Distill on a Diet: Efficient Knowledge Distillation via Learnable Data Pruning

Yifan Wu[1,2*], Yiqi Wang[4*], Xichen Ye[3*], Wenjing Yan[2], Xiaoqiang Li[4(✉)], Cheng Jin[3], Xiangyu Yue[2], and Weizhong Zhang[1(✉)]

[1] School of Data Science, Fudan University
[2] Faculty of Engineering, The Chinese University of Hong Kong
[3] College of Computer Science and Artificial Intelligence, Fudan University
[4] School of Computer Engineering and Science, Shanghai University
victorwu@link.cuhk.edu.hk {wangyq2004,xqli}@shu.edu.cn
xcye25@m.fudan.edu.cn wenjingyan@cuhk.edu.hk xyyue@ie.cuhk.edu.hk
{jc,weizhongzhang}@fudan.edu.cn
https://github.com/yifanwu-victor/Distill-on-a-Diet

Abstract. Knowledge Distillation (KD) is widely used to obtain compact models for efficient inference in resource-constrained environments. Yet the computational overhead of the distillation process itself is often overlooked, raising the question of whether a better student model can be obtained with less data and less compute via data pruning. However, existing data pruning methods are not designed for KD: some introduce substantial overhead (*e.g.*, obtaining training dynamics through retraining), while others rely on heuristic selection rules that fail to capture what KD actually requires, often resulting in suboptimal subsets. To address these issues, we propose **IF-Beta**, an efficient data pruning framework that combines influence function and a learnable sampling policy. Empirically, we first demonstrate that influence functions can serve as an effective and efficient estimator of sample impact in KD settings, where only a pretrained teacher is available. Building on this, our sampling policy is specifically parameterized by a Beta distribution, whose highly flexible two-parameter family allows the policy to adapt to diverse pruning regimes rather than being tied to fixed heuristic forms. Next, we formulate KD pruning as optimizing this policy through a bilevel objective, where the inner loop operates in the teacher's feature space with a KD-aligned objective, enabling fast proxy training, while the outer loop updates the policy parameters to maximize the distillation performance. This design ensures IF-Beta is both computationally efficient and inherently aligned with the goals of KD. Extensive experiments on CIFAR-10/100 and ImageNet show that IF-Beta consistently outperforms other baselines across a wide range of pruning ratios. Remarkably, IF-Beta does enable students trained on less data and less compute to surpass the performance of students distilled on the full dataset.

Keywords: Data Pruning · Knowledge Distillation · Influence Function · Bilevel Optimization

∗ Equal contribution. ✉ Corresponding authors.

1 Introduction

Over the past decade, deep neural networks (DNNs) [12, 16, 23] have achieved remarkable success across a wide range of domains, largely driven by the increasing scale of models and datasets. Yet such large-scale models inevitably incur substantial computation cost at inference, which severely limits their applicability in real-world scenarios. To address this issue, [18] introduced Knowledge Distillation (KD), enabling compact students to inherit the knowledge of cumbersome teachers and thereby achieve efficient inference with competitive performance [14, 20, 39]. Despite its inference benefits, KD does not alleviate the training burden: students are usually distilled on the entire dataset under the guidance of a large teacher, resulting in significant training cost [7].

Several prior methods have been explored to accelerate KD training [6, 25, 33, 35], yet their speedups typically come at the cost of degrading student performance [7]. Alternatively, an emerging line of research [4, 7] points to another promising direction: reducing the number of samples that participate in distillation, *i.e.*, data pruning [26,30], which aims to retain only a subset of training samples while maintaining competitive performance of the models. Baruch et al. [4] was the first to systematically re-evaluate classical pruning methods [29, 31, 34] in KD, demonstrating that data pruning can indeed be effective for KD. More recently, Chen et al. [7] introduced a pruning method specifically tailored for KD, which directly employs the teacher's cross-entropy (CE) loss as a difficulty metric and selects samples whose losses fall within a "medium" range window.

Despite these advances, data pruning in the context of KD remains hindered by two fundamental, yet orthogonal, obstacles: (i) **Efficiency-Effectiveness Trade-off in Score Estimation**. High-fidelity trajectory-based metrics (*e.g.*, EL2N [29], Forgetting [34], and AUM [31]) are computationally prohibitive for KD, as they depend on training-time statistics that are generally not provided with pretrained teachers. Consequently, obtaining these scores necessitates expensive retraining, which largely offsets the efficiency gains intended by pruning. In contrast, the CE loss adopted by Chen et al. [7] does greatly reduce time cost by computing only the per-sample loss. However, it is an unreliable indicator of sample difficulty, as over-parameterized deep networks are capable of overfitting even hard or mislabeled samples [2], resulting in small yet weakly discriminative losses across samples of widely varying difficulty (more details see Fig. 1). (ii) **Rigid and Suboptimal Sampling**. Even with reliable scores, conventional sampling strategies rely heavily on manual heuristics, such as top-k [4], stratified sampling with a fixed cut-off ratio [40], or window-based selection over the score distribution [7, 9]. These hand-crafted designs share a common limitation: they impose a fixed selection region that cannot adapt to the varying data distributions and teacher-student dynamics across different KD settings, inevitably leading to limited performance.

To overcome the aforementioned limitations, we propose an optimized KD pruning framework, **Influence Function based Beta Policy (IF-Beta)**, which unifies IF scoring with a learnable probabilistic pruning policy. In particular, we first revisit the influence function (IF) [21], which is a post-hoc ap-

proach to quantify the impact of each training data. Thanks to the recent advances [36], we show that IF serves as an efficient and effective score estimator for KD-oriented data pruning, replacing prior score-based approaches. Building upon this IF-based scoring, we further introduce a learnable probabilistic pruning policy based on a parameterized Beta distribution, called Beta Policy, which provides a shape-adaptive mechanism for selecting data. Formally, we pose the optimization of our Beta Policy as the following bilevel optimization problem:

$$\min_{\phi} \Phi(\phi) = \mathbb{E}_{\boldsymbol{m}\sim\pi_\phi^r} \widehat{\mathcal{L}}(\theta_S^*(\boldsymbol{m})), \tag{1}$$

$$s.t.\ \theta_S^*(\boldsymbol{m}) = \arg\min_{\theta_S} \mathcal{L}(\theta_S; \theta_T, \boldsymbol{m}), \tag{2}$$

where ϕ parameterizes the Beta Policy π_ϕ^r over the scoring space. The inner-level objective trains the student θ_S on the subset of samples specified by the mask $\boldsymbol{m} \sim \pi_\phi^r$, using the teacher θ_T for guidance through the KD loss $\mathcal{L}(\theta_S; \theta_T, \boldsymbol{m})$. The outer-level objective then updates ϕ, evaluated by the validation objective $\widehat{\mathcal{L}}(\theta_S^*)$, to improve the expected distillation performance. To further realize this efficiently, we simulate the student side of KD using a lightweight linear classifier trained on the frozen teacher features. This proxy-student formulation faithfully captures the knowledge transfer dynamics of KD while keeping the overall computational cost negligible. Overall, we summarize our contributions as follows:

- We investigate the influence function as an efficient and effective estimator for KD pruning, and empirically show that IF can substitute both difficulty-based and uncertainty-based scores utilized in conventional data pruning without requiring retraining.
- We introduce a learnable data pruning policy based on a parameterized Beta distribution, which can be optimized efficiently via a bilevel formulation in the teacher's feature space.
- Extensive experiments demonstrate that IF-Beta achieves consistently strong performance across KD settings, and show that IF-Beta can surpass full-data distillation using less data and less compute.

2 Preliminaries

We first revisit the formulation of knowledge distillation (KD) and influence functions (IF), which together establish the basis in our framework. A more detailed discussion, along with other related works (including knowledge distillation, data pruning and influence function), is provided in the Appendix.

Knowledge Distillation. The goal of KD is to transfer the knowledge of a large teacher network into a compact student model, enabling efficient inference while maintaining competitive performance [18]. In this work, we typically focus on the standard KD framework for image classification tasks, which simply leverages the teacher's predictive distribution to guide the optimization of the student.

Let the training dataset be $D_{\text{tr}} = \{z_n = (x_n, y_n)\}_{n=1}^{N}$. Given a teacher network f_T and a student network f_S, the student is trained on D_{tr} by minimizing a weighted combination of the standard cross-entropy loss and a distillation term:

$$\ell_{\text{KD}} = (1-\alpha)\,\ell_{\text{CE}} + \alpha\,\ell_{\text{KL}}, \tag{3}$$

where $\alpha \in [0,1]$ and the distillation term is the Kullback–Leibler (KL) divergence between the teacher and student predictive distributions: $\ell_{\text{KL}} = \text{KL}(f_T(x) \,\|\, f_S(x))$.

Data Pruning. Data pruning aims to select a subset of training samples that preserves the performance of the full dataset while reducing computational cost. Most existing pruning methods [4, 7, 9, 40] follow a two-stage pipeline: (i) estimating a score for each sample, and (ii) selecting a subset based on these scores under a given pruning ratio.

For score estimation, prior work typically measures either sample difficulty (*e.g.*, EL2N [29], Forgetting [34], AUM [31]) or predictive uncertainty (*e.g.*, Dyn-Unc [17], DUAL [8]), which are derived from training dynamics (*i.e.*, per-epoch predictions) and therefore require full or partial model retraining. In KD, however, only a pretrained teacher is available, without its training trajectory. This motivates the need for effective and efficient post-hoc score estimators that can be computed directly from a pretrained model.

Given the estimated scores, existing methods typically apply heuristic selection rules, such as top-k [4], stratified sampling with manually designed cut-off ratio [40], or sliding-window mechanisms [7,9]. Such heuristic strategies may limit subset optimality under different pruning regimes, motivating more principled and adaptive selection mechanisms.

Influence Functions. To address the limitation of trajectory-dependent scoring, we revisit influence functions as a potential post-hoc estimator, which aims to understand the influence of individual training points on the model's predictions, *i.e.*, how the model's output would change if a particular training point were removed [10, 15].

Consider a model, whose parameters $\theta^\star$ are obtained via empirical risk minimization (ERM) over D_{tr}:

$$\theta^\star := \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \ell(z_n, \theta), \tag{4}$$

where ℓ denotes a per-sample loss function, *e.g.*, the cross-entropy loss for classification tasks.

Next consider a validation set $D_{\text{val}} = \{z_m = (x_m, y_m)\}_{m=1}^{M}$. Following Koh and Liang [21], the influence of a training point $z_{\text{tr}} \in D_{\text{tr}}$ on a specific validation example $z_{\text{val}} \in D_{\text{val}}$ can be expressed as:

$$\mathcal{I}(z_{\text{tr}}; z_{\text{val}}) := g_{z_{\text{val}}}^{\top} H_{\theta^\star}^{-1} g_{z_{\text{tr}}}, \tag{5}$$

where $g_z = \nabla_{\theta^\star} \ell(z, \theta^\star)$ denotes the gradient of the loss w.r.t. $\theta^\star$ for z, and $H_{\text{tr}} = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta^\star}^2 \ell(z_n, \theta^\star)$ is the Hessian matrix computed over D_{tr}. Therefore,

the influence function of z_{tr} on D_{val} can be extended as follows:

$$\mathcal{I}(z_{\text{tr}}; D_{\text{val}}) := \frac{1}{M} \sum_{m=1}^{M} g_{z_m}^{\top} H_{\text{tr}}^{-1} g_{z_{\text{tr}}}, \tag{6}$$

where $z_m \in D_{\text{val}}$.

This allows us to quantify the contribution of each training point. Despite its appealing theoretical foundation, traditional IF estimators were long considered impractical due to their high computational cost [24] and unreliable estimation quality [3, 5].

3 Method

In this section, we first introduce an efficient post-hoc score estimator based on influence functions (Section 3.1). Building upon the estimated scores, we then propose an adaptive sampling strategy via a Beta Policy (Section 3.2). The search for the optimal policy is formulated as a bilevel optimization problem (Section 3.3), for which we further present an efficient solution to this optimization in the feature space under the KD setting (Section 3.4).

3.1 IF Serves as an Efficient Score Estimator

To address the limitation of trajectory-dependent scoring in KD, we revisit influence functions as a post-hoc estimator. Although classical influence estimation suffers from instability and high computational cost in deep neural networks, recent advances [36] show that optimizing towards a flat validation minimum (FVM) significantly stabilizes influence estimation (also observed in Fig. 1).

Following this insight, we fine-tune the pretrained teacher model θ_T on the validation set $\mathcal{D}_{\text{val}}$ using Sharpness-Aware Minimization (SAM) [13]. Typically, we solve:

$$\tilde{\theta}_T := \arg\min_{\theta_T} \hat{R}_{\text{val}}^{\gamma}(\theta_T), \tag{7}$$

$$\text{where } \hat{R}_{\text{val}}^{\gamma}(\theta_T) := \max_{\|\Delta\| \leq \gamma} \hat{R}_{\text{val}}(\theta_T + \Delta). \tag{8}$$

Here, γ is a fixed constant and $\hat{R}_{\text{val}}$ denotes the empirical validation risk. This fine-tuning step is highly efficient because it operates directly on the given teacher model, introducing only modest computational overhead. The SAM-style objective improves the stability of curvature estimation by enforcing local flatness of the loss landscape, thereby enabling IF to achieve more reliable influence estimation than previous methods.

Let $\tilde{g}_z = \nabla_{\tilde{\theta}_T} \ell(z, \tilde{\theta}_T)$ denote the gradient of the loss w.r.t $\tilde{\theta}_T$ for the sample z, and let $\tilde{H}_{\text{val}} = \frac{1}{M} \sum_{m=1}^{M} \nabla_{\tilde{\theta}_T}^2 \ell(z_m, \tilde{\theta}_T)$ represent the Hessian matrix computed

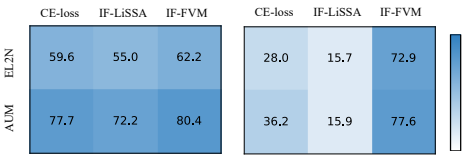
over the validation set D_{val}. Under the FVM condition, the influence of a training sample z_{tr} on the validation set D_{val} can then be derived as:

$$\mathcal{I}(z_{\text{tr}}, D_{\text{val}}) := \tilde{g}_{z_{\text{tr}}}^{\top} \tilde{H}_{\text{val}}^{-1} \tilde{g}_{z_{\text{tr}}}, \tag{9}$$

which we adopt as the score estimator for sample selection. A detailed derivation is provided in the Appendix.

In practice, directly computing Eq. 9 can be inefficient, as it requires estimating the inverse Hessian matrix $\tilde{H}_{\text{val}}^{-1}$. To address this issue, we approximate the inverse Hessian using the inverse of a diagonal Fisher Information matrix, $\text{diag}\left(\frac{1}{|D_{\text{tr}}|} \sum_{z_{\text{tr}} \in D_{\text{tr}}} \tilde{g}_{z_{\text{tr}}} \tilde{g}_{z_{\text{tr}}}^{\top}\right)$, which yields a highly efficient approximation and makes the overall score computation practical at scale.

Fig. 1 validates the reliability of our proposed IF-FVM score estimator. Specifically, we benchmark various post-hoc score estimators by computing the Spearman rank correlation between their scores and two established trajectory-based difficulty metrics: EL2N [29] and AUM [31]. These metrics are widely regarded as faithful proxies for sample hardness. Intuitively, a higher correlation coefficient indicates that the estimator more accurately captures the intrinsic difficulty of samples by effectively aligning with the learning dynamics observed during training. As shown, IF-FVM consistently maintains a strong correlation across datasets, outperforming both the CE-loss baseline [7] and IF-LiSSA [1]. Our results suggest that IF-FVM serves as a computationally efficient yet robust alternative to metrics that otherwise require expensive trajectory tracking.

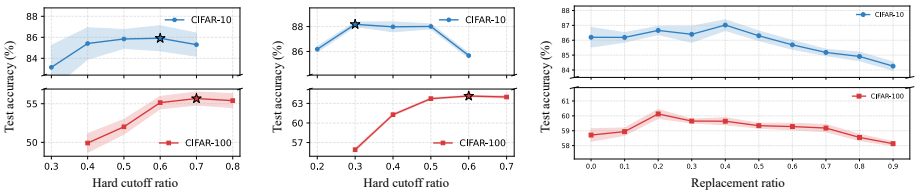


(a) CIFAR-10 **(b)** CIFAR-100

Fig. 1: Spearman rank correlation (%) between post-hoc score estimators and trajectory-based difficulty metrics. Higher values indicate stronger alignment in sample ranking.

3.2 Adaptive Sampling via Beta Policy

Having established that IF under FVM serves as an effective score estimator, we next pose a natural question: *Can existing sampling strategies for sample selection be directly applied to KD scenarios without compromising performance or efficiency?* While applicable in principle, their effectiveness in KD remains suboptimal. Specifically, we identify critical limitations in representative methods, CCS [40] and BWS [9], within KD contexts: CCS exhibits a threshold shift, where the optimal cutoff ratio varies significantly between KD and non-KD settings, rendering pre-defined hyperparameters non-transferable (Figs. 2a and 2b). Meanwhile, BWS is hindered by a suboptimal fixed-window design; as our replacement experiments (Fig. 2c) reveal, introducing randomness into the selection window actually improves performance. These findings (further detailed in the Appendix) underscore the need for an adaptive, learnable sampling strategy that overcomes the rigidity of current heuristic-based selection.



(a) CCS w/o KD. **(b)** CCS w/ KD. **(c)** Best-window w/ replacement.

Fig. 2: Limitations of heuristic sampling methods on ResNet-18 with a pruning ratio 90%. (a,b) CCS with different hard cutoff ratios when w/o KD and W/ KD. (c) BWS with different replacement ratios with outside samples.

To address the aforementioned limitations, we propose a Beta-based sampling policy. Given a precomputed difficulty score s_i for each sample z_i, we apply rank-to-percentile normalization to map the scores onto the interval $[0, 1]$, obtaining normalized values $\hat{s}_i \in [0, 1]$, where smaller values correspond to more difficult samples. Accordingly, we parameterize the selection probability of each sample using the Beta distribution:

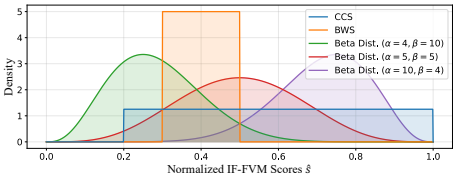


Fig. 3: Comparison of our Beta Policy with CCS and BWS. Our Beta Policy provides a more flexible and adaptive sampling distribution.

$$p_i(\phi) = \frac{1}{Z} \frac{1}{B(\alpha(\phi), \beta(\phi))} \hat{s}_i^{\alpha(\phi)-1} (1 - \hat{s}_i)^{\beta(\phi)-1}, \tag{10}$$

where $B(\cdot, \cdot)$ is the Beta function, and Z is a normalization constant ensuring $\sum_i p_i = 1$. Here, ϕ represents the learnable parameters that jointly control α and β. As illustrated in Fig. 3, our approach provides a continuous and flexible density over the sample space. This contrasts with the rigid, uniform sampling over predefined support sets used in CCS and BWS, allowing for a more adaptive selection strategy.

Based on the parameterized sampling probability $p_i(\phi)$ defined in Eq. 10, we introduce the **Beta Policy** π_ϕ^r, a Beta distribution-based categorical distribution over the training samples conditioned on ϕ and r. This policy enables sampling a binary mask vector $\boldsymbol{m} \in \{0, 1\}^N$ with a fixed pruning ratio $r \in [0, 1]$ according to

$$\boldsymbol{m} \sim \pi_\phi^r := p(\boldsymbol{m} \mid \phi, r), \tag{11}$$

where

$$p(\boldsymbol{m} \mid \phi, r) = \frac{1}{Z_{K_r}} \prod_{i=1}^{N} p_i(\phi)^{m_i} 1_{\|\boldsymbol{m}\|_1 = K_r}(\boldsymbol{m}), \tag{12}$$

$1_{\|\boldsymbol{m}\|_1 = K_r}(\cdot)$ denotes the indicator function and Z_{K_r} is the normalization constant ensuring a valid categorical distribution over all masks satisfying $\|\boldsymbol{m}\|_0 = K_r = (1 - r)N$. The sampled mask vector $\boldsymbol{m}$ then defines a pruned subset of the training data, $\mathcal{D}_{\text{sub}} = \{z_i \mid m_i = 1, z_i \in \mathcal{D}_{\text{tr}}\}$.

3.3 Bilevel Optimization for the Beta Policy

Following Zhou et al. [41], we formulate the search for the optimal policy π_ϕ^r as a bilevel optimization problem:

$$\min_{\phi} \Phi(\phi) = \mathbb{E}_{\boldsymbol{m}\sim\pi_\phi^r} \widehat{\mathcal{L}}(\theta_S^*(\boldsymbol{m})), \tag{13}$$

$$s.t.\ \theta_S^*(\boldsymbol{m}) = \arg\min_{\theta_S} \mathcal{L}(\theta_S; \theta_T, \boldsymbol{m}). \tag{14}$$

Here, θ_S denotes the model parameter of the student model. The inner objective evaluates the training risk of θ on the sampled subset $\mathcal{D}_{\text{sub}}$ using the Knowledge Distillation loss ℓ_{KD}:

$$\mathcal{L}(\theta_S; \theta_T, \boldsymbol{m}) = \frac{1}{K} \sum_{z_i \in D_{\text{sub}}} \ell_{\text{KD}}(z_i, \theta_S, \theta_T), \tag{15}$$

where θ_T represents the parameters of the teacher model. The outer objective measures the validation risk of the optimized model $\theta_S^*(\boldsymbol{m})$ on the validation set $\mathcal{D}_{\text{val}}$ using Cross Entropy loss ℓ_{CE}:

$$\widehat{\mathcal{L}}(\theta_S^*(\boldsymbol{m})) = \frac{1}{M} \sum_{z_i \in D_{\text{val}}} \ell_{\text{CE}}\big(z_i, \theta_S^*(\boldsymbol{m})\big). \tag{16}$$

Intuitively, this bilevel formulation seeks the optimal policy such that the sampled subset induces a model that generalizes best on the validation set.

However, directly solving this bilevel optimization yields a gradient estimate of the form $\nabla_\phi \Phi(\phi) \approx \nabla_\phi \theta_S^*(\boldsymbol{m}) \nabla_{\theta_S} \widehat{\mathcal{L}}(\theta_S^*(\boldsymbol{m}))$, which requires implicit differentiation of the inner optimum, *i.e.*, $\nabla_\phi \theta_S^*(\boldsymbol{m})$, and is thus computationally prohibitive for deep neural networks. By exploiting the probabilistic formulation of our Beta Policy π_ϕ^r, we circumvent this issue by adopting a policy gradient estimator, thereby eliminating the need for costly implicit differentiation.

Particularly, we establish the following result:

Theorem 1 (Unbiased Policy-Gradient Estimator). *Let π_ϕ^r be a differentiable sampling policy over mask vectors $\boldsymbol{m} \in \{0,1\}^N$. Define the policy-gradient estimator*

$$\hat{g}(\boldsymbol{m}) = \widehat{\mathcal{L}}\left(\theta_S^*(\boldsymbol{m})\right) \nabla_\phi \ln p(\boldsymbol{m}|\phi, r). \tag{17}$$

Then $\hat{g}(\boldsymbol{m})$ is an unbiased estimator of the true gradient $\nabla_\phi \Phi(\phi)$:

$$\mathbb{E}_{\boldsymbol{m}\sim\pi_\phi^r}\left[\hat{g}(\boldsymbol{m})\right] = \nabla_\phi \Phi(\phi). \tag{18}$$

The proof is provided in the Appendix. Theorem 1 ensures that $\hat{g}(\boldsymbol{m})$ provides an unbiased stochastic gradient for optimizing ϕ. Consequently, given the inner-loop optimum $\theta_S^*(\boldsymbol{m})$, the policy parameter ϕ can then be updated via stochastic gradient descent:

$$\phi \leftarrow \phi - \eta\, \widehat{\mathcal{L}}\left(\theta_S^*(\boldsymbol{m})\right)\, \nabla_\phi \ln p(\boldsymbol{m}|\phi, r), \tag{19}$$

where η denotes the learning rate.

3.4 Efficient Bilevel Optimization in Feature Space

While the outer-loop objective in Eq. 13 can be efficiently optimized via the policy gradient estimator (Eq. 19), the inner-loop optimization in Eq. 14 remains computationally expensive, as it requires retraining the student network to convergence for every policy update.

To address this issue, we leverage the structure of the KD setting. Remarkably, instead of training the student model's full parameter set θ from scratch, we utilize a surrogate student by attaching a trainable linear classification head f_L to the fixed pretrained teacher model with parameters θ_T. Let $f_T(\cdot)$ denote the fixed feature extractor parameterized by θ_T. Each training sample x_i is thus represented by its frozen feature $\mathbf{f}_i = f_T(x_i) \in \mathbb{R}^d$. Accordingly, the student model is formulated as:

$$f_{\theta_S}(x_i) := f_L(\mathbf{f}_i) = \boldsymbol{C}\,\mathbf{f}_i, \tag{20}$$

where $\boldsymbol{C} \in \mathbb{R}^c \times \mathbb{R}^d$ denotes the parameters of the linear classifier that maps feature vectors to class logits.

Finally, the inner-loop objective in Eq. 15 can be reformulated as

$$\boldsymbol{C}^*(\boldsymbol{m}) = \arg\min_{\boldsymbol{C}} \frac{1}{K} \sum_{z_i \in D_{\text{sub}}} \ell_{\text{KD}}(z_i, \boldsymbol{C}, \theta_T), \tag{21}$$

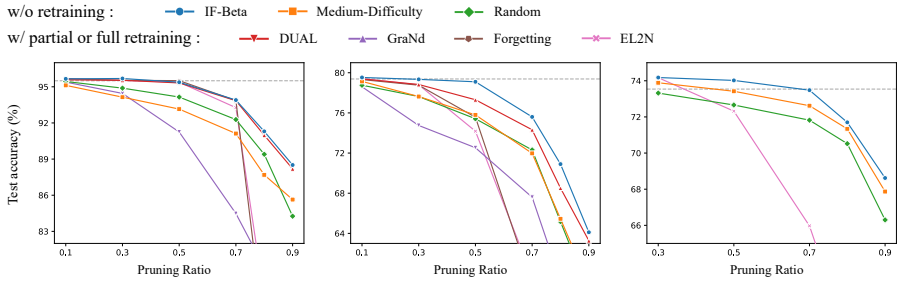
which can be efficiently optimized since it involves learning only a single linear layer. Further detailed implementation and the complete pipeline of IF-Beta are provided in the Appendix.

Remark. It is worth noting that the bilevel formulation in Eq. 15 presented in Section 3.3 is general and can be directly applied to both KD scenarios [18] and standard supervised classification [11, 22]. In particular, by replacing the inner objective in Eq. 15 with a standard CE loss, our approach naturally transfers to a general data pruning framework. To demonstrate this versatility, we further conduct standard data pruning experiments following prior works [8, 9, 40], with the results reported in Section 4.4.

4 Experiments

In this section, we present extensive experiments to evaluate the effectiveness and generality of IF-Beta. We first compare existing data pruning methods for distillation across three representative datasets (CIFAR-10/100 [22] and ImageNet [11]) under various pruning ratios, and with different teacher architectures (*e.g.*, ResNet [16], WideResNet [37], ViT [12]).

Baselines. The baselines considered in this work can be grouped into two categories: (1) methods that require partial or full retraining to obtain training dynamics, including EL2N [29], GraNd [29], Forgetting [34], and DUAL [8]; and (2) methods that do not require retraining, including Random [4] and Medium-Difficulty [7]. In addition, to fairly compare with recent score-based sampling



(a) CIFAR-10 **(b)** CIFAR-100 **(c)** ImageNet

Fig. 4: Performance comparison between IF-Beta and other baselines on CIFAR-10/100 and ImageNet under the KD setting, where for CIFAR-10/100 both teacher and student are ResNet-18, and for ImageNet both are ResNet-50. The pruning ratio is the fraction of examples removed from the original datasets. The dashed horizontal line denotes the student distilled on the full dataset (without pruning). Detailed numerical results are provided in the Appendix.

strategies (CCS [40] and BWS [9]), we implement them with our introduced IF-based scores, resulting in IF-CCS and IF-BWS, which also do not require retraining.

Training Details. For distillation, all models are trained with the loss in Eq. 3. Noting that Chen et al. [7] also introduce a new distillation loss, we additionally report results under this loss and denote it as Medium-Difficulty*. Due to computational constraints, on ImageNet we only compare against EL2N and the six baselines that do not require retraining (including Medium-Difficulty*). For CIFAR-10/100, all models (across all pruning ratios) are trained with a batch size of 128 for 40K iterations (about 200 epochs over the entire dataset), while for ImageNet all student models are trained with a batch size of 128 for 300K iterations (about 30 epochs over the entire dataset). All results are averaged over three random seeds (0, 1, 2). Additional experimental details are provided in the Appendix.

4.1 Main Results: Data Pruning for Knowledge Distillation

We first evaluate data pruning under the KD setting where the teacher and student share the same architecture. As shown in Fig. 4, IF-Beta consistently outperforms all baselines across all datasets and pruning ratios. Notably, even compared to the recent method DUAL [8], which requires partial retraining (60 epochs) to extract training dynamics, IF-Beta consistently achieves superior accuracy. Compared with recent Medium-Difficulty [7], which also does not require retraining, IF-Beta even exhibits a clear performance advantage over it.

Remarkably, IF-Beta can surpass full-data distillation using only a subset of the training data: retaining 70% on CIFAR-10 achieves 95.69% vs. 95.50%, 90% on CIFAR-100 reaches 79.53% vs. 79.38%, and 50% on ImageNet already yields 74.02% vs. 73.54%. This suggests that in KD, not all training samples

Table 1: Performance comparison between full-data distillation, IF-Beta, and Random (pruning 10%, 30%, 50%) under limited training budgets on CIFAR-10/100 (10K iterations) and ImageNet (150K iterations). **Bold** denotes the best results.

Prune	Method	CIFAR-10	CIFAR-100	ImageNet
0%	Full-data	86.62± 1.67	63.54± 2.21	57.16± 0.14
10%	IF-Beta	**87.81**± 0.58	**66.26**± 0.91	–
	Random	87.30± 0.62	64.95± 0.42	–
30%	IF-Beta	**89.23**± 1.05	**66.53**± 1.11	**61.35**± 0.09
	Random	88.03± 0.83	63.50± 2.42	57.48± 0.17
50%	IF-Beta	**90.80**± 1.29	**67.31**± 1.52	**60.70**± 0.12
	Random	86.37± 0.85	63.39± 1.29	57.12± 0.15

are equally beneficial—some may actively impede student generalization, and carefully pruning them away yields a cleaner and more informative training signal.

Finally, IF-Beta is further validated under heterogeneous teacher architectures spanning ResNet-50/101 [16], WideResNet [37], and ViT-Base [12], consistently outperforming all retraining-free baselines across all settings (see Appendix).

4.2 Further Analysis of IF-Beta

Beyond the main results, we further examine IF-Beta from four complementary perspectives: (i) sample efficiency under constrained training budgets, (ii) end-to-end compute savings relative to full-data training, (iii) advantages over other efficient KD methods, and (iv) generality across diverse distillation objectives.

Sample Efficiency under Constrained Budgets. As shown in Tab. 1, we evaluate IF-Beta under severely constrained training budgets (10K iterations for CIFAR-10/100 and 150K for ImageNet). IF-Beta consistently outperforms both full-data distillation and random pruning across all pruning ratios and datasets—with 50% data pruned, IF-Beta achieves 90.80% on CIFAR-10, 67.31% on CIFAR-100, and 60.70% on ImageNet, all exceeding the full-data baselines. Crucially, the consistent margin over random pruning demonstrates that these gains are not solely attributable to a more favorable training regime induced by fewer samples,[5] but stem from effective sample selection itself.

End-to-End Compute Savings. We then examine how much time is actually needed (with less data) to reach the final performance of the full–data student. As shown in Tab. 2, across all three datasets, IF–Beta can surpass the final full–data student using fewer samples and in much lower time cost, even when the cost of IF–Beta pruning process is included.

Comparison with Efficient KD Methods. Following the evaluation protocol of Chen et al. [7], we compare IF-Beta against other efficient KD methods

[5] With a fixed iteration budget, training on a smaller subset effectively increases the number of passes over each sample, which alone can improve generalization.

Table 2: Compute comparison between IF–Beta and full–data training when KD. Relative cost denotes the ratio of time between: (i) the total number of iterations IF–Beta actually consumes until surpassing the final full–data students' performance (including the pruning process), and (ii) the standard full–data training schedule. **Bold** denotes the best results.

Dataset	Prune	Relative Cost	**IF-Beta**	**Full Data**
CIFAR-10	30%	0.806×	**95.64**	95.50
CIFAR-100	10%	0.845×	**79.45**	79.38
ImageNet	50%	0.915×	**73.67**	73.54

Table 3: Test accuracy and training time on ImageNet of prior efficient KD methods and our IF-Beta using 70% of samples, where the teacher is ResNet-34 and the student is MobileNet. **Bold** denotes the best results and underline denotes the second-best.

	Baseline [19]	Zipf's LS [25]	USKD [35]	Medium-Difficulty* [7]	IF-Beta
Acc (%)	69.57	69.59	70.38	<u>70.92</u>	**71.20**
Time (h)	39.86	> 39.86	> 39.86	<u>36.98</u>	**19.24**

by distilling ResNet-34 into MobileNet on ImageNet using a single NVIDIA A100 GPU. As shown in Tab. 3, IF-Beta achieves the best accuracy of 71.20% while requiring only 19.24 hours—less than half the cost of the full-data Baseline (39.86h) and Medium-Difficulty* (36.98h). The efficiency gains stem from three factors: reduced data volume (70% subset), no repeated teacher queries per epoch, and a higher-quality subset that preserves distillation effectiveness even under one-shot teacher logits. Further detailed discussion is provided in the Appendix.

Generality Across Distillation Objectives. To verify that IF-Beta is not tied to a specific distillation loss, we evaluate it under three representative KD objectives on CIFAR-10/100: Relational KD [27] (relation-based), FitNets [32] (feature-based), and Attention Transfer [38] (attention-based). As shown in Tab. 4, IF-Beta consistently outperforms both Random pruning and Medium-Difficulty

Table 4: Performance under different KD losses and pruning ratios on CIFAR-10/100. Med-Dif denotes Medium Difficulty. **Bold** denotes the best results

Prune	Method	Relational KD		FitNets		Attention Transfer	
		CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
0%	Full-data	95.52	78.56	95.35	78.32	95.70	78.77
30%	Random	95.07	76.19	94.57	75.08	94.52	75.63
	Med-Dif*	94.87	77.04	93.97	75.53	94.02	75.77
	IF-Beta	**95.24**	**78.59**	**95.05**	**77.94**	**95.34**	**78.52**
50%	Random	94.64	74.09	93.65	71.46	93.65	71.73
	Med-Dif*	94.11	74.26	92.79	71.33	92.96	72.65
	IF-Beta	**95.67**	**77.82**	**95.16**	**77.56**	**95.04**	**77.58**
70%	Random	93.70	69.21	91.54	64.99	92.06	66.08
	Med-Dif*	93.18	69.54	90.49	65.04	91.22	66.69
	IF-Beta	**94.53**	**73.36**	**93.49**	**71.52**	**93.64**	**72.26**

Table 5: Performance comparison of different sampling strategies (IF-Beta vs. IF-CCS and IF-BWS) under the KD setting across various pruning ratios on CIFAR-10/100 and ImageNet. **Bold** denotes the best results and underline denotes the second-best.

Method	90%	80%	70%	50%	30%	10%
CIFAR-10						
IF-CCS	88.18	90.87	93.39	93.77	94.65	95.40
IF-BWS	86.20	86.50	89.80	94.82	95.51	95.47
IF-Beta	**88.51**	**91.30**	**93.90**	**95.38**	**95.69**	**95.66**
CIFAR-100						
IF-CCS	63.73	70.42	72.22	77.64	78.47	78.58
IF-BWS	58.71	63.91	71.75	75.81	78.85	78.95
IF-Beta	**64.11**	**70.90**	**75.60**	**79.10**	**79.34**	**79.53**
ImageNet						
IF-CCS	65.73	68.31	71.68	72.98	73.32	-
IF-BWS	67.90	70.56	72.86	73.63	73.66	-
IF-Beta	**68.62**	**71.70**	**73.48**	**74.02**	**74.18**	-

across all losses, pruning ratios, and datasets, demonstrating its effectiveness and generality.

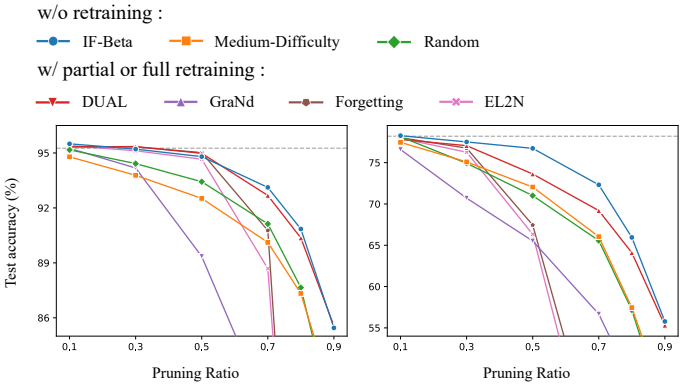
Remark. Taken together, these results suggest a broader perspective on efficient KD: the *quality* of the training data may matter more than the *specifics* of the distillation algorithm. Across diverse training budgets, distillation objectives, and teacher-student pairs, IF-Beta consistently benefits from selecting a more informative subset—suggesting that careful data curation is a universally effective and complementary strategy for improving both the efficiency and effectiveness of knowledge distillation.

4.3 Ablation Study: Advantages of the Learnable Beta Policy

We also examine the performance of CCS [40] and BWS [9] when combined with our influence scores (*i.e.*, IF-CCS and IF-BWS) on CIFAR-10/100 and ImageNet. As shown in Tab. 5, IF-Beta consistently outperforms both variants across all pruning ratios and datasets, demonstrating that the learnable Beta Policy provides clear advantages over fixed-threshold and fixed-window selection, consistent with our analysis in Section 3.2.

4.4 Data pruning for non-KD Scenario

To further verify the generality of IF-Beta beyond KD, we also evaluate it under the standard data pruning setting following CCS [40]. In this case, the inner objective in Eq. 15 is replaced with a standard CE loss, corresponding to conventional supervised training rather than distillation. As shown in Fig. 5, IF-Beta remains competitive and consistently outperforms other baselines across all pruning ratios on CIFAR–10/100. These results indicate that IF-Beta is not tied to KD, but can also serve as a general-purpose pruning mechanism under standard supervised learning, without relying on training-dynamics trajectories.



(a) CIFAR-10 **(b)** CIFAR-100

Fig. 5: Performance comparison between IF-Beta (w/o KD) and other baselines with ResNet-18 on CIFAR-10/100 under standard data pruning setting (*i.e.*, training without KD). The pruning ratio is the fraction of examples removed from the original datasets. The dashed horizontal line denotes the model trained on the full dataset. Detailed numerical results are provided in the Appendix.

Further Discussion. IF-Beta is originally designed for KD, where a pretrained teacher model is naturally available to compute influence scores. However, under standard data pruning, such in-domain pretrained models are typically not provided unless one retrains a model first. Fortunately, thanks to the nature of FVM [36], we can simply replace the in-domain model with a generic off-the-shelf pretrained model (*e.g.*, trained on ImageNet) to compute influence scores. As shown in Tab. 6, under the setting of conventional training without KD, using only an ImageNet-pretrained ResNet-18 from the PyTorch model zoo [28] to compute IF on CIFAR-10/100 still yields competitive pruning performance, with only minor degradation. This means that, for standard data pruning, IF-Beta can completely avoid any retraining cost while still providing strong pruning quality.

Table 6: Performance comparison under standard data pruning with various data pruning ratios on CIFAR-10/100. IF-Beta (w/o KD) uses an in-domain pretrained model to compute influence scores, while IF-Beta* (w/o KD) uses a generic ImageNet-pretrained ResNet-18 (from PyTorch model zoo [28]). **Bold** denotes the best results and underline denotes the second-best.

Method	90%	80%	70%	50%	30%	10%
CIFAR-10						
Random	80.17	87.65	91.13	93.43	94.42	95.17
IF-Beta	**85.45**	**90.85**	**93.12**	**94.80**	**95.21**	**95.50**
IF-Beta*	85.34	90.22	92.22	94.59	95.14	95.37
CIFAR-100						
Random	44.58	57.16	65.57	71.01	74.93	78.14
IF-Beta	**55.78**	**65.95**	**72.32**	**76.73**	**77.51**	**78.27**
IF-Beta*	53.31	63.09	69.50	74.79	76.54	77.60

5 Conclusion

This paper proposes IF-Beta, a data-pruning framework for knowledge distillation that accelerates student training by identifying the most informative samples for distillation. IF-Beta unifies influence function-based scoring with a learnable Beta Policy parameterized by a Beta distribution, which can be efficiently optimized through a bilevel formulation in the teacher's feature space. The influence function provides a reliable and retraining-free estimate of sample importance, while the Beta Policy overcomes the rigidity of heuristic selection strategies by adaptively learning the optimal sampling distribution. Extensive experiments on CIFAR-10/100 and ImageNet demonstrate that IF-Beta consistently outperforms existing baselines across a wide range of pruning ratios, and remarkably enables students trained on substantially less data and under lower compute budgets to surpass full-data distillation. Beyond KD, IF-Beta generalizes naturally to standard data pruning settings, demonstrating its effectiveness as a broadly applicable pruning strategy regardless of the training objective. These results suggest that *which* samples to distill may matter more than *how* to distill them, positioning data curation as a principled and practical lever for improving efficiency in modern training pipelines.

Acknowledgements

This work was supported by the National Nature Science Foundation of China (62472097), Shanghai Municipal Science and Technology Commission (Grant No.25511102200), Fudan Kunpeng & Ascend Center of Cultivation, and Shanghai science and technology committee (Grant No.25511106200). The computations in this research were performed on the CFFF platform of Fudan University.

References

1. Agarwal, N., Bullins, B., Hazan, E.: Second-order stochastic optimization for machine learning in linear time. J. Mach. Learn. Res. **18**, 116:1–116:40 (2017)
2. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A.C., Bengio, Y., Lacoste-Julien, S.: A closer look at memorization in deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 233–242. PMLR (2017)
3. Bae, J., Ng, N., Lo, A., Ghassemi, M., Grosse, R.B.: If influence functions are the answer, then what is the question? In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
4. Baruch, E.B., Botach, A., Kviatkovsky, I., Aggarwal, M., Medioni, G.: Distilling the knowledge in data pruning. In: Forty-second International Conference on Machine Learning (2025)

5. Basu, S., Pope, P., Feizi, S.: Influence functions in deep learning are fragile. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
6. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10915–10924. IEEE (2022)
7. Chen, Y., Xu, X., de Hoog, F., Liu, J., Wang, S.: Medium-difficulty samples constitute smoothed decision boundary for knowledge distillation on pruned datasets. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025)
8. Cho, Y., Shin, B., Kang, C., Yun, C.: Lightweight dataset pruning without full training via example difficulty and prediction uncertainty. In: Proceedings of the 42th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, PMLR (2025)
9. Choi, H., Ki, N., Chung, H.W.: BWS: best window selection based on sample scores for data pruning across broad ranges. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024)
10. Cook, R.D., Weisberg, S.: Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics **22**(4), 495–508 (1980)
11. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 248–255. IEEE Computer Society (2009)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
13. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
14. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International journal of computer vision **129**(6), 1789–1819 (2021)
15. Hampel, F.R.: The influence curve and its role in robust estimation. Journal of the american statistical association **69**(346), 383–393 (1974)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016)
17. He, M., Yang, S., Huang, T., Zhao, B.: Large-scale dataset pruning with dynamic uncertainty. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024. pp. 7713–7722. IEEE (2024)
18. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015)
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017)

20. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling BERT for natural language understanding. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. Findings of ACL, vol. EMNLP 2020, pp. 4163–4174. Association for Computational Linguistics (2020)
21. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 1885–1894. PMLR (2017)
22. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. pp. 1106–1114 (2012)
24. Kwon, Y., Wu, E., Wu, K., Zou, J.: Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
25. Liang, J., Li, L., Bing, Z., Zhao, B., Tang, Y., Lin, B., Fan, H.: Efficient one pass self-distillation with zipf's label smoothing. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XI. Lecture Notes in Computer Science, vol. 13671, pp. 104–119. Springer (2022)
26. Moser, B.B., Shanbhag, A.S., Frolov, S., Raue, F., Folz, J., Dengel, A.: A coreset selection of coreset selection literature: Introduction and recent advances (2025)
27. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 3967–3976. Computer Vision Foundation / IEEE (2019)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 8024–8035 (2019)
29. Paul, M., Ganguli, S., Dziugaite, G.K.: Deep learning on a data diet: Finding important examples early in training. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 20596–20607 (2021)
30. Phillips, J.M.: Coresets and sketches. In: Handbook of discrete and computational geometry, pp. 1269–1288. Chapman and Hall/CRC (2017)
31. Pleiss, G., Zhang, T., Elenberg, E.R., Weinberger, K.Q.: Identifying mislabeled data using the area under the margin ranking. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing

Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
32. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
33. Shen, Z., Xing, E.P.: A fast knowledge distillation framework for visual recognition. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV. Lecture Notes in Computer Science, vol. 13684, pp. 673–690. Springer (2022)
34. Toneva, M., Sordoni, A., des Combes, R.T., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019)
35. Yang, Z., Zeng, A., Li, Z., Zhang, T., Yuan, C., Li, Y.: From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 17139–17148. IEEE (2023)
36. Ye, X., Wu, Y., Zhang, W., Jin, C., Chen, Y.: Towards robust influence functions with flat validation minima. In: Proceedings of the 42th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, PMLR (2025)
37. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: British Machine Vision Conference 2016. British Machine Vision Association (2016)
38. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017)
39. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)
40. Zheng, H., Liu, R., Lai, F., Prakash, A.: Coverage-centric coreset selection for high pruning rates. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023)
41. Zhou, X., Pi, R., Zhang, W., Lin, Y., Chen, Z., Zhang, T.: Probabilistic bilevel coreset selection. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 27287–27302. PMLR (2022)