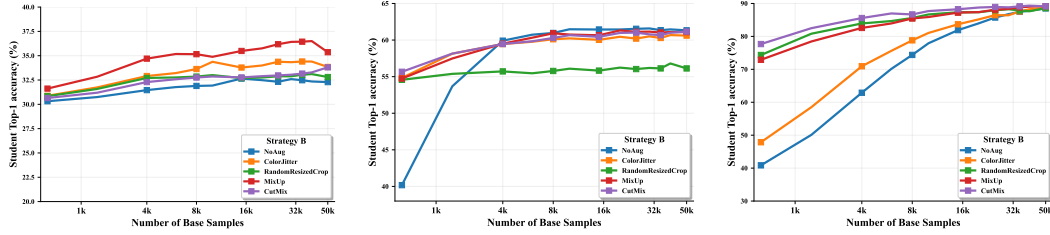


Figure 12: **The scaling behavior across different teacher using different data augmentation under STRATEGY B.** From left to right, the teacher models have accuracies of 10%, 50%, and 90%, respectively. For a teacher model with moderate or high accuracy, applying different data augmentation strategies to the student model results in their performance converging to nearly the same value (except under the 50% teacher model where using the RandomResizedCrop strategy for the student model is insensitive to changes in the number of samples). However, for a low-accuracy teacher model (in this case, a completely random initialized teacher model), applying different data augmentation strategies to the student model shows slight variations in effect as the number of samples increases.

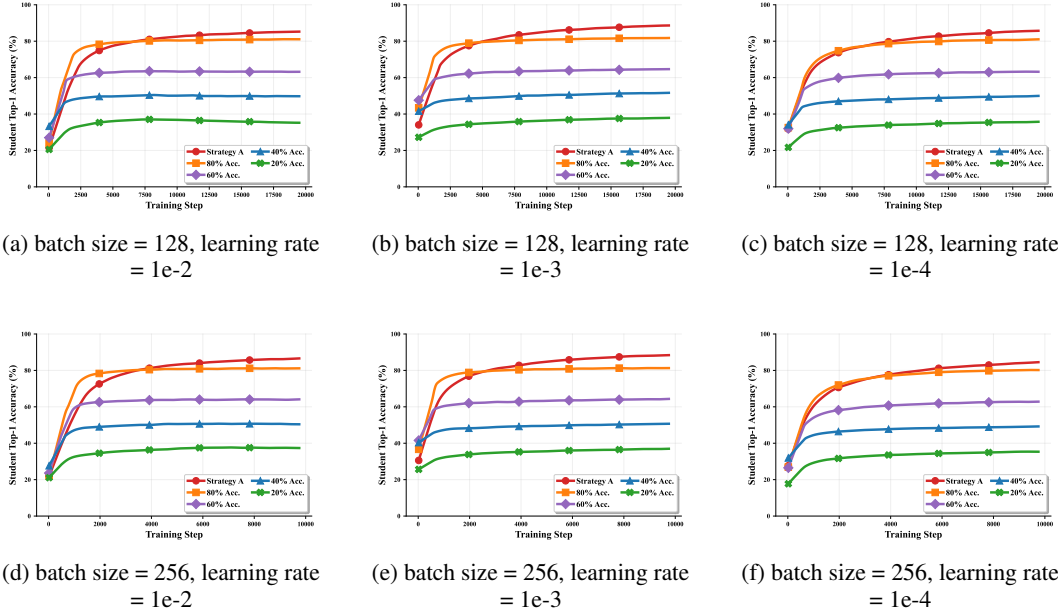


Figure 13: **The comparison of strategies STRATEGY A and STRATEGY B under different learning rate and batch size settings.** All results show that the STRATEGY B has a short term advantage while STRATEGY A have a long term advantage.

Table 5: **The results of student models trained with different batch sizes and learning rates under STRATEGY A and STRATEGY B, evaluated across varying levels of teacher model accuracy.** For each configuration, we report the mean Top-1 accuracy of the student model. The results are consistent for each setting.

batch size	learning rate	Teacher Top-1 Accuracy (%) $\pm 1\%$				
		20%	40%	60%	80%	NoTeach
128	1e-2	35.42 ± 0.5	49.54 ± 0.9	63.23 ± 0.5	81.18 ± 0.4	85.52 ± 0.5
	1e-3	37.86 ± 0.3	51.44 ± 0.3	64.54 ± 0.8	81.68 ± 0.3	88.59 ± 0.5
	1e-4	34.88 ± 0.5	49.63 ± 0.5	63.20 ± 0.7	81.09 ± 0.3	86.01 ± 0.4
256	1e-2	37.39 ± 0.3	50.70 ± 0.4	64.42 ± 0.6	81.78 ± 0.4	86.50 ± 0.7
	1e-3	37.06 ± 0.6	50.35 ± 0.4	64.29 ± 0.5	81.69 ± 0.4	88.53 ± 0.2
	1e-4	34.88 ± 0.5	48.66 ± 0.4	63.15 ± 0.5	76.73 ± 8.2	84.42 ± 0.6

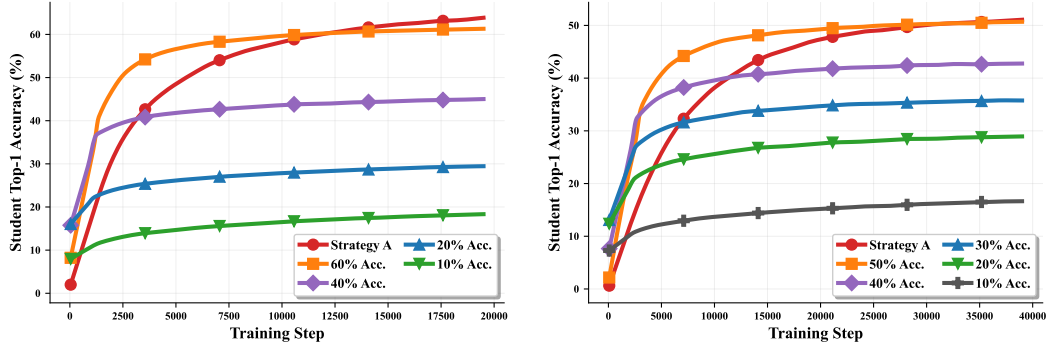


Figure 14: Comparison of STRATEGY A and STRATEGY B using different teachers on CIFAR100 (Left) and TinyImageNet (Right). Both show the long-term advantages of STRATEGY A, and the weaker teacher models exhibit more significant improvements in the performance of the student models.

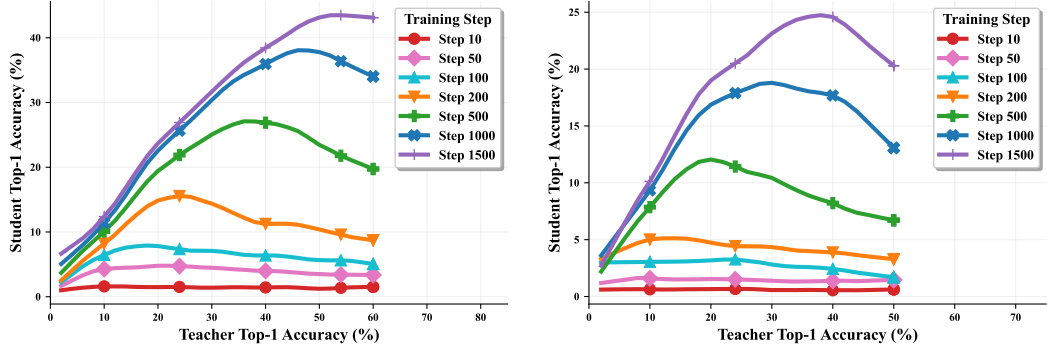


Figure 15: The different behavior across teachers in different early-stage training steps on CIFAR100 (Left) and TinyImageNet (Right). For a teacher model with moderate accuracy, there is an advantage in the early stages of training the student model.

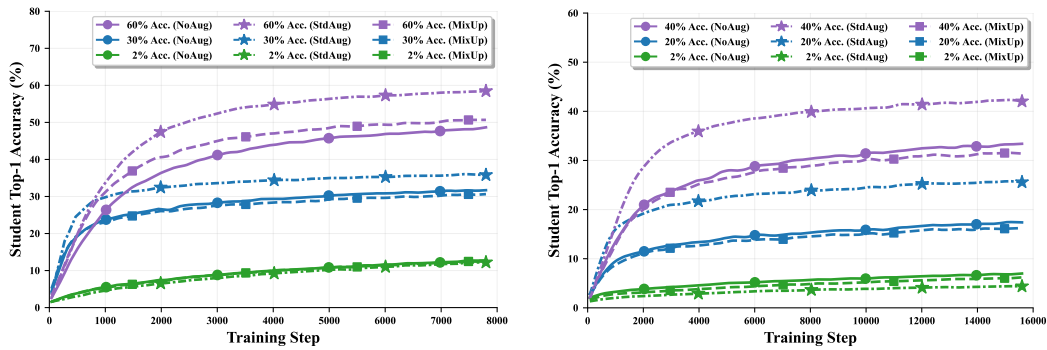


Figure 16: The different impacts on student model training when applying different image augmentation strategies to the teacher model under STRATEGY B, with the left figure representing CIFAR100 and the right figure representing TinyImageNet. As the accuracy of the teacher model increases, aggressive data augmentation strategies like MixUp show a stronger advantage. However, when the accuracy of the teacher model is low, MixUp actually exhibits a more pronounced disadvantage.

Table 6: Comparing the student models trained with STRATEGY B and STRATEGY C on different datasets. Both show that under STRATEGY C, teacher models with moderate or low accuracy are more beneficial for improving the performance of student models.

Dataset	Step	Strategy	Teacher Top-1 Accuracy (%) $\pm 1\%$						
			2%	10%	20%	30%	40%	50%	60%
CIFAR100	1950	B	7.20 ± 0.7	12.39 ± 0.7	23.51 ± 0.6	31.61 ± 0.5	36.79 ± 0.5	42.21 ± 0.6	43.82 ± 0.7
		C	5.83 ± 1.1	13.11 ± 0.8	22.88 ± 0.7	27.92 ± 0.3	31.52 ± 0.4	34.47 ± 0.6	35.56 ± 0.5
		Gain	0.81	1.06	0.97	0.88	0.86	0.82	0.81
	7800	B	12.80 ± 0.7	16.13 ± 0.6	27.19 ± 0.8	35.99 ± 0.7	43.11 ± 0.5	51.80 ± 0.3	58.83 ± 0.6
		C	9.99 ± 0.6	19.87 ± 0.7	31.41 ± 0.9	37.23 ± 0.8	42.82 ± 0.6	48.85 ± 0.5	53.25 ± 0.5
		Gain	0.78	1.23	1.16	1.03	0.99	0.94	0.91
	39000	B	17.69 ± 0.4	20.30 ± 0.5	31.03 ± 0.5	38.99 ± 0.6	45.84 ± 0.4	55.04 ± 0.3	61.74 ± 0.4
		C	18.91 ± 0.7	32.49 ± 0.5	39.75 ± 0.7	44.11 ± 0.4	49.74 ± 0.5	56.53 ± 0.5	62.48 ± 0.5
		Gain	1.07	1.60	1.28	1.13	1.08	1.03	1.01
Tiny-ImageNet	3900	B	2.63 ± 0.9	11.82 ± 0.8	22.31 ± 0.7	28.15 ± 0.3	32.48 ± 0.5	34.68 ± 0.4	-
		C	2.68 ± 0.6	11.74 ± 0.9	20.34 ± 0.8	24.33 ± 0.6	26.48 ± 0.5	27.90 ± 0.6	-
		Gain	1.02	0.99	0.91	0.86	0.82	0.81	-
	15600	B	3.92 ± 0.6	14.76 ± 0.3	27.13 ± 0.6	34.21 ± 0.8	41.20 ± 0.5	48.55 ± 0.6	-
		C	4.59 ± 0.7	17.78 ± 0.5	28.90 ± 0.4	34.61 ± 0.7	39.59 ± 0.8	44.59 ± 0.3	-
		Gain	1.17	1.21	1.07	1.01	0.96	0.92	-
	78000	B	5.63 ± 0.4	18.62 ± 0.4	30.07 ± 0.3	36.56 ± 0.5	43.60 ± 0.5	51.38 ± 0.6	-
		C	8.22 ± 0.7	24.64 ± 0.6	35.21 ± 0.6	41.22 ± 0.6	46.23 ± 0.5	51.86 ± 0.4	-
		Gain	1.46	1.32	1.17	1.13	1.06	1.01	-

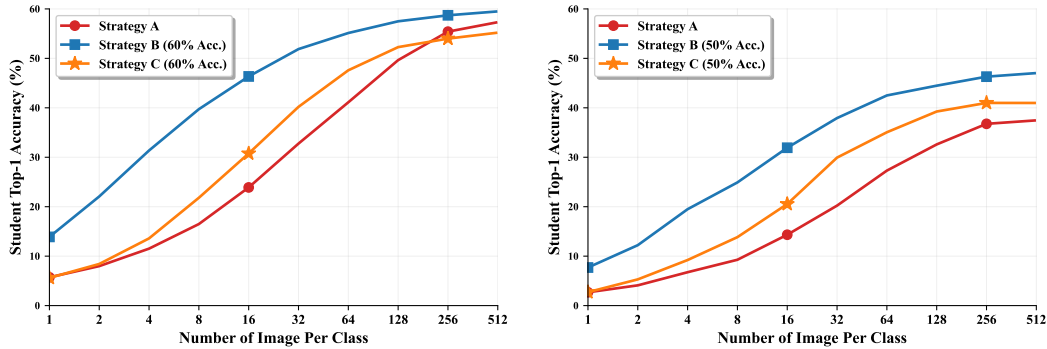


Figure 17: The scaling behavior using different strategy for training on CIFAR100 (Left) and TinyImageNet (Right). In both datasets, STRATEGY B consistently shows a greater advantage, STRATEGY C has some advantage when the sample size is small. And for CIFAR100, when the sample size is sufficient enough (i.e., using all samples where IPC=500), the performance of STRATEGY C is eventually surpassed by STRATEGY A.

Table 7: **Comparing the student models trained with STRATEGY B and STRATEGY C on different backbones at different training stages on ImageNet.** Strategy C consistently achieves higher final accuracy than Strategy B across all teacher models.

Backbone	Step	Strategy	Teacher Top-1 Accuracy (%) $\pm 1\%$		
			12%	28%	50%
RN50	1k	B	11.21 \pm 0.7	14.47 \pm 0.7	7.25 \pm 0.5
		C	6.74 \pm 0.6	7.96 \pm 0.9	5.98 \pm 0.7
		Gain	0.60	0.55	0.83
	50k	B	25.63 \pm 0.5	41.85 \pm 0.6	56.92 \pm 0.5
		C	29.67 \pm 0.6	43.00 \pm 0.7	54.63 \pm 0.3
		Gain	1.16	1.03	0.96
	150k	B	27.30 \pm 0.4	42.39 \pm 0.5	58.60 \pm 0.5
		C	34.75 \pm 0.5	46.24 \pm 0.3	58.04 \pm 0.6
		Gain	1.27	1.09	0.99
	250k	B	27.00 \pm 0.3	42.64 \pm 0.4	58.02 \pm 0.4
		C	36.05 \pm 0.4	47.71 \pm 0.5	58.60 \pm 0.2
		Gain	1.34	1.12	1.01
ViT	1k	B	7.95 \pm 0.7	6.57 \pm 0.6	4.96 \pm 0.5
		C	5.18 \pm 1.0	5.08 \pm 0.5	4.07 \pm 0.4
		Gain	0.65	0.77	0.82
	50k	B	17.20 \pm 0.6	29.93 \pm 0.7	50.05 \pm 0.5
		C	20.83 \pm 0.5	31.94 \pm 0.6	45.37 \pm 0.6
		Gain	1.21	1.07	0.91
	150k	B	17.70 \pm 0.5	30.34 \pm 0.4	50.98 \pm 0.3
		C	24.68 \pm 0.3	35.10 \pm 0.5	51.66 \pm 0.4
		Gain	1.39	1.16	1.01
	250k	B	17.63 \pm 0.4	30.92 \pm 0.5	51.00 \pm 0.5
		C	25.64 \pm 0.6	36.64 \pm 0.4	52.58 \pm 0.3
		Gain	1.45	1.18	1.03

Table 8: Comparing the student models trained with STRATEGY A and STRATEGY C at different training stages on ImageNet using ResNet50 as the backbone. Strategy C significantly accelerates convergence within the first 2.5k steps, regardless of the teacher model used.

Step	Accuracy	Teacher Top-1 Accuracy (%) $\pm 1\%$				
		NoTeach(A)	12%(C)	28%(C)	50%(C)	62%(C)
0.5k	Stu-Acc.	0.87 ± 0.8	3.39 ± 1.0	3.07 ± 0.4	2.06 ± 0.6	2.10 ± 0.5
	Gain	–	3.89	3.53	2.37	2.41
1k	Stu-Acc.	2.92 ± 0.6	6.74 ± 0.4	7.96 ± 0.8	5.98 ± 0.4	5.99 ± 0.3
	Gain	–	2.31	2.73	2.05	2.05
1.5k	Stu-Acc.	6.08 ± 0.4	10.05 ± 0.3	12.34 ± 0.5	10.99 ± 0.4	10.67 ± 0.5
	Gain	–	1.65	2.03	1.81	1.76
2k	Stu-Acc.	9.35 ± 0.6	11.32 ± 0.4	17.09 ± 0.6	14.81 ± 0.6	17.09 ± 0.3
	Gain	–	1.21	1.83	1.58	1.83
2.5k	Stu-Acc.	9.86 ± 0.3	13.42 ± 0.6	18.57 ± 0.4	19.89 ± 0.3	19.94 ± 0.4
	Gain	–	1.36	1.88	2.02	2.02
3k	Stu-Acc.	19.15 ± 0.4	13.50 ± 0.4	22.14 ± 0.3	23.13 ± 0.3	24.59 ± 0.6
	Gain	–	0.70	1.16	1.21	1.28
3.5k	Stu-Acc.	20.73 ± 0.6	14.53 ± 0.3	22.36 ± 0.3	26.37 ± 0.6	27.43 ± 0.5
	Gain	–	0.70	1.08	1.27	1.32
4k	Stu-Acc.	23.93 ± 0.4	14.56 ± 0.5	24.73 ± 0.5	24.80 ± 0.4	28.51 ± 0.4
	Gain	–	0.61	1.03	1.04	1.19
4.5k	Stu-Acc.	27.30 ± 0.5	16.15 ± 0.4	25.70 ± 0.5	32.00 ± 0.5	30.56 ± 0.5
	Gain	–	0.59	0.94	1.17	1.12
5k	Stu-Acc.	30.28 ± 0.4	16.99 ± 0.3	27.13 ± 0.3	32.73 ± 0.2	34.43 ± 0.3
	Gain	–	0.56	0.90	1.08	1.14

Table 9: **Comparison of different augmentation method applied for the teacher model using STRATEGY B on different dataset (7,800 steps for CIFAR100 and 15,600 steps for TinyImageNet).** On both datasets, the RandomResizedCrop strategy and the standard data augmentation strategy have a more pronounced effect compared to using the MixUp strategy. And only when the teacher model trained with these aggressive data augmentation strategies has a high accuracy does it provide a greater advantage for the student model, whereas for a weak teacher model, the advantage is not as apparent.

Dataset	Augmentation (on teacher)	Teacher Top-1 Accuracy (%) $\pm 1\%$					
		2%	10%	20%	30%	40%	50%
CIFAR100	NoAug	12.85	16.89	24.21	31.82	38.13	46.87
	ColorJitter	11.09	16.65	26.75	32.65	39.65	47.64
	Crop	12.68	15.20	25.82	33.80	42.76	50.58
	MixUp	12.64	16.56	25.19	30.78	37.51	45.59
	CutMix	9.62	14.80	22.33	29.34	36.07	43.05
	StdAug	12.34	16.02	25.97	35.95	43.97	51.19
TinyImageNet	NoAug	6.52	10.82	17.55	26.35	33.63	-
	ColorJitter	4.93	9.59	17.70	27.34	34.80	-
	Crop	5.51	14.87	24.31	32.31	38.65	-
	MixUp	6.18	8.99	16.54	25.28	31.62	-
	CutMix	5.62	7.40	16.48	24.07	31.62	-
	StdAug	4.45	17.66	25.79	33.37	42.20	-

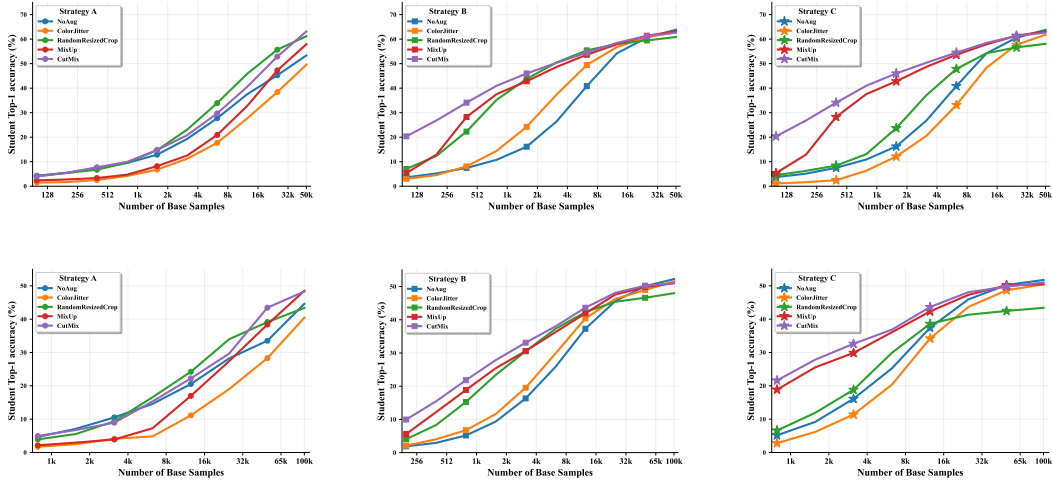


Figure 18: **The scaling behavior of the student model using different image augmentation methods under the three strategies.** The three images above show the results of CIFAR100, while the three images below show the results of TinyImageNet, all of which demonstrate the advantages of STRATEGY B and STRATEGY C compared to the traditional STRATEGY A under limited samples.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We show them in Abstract and Section 1.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We show them in Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We show them in Appendix [B](#).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use the public dataset and provide our code in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We clearly state our experimental setup in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We show them in the tables.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: We provide most of them in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This research fully complies with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper does not provide discussion of potential societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper does not provide the safeguards.

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All code, data, and models used in this paper have been properly attributed to their original creators or owners.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM is used only for editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.