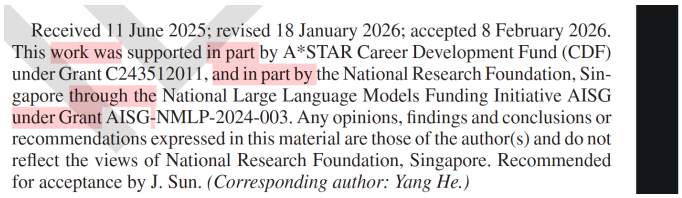
Author Response:

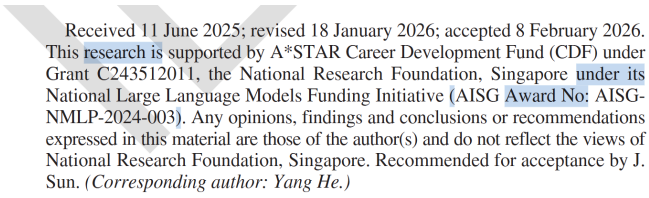
Please revise the funding content to match the following statement:

This research is supported by A*STAR Career Development Fund (CDF) under Grant C243512011, the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.



Received 11 June 2025; revised 18 January 2026; accepted 8 February 2026. This work was supported in part by A*STAR Career Development Fund (CDF) under Grant C243512011, and in part by the National Research Foundation, Singapore through the National Large Language Models Funding Initiative AISG under Grant AISG-NMLP-2024-003. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Recommended for acceptance by J. Sun. *(Corresponding author: Yang He.)*

Left: Current Funding Information



Received 11 June 2025; revised 18 January 2026; accepted 8 February 2026. This research is supported by A*STAR Career Development Fund (CDF) under Grant C243512011, the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Recommended for acceptance by J. Sun. *(Corresponding author: Yang He.)*

Right: Expected Funding Information

Soft Label Pruning and Quantization for Large-Scale Dataset Distillation

Lingao Xiao and Yang He, *Member, IEEE*

***Abstract*—Large-scale dataset distillation requires storing auxiliary soft labels that can be 30-40× (ImageNet-1K) or 200× (ImageNet-21K) larger than the condensed images, undermining the goal of dataset compression. We identify two fundamental issues necessitating such extensive labels: (1) *insufficient image diversity*, where high within-class similarity in synthetic images requires extensive augmentation, and (2) *insufficient supervision diversity*, where limited variety in supervisory signals during training leads to performance degradation at high compression rates. To address these challenges, we propose Label Pruning and Quantization for Large-scale Distillation (LPQLD). We enhance *image diversity* via class-wise batching and BN supervision during synthesis. For *supervision diversity*, we introduce Label Pruning with Dynamic Knowledge Reuse to enhance *label-per-augmentation diversity*, and Label Quantization with Calibrated Student-Teacher Alignment to enhance *augmentation-per-image diversity*. Our approach reduces soft label storage by 78× on ImageNet-1K and 500× on ImageNet-21K while improving accuracy by up to 7.2% and 2.8%, respectively. Extensive experiments validate the superiority of LPQLD across different network architectures and other dataset distillation methods.**

***Index Terms*—Dataset distillation, dataset condensation, data compression, efficient learning.**

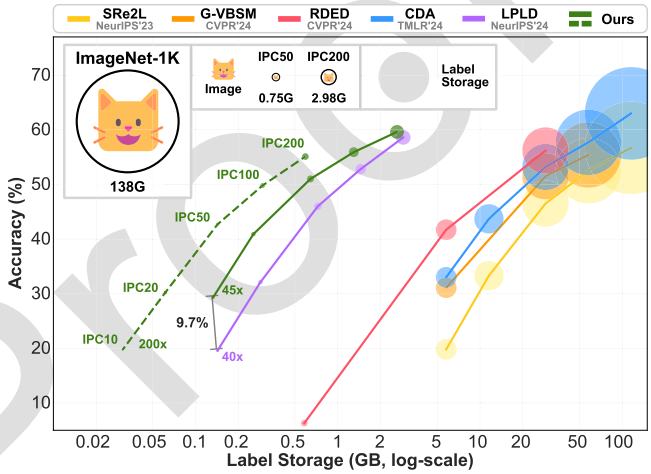


Fig. 1. The relationship between performance and label storage needed for ImageNet-1 K distillation. Our method creates a new Pareto front for the performance-storage trade-off, having **fewer soft labels than images**. Baselines [6], [7], [8], [9], [11] are shown for comparison.

I. Introduction

WE ARE advancing into the era of **ImageNet-level dataset condensation**, aiming to compress large datasets into much smaller, efficient counterparts [1], [2], [3], [4], [5]. While prior works struggled to scale due to memory constraints, recent methods [6], [7] employing a three-phase approach (squeeze, recover, relabel) have shown promise. However, a critical bottleneck emerges in the 'relabel' phase: the auxiliary data, primarily **soft labels** generated under numerous augmentations across training epochs [6], [7], [8], [9], [10], demand enormous storage as shown in Fig. 1. For ImageNet-1 K, this auxiliary storage can be **30-40**× larger than the condensed images themselves, and for ImageNet-21K, this ratio balloons to over **200**×, severely undermining the goal of dataset compression and making large-scale distillation impractical.

We identify two fundamental issues that necessitate such extensive soft label storage. The first is **insufficient image diversity**, where synthesized images exhibit high within-class similarity (as observed in SRe^{2}L [7]). We attribute this to the common practice of constructing batches with samples from different classes to match global Batch Normalization (BN) statistics during image synthesis (the 'recover' phase); this is confirmed by Feature Cosine Similarity and MMD (Section III-B). The second is **insufficient supervision diversity**, where repeated supervision with the same soft labels across epochs limits the variety of supervisory signals and causes severe performance degradation (Fig. 3). This is the core problem of LPLD [11] at high compression rates. Notably, these two issues are interconnected: low image diversity drives heavier augmentation during student training, which in turn demands more diverse supervision to maintain performance.

To address these dual challenges, we propose **Label Pruning and Quantization for Large-scale Distillation (LPQLD)**, a framework that systematically tackles both image diversity and supervision diversity issues.

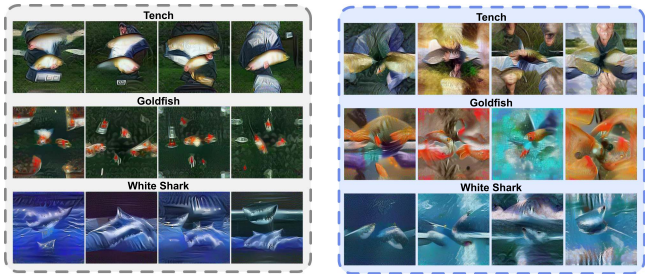
Received 11 June 2025; revised 18 January 2026; accepted 8 February 2026. This work was supported in part by A*STAR Career Development Fund (CDF) under Grant C243512011, and in part by the National Research Foundation, Singapore through the National Large Language Models Funding Initiative AISG under Grant AISG-NMLP-2024-003. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Recommended for acceptance by J. Sun. *(Corresponding author: Yang He.)*

The authors are with the CFAR, Agency for Science, Technology and Research, Singapore 138632, also with the IHPC, Agency for Science, Technology and Research, Singapore 138632, and also with the National University of Singapore, Singapore 119077 (e-mail: xiao_lingao@outlook.com; hyhy1992@gmail.com).

Code is available at https://github.com/he-y/soft-label-pruning-quantization-for-dataset-distillation.

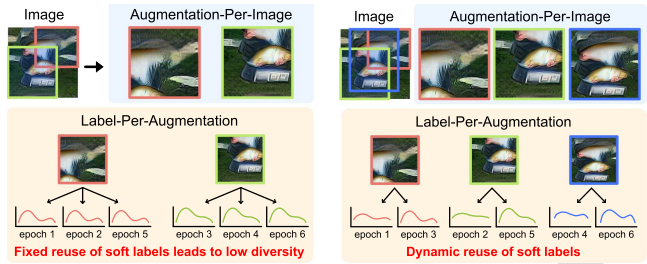
This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2026.3664488, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2026.3664488



(a) SRe^{2}L [6]. (b) Ours (IPC=10).

Fig. 2. Visual comparison of synthesized images (IPC=10) for ImageNet-1 K classes. Our method exhibits noticeably higher **within-class visual diversity**.



(a) LPLD [11]. (b) Ours.

Fig. 3. Illustration of supervision diversity of LPLD and LPQLD (Ours) under the same storage and training budget.

TABLE I
COMPARISON BETWEEN LPLD AND LPQLD

Category	Feature	LPLD	LPQLD
Image Diversity	**Class-wise Batching & Supervision** Enhances within-class diversity	✓	✓
Label Efficiency	**Label Pruning** Reduces number of augmentation pairs	✓	✓
	Label Quantization (Top-k) Compresses individual label size	✗	✓
Supervision Diversity	**Dynamic Knowledge Reuse (DKR)** Extracts diverse signals from pruned labels	✗	✓
	Calibrated Student-Teacher Alignment (CA) Aligns student with quantized teacher	✗	✓

TABLE II
NECESSITY OF LPQLD FRAMEWORK. P: PRUNING, Q: QUANTIZATION, DKR: DYNAMIC KNOWLEDGE REUSE, CA: CALIBRATED ALIGNMENT.

Method	Components	Bottleneck
Baseline (e.g., SRe^{2}L)	Full Soft Labels	**Raw Data Size** Label Storage > Image Storage by 200×
Single Technique	**P** *or* **Q**	**Auxiliary Data Dominance** Indices/Augmented Info. dominate total storage
Naive Combine	P + Q	**Accuracy Collapse** Sparse supervision & mismatched distributions
LPQLD (Ours)	P (+**DKR**) + Q (+**CA**)	**Both Resolved** DKR fixes sparsity, CA fixes mismatch

Addressing Image Diversity: Following LPLD [11], we enhance image diversity during synthesis by batching images within the same class and introducing class-wise BN supervision, encouraging images to collectively capture diverse class characteristics. As illustrated in Fig. 2(b) and quantified in Table III, this approach yields synthetic images with lower within-class similarity and better representation of the original dataset (Fig. 6).

Addressing Supervision Diversity: We tackle supervision diversity from two complementary angles: label-per-augmentation diversity and augmentation-per-image diversity.

a) Label-per-Augmentation Diversity via Dynamic Knowledge Reuse: At high compression rates with label pruning, each augmentation is repeatedly supervised by the same soft label across training epochs, severely limiting supervision diversity. To address this, we propose Dynamic Knowledge Reuse (DKR), a theoretically-grounded strategy employing temperature annealing to extract different supervisory signals from the same label across training epochs (Section III-D). This enables each retained label to provide varied supervision throughout training without additional storage overhead.

b) Augmentation-per-Image Diversity via Label Quantization with Calibrated Alignment: Relying solely on label pruning drastically reduces augmentations per image, potentially limiting each image to only one augmentation. We introduce Label Quantization (Q), which stores only top-k pre-softmax logits per label to enable more augmentation-label pairs under the same storage budget. However, quantization alters the teacher's probability distribution. Our Calibrated Student-Teacher Alignment (CA) addresses this by dynamically adjusting student distribution via grid-search optimization (Section III-E).

c) Balancing the Trade-off: With a fixed storage budget, increasing augmentation-per-image diversity (via quantization) necessarily reduces the number of labels available per augmentation (via pruning). We provide a comprehensive Pareto frontier analysis to identify optimal pruning-quantization configurations that balance these two types of supervision diversity for different dataset scales (Section IV-D), offering practical guidance for deploying LPQLD across diverse scenarios.

Table I summarizes the key differences between LPLD and LPQLD across three categories, highlighting LPQLD's innovations in label efficiency (quantization) and supervision diversity (DKR and CA). Table II shows why these components are essential solutions to fundamental bottlenecks. The baseline suffers from impractical label storage (200× larger than images) and single techniques (**P** or **Q**) hit auxiliary data dominance. Naively combining them (P + Q) is possible, but it leads to accuracy collapse due to sparse supervision and mismatched distributions. Our LPQLD resolves both bottlenecks: P (+**DKR**) addresses sparsity by extracting varied supervisory signals via temperature annealing, while Q (+**CA**) fixes distribution mismatch by dynamically adjusting student temperature. This establishes LPQLD as a systematic framework where each component addresses specific bottlenecks induced by the interaction of pruning and quantization.

The contributions of this work can be summarized as follows:

1) We identify **supervision diversity** as the fundamental bottleneck at high compression rates and provide a unified LPQLD framework for soft label compression.
2) We decompose **supervision diversity** into **augmentation-per-image diversity** and **label-per-augmentation diversity**, and propose two modules

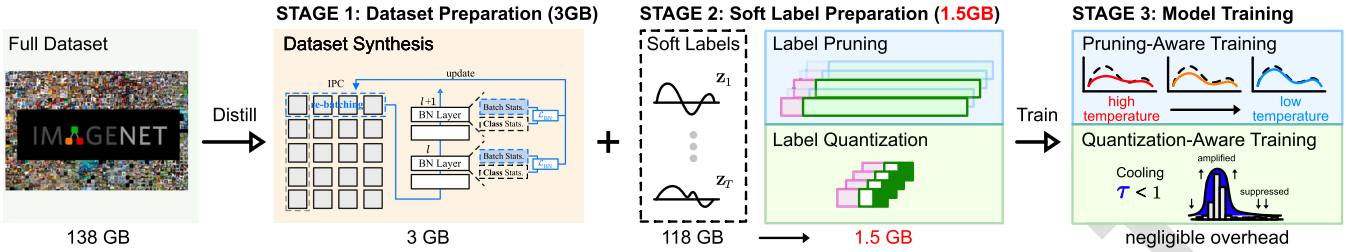


Fig. 4. Pipeline overview of LPQLD. In STAGE 1, we generate a synthetic dataset by matching class-wise BN statistics. In STAGE 2, we generate corresponding soft labels. To reduce soft label storage, label pruning and label quantization are conducted. In STAGE 3, compression-aware model training is performed with distilled datasets and compressed soft labels.

combined with Pareto front analysis to enhance both types of diversity.

3) LPQLD achieves unprecedented storage efficiency (**78**× on ImageNet-1K, **500**× on ImageNet-21K) while maintaining or exceeding LPLD's performance across diverse architectures and distillation methods.

II. Related Works

Large-Scale Dataset Distillation: Recent advancements in dataset distillation have pushed towards handling **large-scale datasets** like ImageNet-1K [12] and beyond. While early methods struggled with the memory demands and scale [3], [13], techniques like TESLA [14] and particularly multi-stage methods such as SRe^{2}L [6] have enabled distillation of full ImageNet-1K. SRe^{2}L notably decouples the process into squeezing, recovering (image synthesis), and relabeling phases. Consequently, significant research effort has concentrated on improving the **quality of the synthesized images** during the recovery phase, exploring various strategies to generate more informative condensed datasets including through optimization [7], [8], [9], [10], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] or through generative methods [25], [26], [27]. To attain decent performance, the **relabeling phase**, responsible for generating extensive soft labels from a teacher model under numerous augmentations, plays a crucial role. However, soft labels have received comparatively less attention regarding their efficiency [28], [29]. The relabeling phase, while crucial for achieving state-of-the-art performance, introduces a significant **storage bottleneck** due to the massive volume of soft label data required, diverging from the goal of dataset distillation.

Soft Label Compression: Despite the extensive storage requirements of soft labels in methods like SRe^{2}L [6], directly addressing this **label storage bloat** has been largely overlooked in dataset distillation research. Some works have investigated how to better leverage existing soft labels during student training [30], [31], but without focusing on reducing their storage footprint, and in some cases even increasing it [32]. Teddy [33] proposes using a proxy model for dynamic label generation, which eliminates pre-computation storage but introduces efficiency trade-offs and additional training overhead. FKD [34], though designed for traditional knowledge distillation, employs label quantization by storing both indices and quantized post-softmax logits for recovery during training. However, it has limitations: (1) it was designed for full datasets rather than distilled ones, creating unique challenges in large-scale dataset distillation where limited image supervision is available, and (2) it does not address the storage requirements of auxiliary augmentation information. In this paper, we follow the naming of "label quantization" to be consistent with previous works [11], [34].

III. Method

In this section, we present our Label Pruning and Quantization for Large-scale Distillation (LPQLD) approach, which addresses the soft label storage bottleneck while maintaining or improving performance. As illustrated in Fig. 4, our method consists of three key components: (1) diverse synthetic dataset generation to enhance *image diversity*, (2) soft label compression via pruning and quantization to enhance *supervision diversity*, and (3) compression-aware training with dynamic knowledge reuse and calibrated student-teacher alignment. This integrated pipeline effectively reduces storage requirements while preserving the knowledge contained in soft labels.

A. Preliminary

Synthetic Dataset Generation: We begin by considering the standard Batch Normalization (BN) transformation:

$$y = \gamma \left(\frac{\boldsymbol{x} - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta, \tag{1}$$

where $\boldsymbol{x}$ represents the input feature maps to the BN layer, y is the normalized output, γ and β are learnable scaling and shifting parameters, while μ and σ^2 represent the mean and variance of the inputs. The constant ϵ is introduced for numerical stability to avoid division by zero. During training, running statistics (mean and variance) are maintained and later used during inference to approximate the unavailable true statistics of testing data.

Inspired by SRe^{2}L [7] and DeepInversion [35], we generate synthetic imagery by aligning the BN statistics of intermediary layers. In particular, the target is to minimize the discrepancy between the BN statistics computed from the synthetic images $\widetilde{\boldsymbol{x}}$ and the stored running estimates from the full dataset $\mathcal{T}$:

$$\mathcal{L}_{\text{BN}}(\widetilde{\boldsymbol{x}}) = \sum_l \left\| \mu_l(\widetilde{\boldsymbol{x}}) - \mathbb{E}\left(\mu_l \mid \mathcal{T}\right) \right\|_2 + \sum_l \left\| \sigma_l^2(\widetilde{\boldsymbol{x}}) - \mathbb{E}\left(\sigma_l^2 \mid \mathcal{T}\right) \right\|_2$$

$$\approx \sum_l \left\|\mu_l(\widetilde{\boldsymbol{x}}) - \mathbf{BN}_l^{\mathrm{RM}}\right\|_2 + \sum_l \left\|\sigma_l^2(\widetilde{\boldsymbol{x}}) - \mathbf{BN}_l^{\mathrm{RV}}\right\|_2, \tag{2}$$

where $\mathbf{BN}_l^{\mathrm{RM}}$ and $\mathbf{BN}_l^{\mathrm{RV}}$ denote the running mean and variance that serve as surrogates for the expected values. Here, $\mu_l(\widetilde{\boldsymbol{x}})$ and $\sigma_l^2(\widetilde{\boldsymbol{x}})$ are the per-layer statistics computed on a batch of synthetic images $\widetilde{\boldsymbol{x}}$.

This BN loss acts as a regularizer to the classification loss $\mathcal{L}_{\mathrm{CE}}$, yielding the joint objective:

$$\arg\min_{\widetilde{\boldsymbol{x}}} \quad \underbrace{\ell\left(\boldsymbol{\theta}_{\mathcal{T}}\left(\widetilde{\boldsymbol{x}}\right), \boldsymbol{y}\right)}_{\mathcal{L}_{\mathrm{CE}}} + \alpha \cdot \mathcal{L}_{\mathrm{BN}}\left(\widetilde{\boldsymbol{x}}\right), \tag{3}$$

where $\boldsymbol{\theta}_{\mathcal{T}}$ is the model pretrained on the original dataset and the scalar α adjusts the influence of the BN regularization.

Image Relabeling: In model distillation, huge teachers impose additional training burdens to computing demands on top of the existing two bottlenecks, deep network propagation and data loading [34]. FKD [34] introduces the relabel phase that precomputes and stores hyper-parameters of augmentations $\{\mathcal{F}\}$ together with soft label sets $\{\mathcal{P}\}$, where $\{\mathcal{P}\} = \mathrm{softmax}(\mathbf{z}, \tau)$ represents probability distributions derived from the logit outputs $\mathbf{z}$ of the teacher model through softmax with temperature τ. The storage size of augmentation hyper-parameters $|\{\mathcal{F}\}|$ is negligible compared to the size of full soft labels $|\{\mathcal{P}\}|$. During the model training phase, the stored label files are loaded; images undergo the same augmentation by applying $\{\mathcal{F}\}$, making the supervision from $\{\mathcal{P}\}$ valid. To efficiently store the soft labels, label quantization is applied, and merely top-k (e.g., $k = 5$) soft labels are selected for each image per augmentation. Note that both indices and values are stored, so the actual storage for top-k is equivalent to $2 \times k$. For the ImageNet-1 K dataset, this cuts down the soft label storage $\{\mathcal{P}\}$ to 100× by taking top-5 predictions.

SRe^{2}L [6] adopts the idea of relabeling process [34] into dataset distillation, and it is the first to achieve effective large-scale (ImageNet-1K-scale) dataset distillation. However, due to the large discrepancy in the dataset size between full ImageNet-1 K and distilled ImageNet-1 K, the label quantization process does not effectively apply. The full ImageNet-1 K has over 1000 K images to accommodate information loss from each soft label, while distilled ImageNet-1 K has only 10 K images (i.e., 10 images per class). As a consequence, by giving up the label quantization, soft label size $|\{\mathcal{P}\}|$ is about 40× of the distilled dataset size $|\mathcal{S}|$.

B. Image Diversity Via Diverse Synthetic Dataset Generation

The excessive storage requirements for soft labels in current dataset distillation methods fundamentally contradict the core motivation of dataset distillation—creating compact, efficient dataset representations. We hypothesize that a key underlying factor is *insufficient image diversity* within synthetic image classes, which necessitates extensive augmentation and corresponding soft labels to provide varied supervision. Before introducing our compression techniques, we first establish metrics to quantify this diversity problem and then propose methods to address it at the image synthesis stage.

1) Intra-Class Diversity. Feature Cosine Similarity: A key measure of diversity is the similarity of images within the same class. We quantify this by calculating the cosine similarity of their feature representations. Lower cosine similarity indicates that the images are less alike, thus the intra-class diversity is higher. This observation is formally stated as:

Proposition 1: A reduction in cosine similarity among features of images within the same class reflects increased diversity.

Mathematically, the cosine similarity between the features of two images is expressed as:

$$\text{cos similarity} := \frac{\sum_{i=1}^{n} f(\widetilde{\boldsymbol{x}}_{\boldsymbol{c},\boldsymbol{i}})\, f(\widetilde{\boldsymbol{x}}'_{c,i})}{\sqrt{\sum_{i=1}^{n} f(\widetilde{\boldsymbol{x}}_{\boldsymbol{c},\boldsymbol{i}})^2}\sqrt{\sum_{i=1}^{n} f(\widetilde{\boldsymbol{x}}'_{c,i})^2}}, \tag{4}$$

where $\widetilde{\boldsymbol{x}}_{\boldsymbol{c}}$ and $\widetilde{\boldsymbol{x}}'_{c}$ are two images from class c, $f(\cdot)$ is the feature extractor, and n is the dimensionality of the feature vector.

2) Inter-Dataset Consistency. Maximum Mean Discrepancy (MMD): While intra-class differences are important, it is equally crucial that the synthetic dataset faithfully captures the original data distribution. We assess this through Maximum Mean Discrepancy (MMD), which compares the feature distributions of the synthetic and real datasets:

Proposition 2: A lower MMD value indicates that the synthetic dataset closely approximates the original dataset's range of features, thus reflecting higher diversity.

The empirical estimation of MMD is defined as [36], [37]:

$$\mathrm{MMD}^2\left(\boldsymbol{P}_{\mathcal{T}}, \boldsymbol{P}_{\mathcal{S}}\right) = \hat{\mathcal{K}}_{\mathcal{T},\mathcal{T}} + \hat{\mathcal{K}}_{\mathcal{S},\mathcal{S}} - 2\hat{\mathcal{K}}_{\mathcal{T},\mathcal{S}}, \tag{5}$$

$$\hat{\mathcal{K}}_{X,Y} = \frac{1}{|X| \cdot |Y|} \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} \mathcal{K}\left(f\left(x_i\right), f\left(y_j\right)\right), \tag{6}$$

where $\{x_i\}_{i=1}^{|X|} \sim X$, $\{y_j\}_{j=1}^{|Y|} \sim Y$, $\mathcal{T}$ and $\mathcal{S}$ denote the real and synthetic datasets, $\mathcal{K}$ is a reproducing kernel (e.g., Gaussian), and $f(\cdot)$ maps inputs to their feature representations.

3) Diversifying the Datasets by Class-Wise Supervision: In previous objective (3), a subset of classes $\mathcal{B}_c$ is sampled to match global BN statistics, leading to independent generation for each image of a class and, consequently, lower intra-class diversity. Drawing inspiration from He et al. [38], we argue that images in the same class can benefit from working cooperatively when generating multiple images per class.

Step 1. Intra-Class Re-batching: To promote image collaboration, we sample multiple images from the same class and apply class-specific supervision [4], [13]. This change, illustrated in Fig. 5, ensures that images are updated together rather than in isolation.

Step 2. Incorporating Class-wise BN Supervision: Since the global running statistics may not be optimal for class-specific information, we propose maintaining separate BN statistics for each class. The additional storage cost is minimal even for large-scale datasets (e.g., ImageNet-1 K; see Appendix C-A, available online).

Step 3. Designing the Class-wise Objective: We adjust the original objective by introducing two losses. First, we compute the classification (cross-entropy) loss using global BN statistics

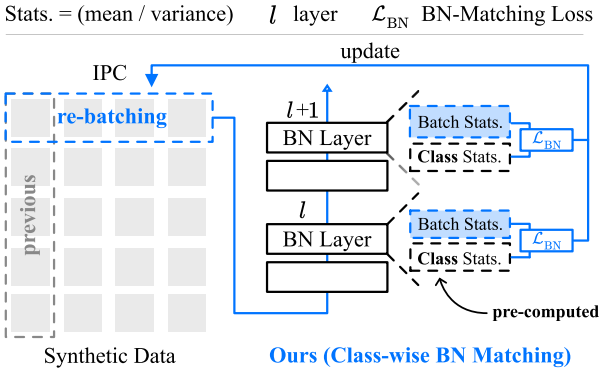


Fig. 5. Illustration of our diverse synthetic dataset generation. Our method encourages collaboration among images of the same class, yielding a natural separation between classes. Here, the classification loss is omitted for illustration.

TABLE III
THE COSINE SIMILARITY BETWEEN IMAGE FEATURES. IT IS THE AVERAGE OF 1 K CLASS ON THE SYNTHETIC IMAGENET-1 K DATASET. FEATURES ARE EXTRACTED USING PRETRAINED RESNET-18.

IPC	SRe^{2}L	CDA	Ours	Full Dataset
50	0.841 ± 0.023	0.816 ± 0.026	0.796 ± 0.029	0.695 ± 0.045
100	0.840 ± 0.016	0.814 ± 0.019	0.794 ± 0.021	
200	0.839 ± 0.011	0.813 ± 0.013	0.793 ± 0.015	

to ensure overall effective supervision. Second, the BN loss is now defined by comparing the synthetic images' BN statistics with the class-specific running statistics. The updated class-wise objective is:

$$\arg\min_{\tilde{\boldsymbol{x}}_c} \overbrace{\left(-\sum_{i=1}^{N} y_{c,i} \log \left(\text{softmax}\left(\boldsymbol{\theta}_{\mathcal{T}}\left(\frac{\tilde{\boldsymbol{x}}_{c,i}-\mathbf{BN}^{\text{RM}}_{\text{global}}}{\sqrt{\mathbf{BN}^{\text{RV}}_{\text{global}}+\epsilon}}\right)\right)_c\right)}^{\text{Cross-Entropy Loss with Global BN}} + \alpha \cdot \underbrace{\sum_{l}\left(\left\|\mu_l(\tilde{\boldsymbol{x}}_c)-\mathbf{BN}^{\text{RM}}_{l,c}\right\|_2+\left\|\sigma_l^2(\tilde{\boldsymbol{x}}_c)-\mathbf{BN}^{\text{RV}}_{l,c}\right\|_2\right)}_{\text{BN Loss with Class-wise BN}}. \tag{7}$$

Note that while the BN loss utilizes class-specific statistics, the computation of logits for the cross-entropy loss still relies on global statistics. This design choice is crucial because adjusting only μ and σ without refining γ and β could degrade model performance.

Superior Performance in Similarity Metrics: Thanks to the adjustments described in Section III-B3, our synthetic dataset demonstrates both lower intra-class feature cosine similarity (as evidenced in Table III) and a much reduced MMD (see Fig. 6) compared to prior methods. These improvements indicate enhanced *image diversity*, where the synthetic images capture a broader and more representative feature distribution of the original dataset. With this diverse synthetic dataset established, we next address the issue of superfluous soft labels.

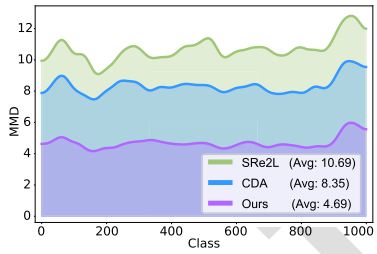


Fig. 6. Comparison of Maximum Mean Discrepancy (MMD).

C. Soft Label Pruning and Quantization

Label Pruning: As illustrated in Fig. 7 (blue part), label pruning prunes both augmentation hyper-parameters $\{\mathcal{F}\}$ and soft labels $\{\mathcal{P}\}$. For a model distillation training of $T = 300$ epochs, the theoretical maximum pruning rate would be $300\times$ to ensure each image undergoes at least a single augmentation. Benefiting from the enhanced *image diversity* that implicitly reduces the number of different augmentations needed, we adopt random label pruning. Compared to metric-based label pruning that requires selection from T epochs, random pruning enjoys acceleration in the label generation phase. In particular, we generate labels for $(1-p)\times T$ epochs, meaning if there are 300 epochs T of training in total and pruning rate p is 0.9, 3 epochs of model propagation are required and actual $10\times$ speed-up is achieved. The soft label generation process can be time-consuming especially for limited CPU resources and a low-speed disk since extensive memory movement and system IO operations are executed.

Label Quantization: Orthogonal to label pruning, label quantization further reduces storage requirements to enhance *augmentation-per-image diversity*. FKD [34] introduces different label quantization approaches, including Hardening, Smoothing, Marginal Smoothing (MS), and Marginal Renormalization (MR). Our method stores the set of top-k pre-softmax logits $\{\mathbf{z}\}$ to enable dynamic temperature scheduling for enhancing *label-per-augmentation diversity*. By incorporating the top-k selection, the post-softmax distribution $\mathcal{P}_i(\mathbf{z}_i^k, \mathcal{F}_i, \tau)$ is defined as:

$$\begin{aligned}\mathcal{P}_i(\mathbf{z}_i^k, \mathcal{F}_i, \tau) &= \text{softmax}(\mathbf{z}_i^k, \tau), \quad \text{where} \\ \mathbf{z}_i^k &= \text{topk}(\theta_{\mathcal{T}}(\mathcal{A}(x_i, \mathcal{F}_i)), k).\end{aligned} \tag{8}$$

D. Supervision Diversity 1 Via Dynamic Knowledge Reuse

After label pruning, the remaining labels must be reused across multiple training epochs to provide sufficient supervision throughout the entire training process. Simply repeating the same supervision signals is possible but limits *label-per-augmentation diversity*. The challenge is maximizing knowledge transfer from this reduced set of labels.

Dynamic Knowledge Reuse: We propose Dynamic Knowledge Reuse, which leverages temperature annealing to generate diverse supervision signals from the same pruned labels, enhancing *label-per-augmentation diversity*. Temperature in softmax controls the entropy of probability distributions, higher

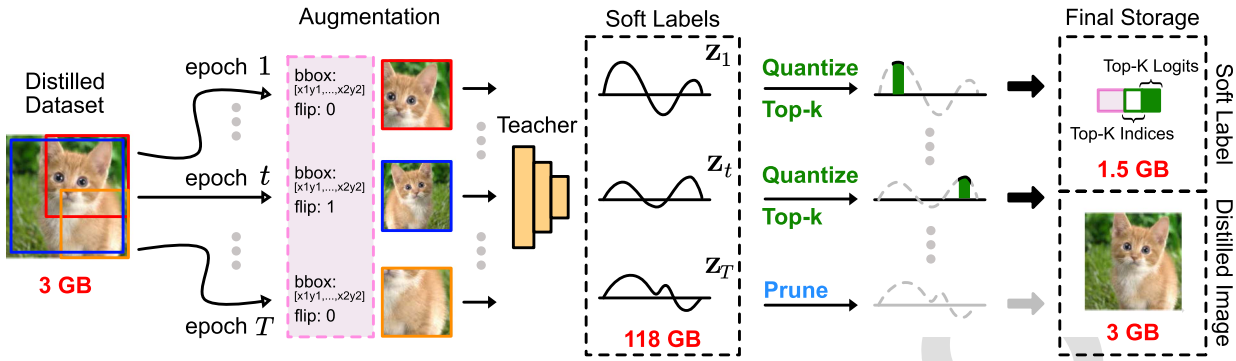


Fig. 7. Illustration of the soft label generation process incorporating label pruning and label quantization techniques. The process begins by augmenting each dataset image for T epochs, with each augmented image producing different outputs from a fixed pretrained teacher. Traditional approaches store all augmentations information and soft labels across all T epochs. **Label pruning** retains only $(1-p)\times T$ epochs out of the total T epochs, where $p<1$ represents the pruning rate. **Label quantization**, an orthogonal process, reduces the storage by selecting only the Top-k logits $\mathbf{z}$ to store. Note that we store the *before softmax* output logits.

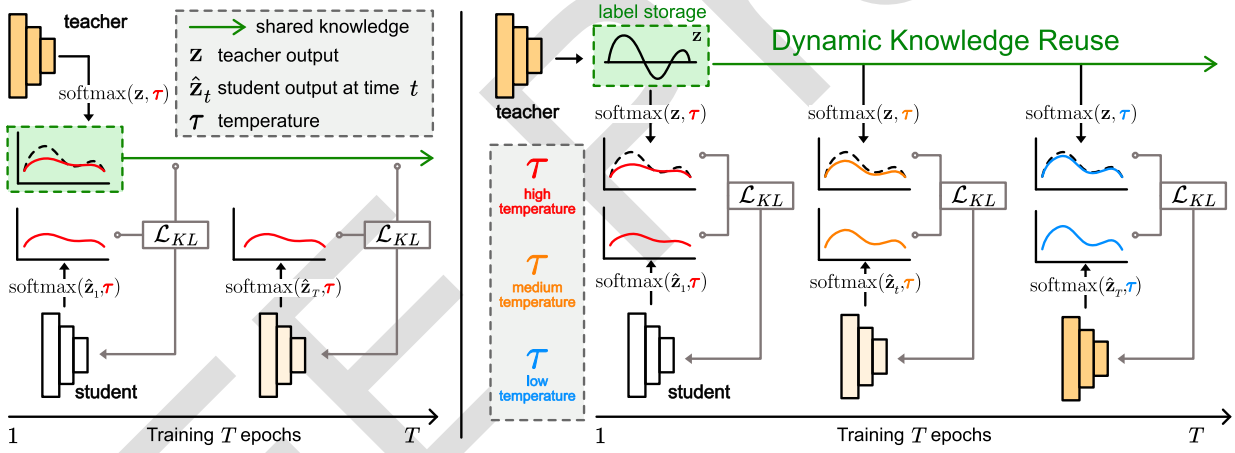


Fig. 8. Illustration of knowledge reuse due to pruning that leads to insufficient amount of teacher information for all T epochs of model distillation training. (**Left**) Naive Knowledge Reuse: despite the training stages, the supervision from teachers is fixed for a specific pre-stored augmentation. (**Right**) Dynamic Knowledge Reuse: to allow the same augmentation to have diverse supervision signals, we set a different temperature during *softmax* operation for the same augmentation at different training epochs. In particular, we use an annealing temperature scheduler. Note: *before softmax* output logits are stored, and such dynamic temperature assignment does not introduce additional storage costs.

temperatures produce smoother distributions that reveal more class similarities, while lower temperatures create sharper distributions focusing on likely classes.

The key insight is that by varying temperature over time, we can create a curriculum learning effect that enhances *label-per-augmentation diversity:* starting with higher temperatures to learn broad class relationships, then gradually transitioning to lower temperatures to refine specific class boundaries.

Storage-Efficient Implementation: Naively implementing temperature annealing would require storing probability distributions for each temperature, which is prohibitively expensive. Instead, we store only the pre-softmax logits $\mathbf{z}_i = \theta_{\mathcal{T}}(\mathcal{A}(x_i,\mathcal{F}_i))$ once and dynamically compute the probability distribution for any temperature during training as shown in Fig. 8:

$$p_i^{(\tau_t)} = \frac{\exp(z_i/\tau_t)}{\sum_{j=1}^{C}\exp(z_j/\tau_t)}. \tag{9}$$

This approach provides significant storage savings while enabling exact recovery of any temperature-specific distribution for *label-per-augmentation diversity*. For a C-class problem, we store only C values per image instead of C values for each of the many temperature settings we would use throughout training.

Theoretical Justification for Label-Pruned Temperature Annealing: Annealing KD [39] established that temperature annealing creates a beneficial learning curriculum by leveraging the VC-dimension theory [40], [41]. The key theoretical advantage they proved was that temperature annealing leads to better convergence compared to standard KD. This can be expressed through the inequality [39], [42]:

$$R_\tau(f_s^\tau) - R_\tau(f_t^\tau) \le O\left(\frac{|\mathcal{H}_s|_c}{N^{\alpha_{st}^\tau}}\right) + \varepsilon_{st}^\tau \le O\left(\frac{|\mathcal{H}_s|_c}{N^{\alpha_{st}}}\right) + \varepsilon_{st} \tag{10}$$

where $R(\cdot)$ is the expected error, f_s^τ and f_t^τ represent student and teacher functions at temperature τ, and $\alpha_{st}^\tau \ge \alpha_{st}$ indicates better convergence rates with annealing.

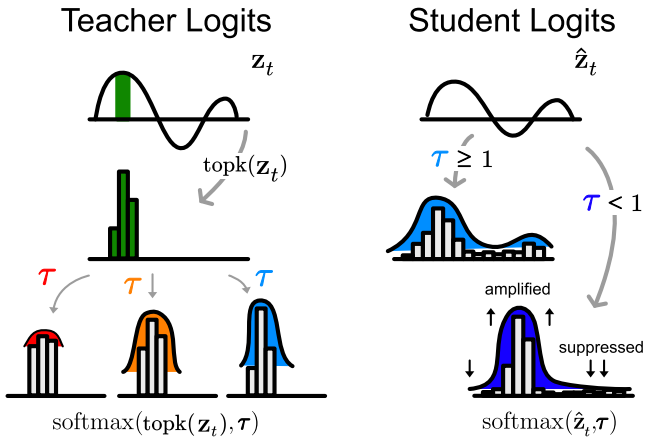


Fig. 9. Illustration of distribution-aware temperature scaling for student networks. (**Left**) Soft label distribution of teacher after *softmax*. *Softmax* applies to Top-k selected logits, and remaining logits assign a probability of 0. (**Right**) Assigning students a 'colder' temperature ($\tau < 1$) better aligns with the soft label distribution of the teacher.

In our label-pruned setting where we reuse a fraction λ of teacher predictions, we can adapt this inequality to account for repeated supervision:

$$R_\tau(f_s^\tau) - R_\tau(f_t^\tau) \leq O\left(\frac{|\mathcal{H}_s|_c}{(k(\tau)\cdot|\mathcal{D}_t^\lambda|)^{\alpha_{st}^\tau}}\right) + \varepsilon_{st}^\tau \qquad (11)$$

where $\mathcal{D}_t^\lambda$ is our pruned subset with pruning ratio λ and $k(\tau)$ represents the effective number of times samples are reused up to temperature τ. This adapted bound thus suggests that the benefits of annealing are preserved, as the effective sample reuse $k(\tau)$ helps to maintain a strong learning signal.

The benefits of Annealing KD remain valid in our label-pruned setting for two main reasons: **First**, the core mechanism of temperature annealing does not depend on having unique teacher predictions for each training instance. The same logits viewed through different temperatures still create the desired curriculum effect from smooth to sharp distributions, effectively creating *label-per-augmentation diversity* in the prediction space. **Second**, high *label-per-augmentation diversity* from repeatedly using the same predictions through different temperatures compensates for the reduced quantity of supervision. While we store only a fraction λ of the original teacher predictions, reusing them with an annealing temperature schedule allows the student to progressively learn more difficult aspects of the same predictions.

E. Supervision Diversity 2 Via Calibrated Alignment

Label quantization enables enhanced *augmentation-per-image diversity* by reducing per-label storage, allowing more augmentation-label pairs under the same storage budget. Unlike label pruning, quantization fundamentally changes the teacher's prediction distribution $\mathcal{P}$. By applying softmax on quantized logits $\mathbf{z}_i^k$, probabilities are re-normalized within the top-k selected classes, creating a sharper distribution with lower entropy than the unquantized distribution.

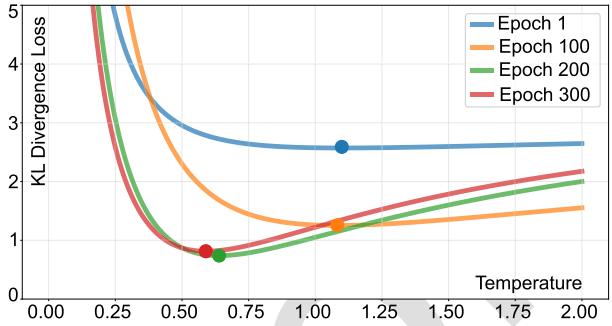


Fig. 10. Loss landscape of KL divergence showing a clear minimum within $\tau \in [0, 1]$.

Theoretical justification for $\hat{\tau} < 1$: For the student to match the quantized teacher distribution when minimizing KL divergence, the following ratio equality must hold for any two classes i, j in the top-k set (see Appendix A, available online for detailed proof):

$$\frac{\hat{\mathcal{P}}_i(\hat{\mathbf{z}}_i, \mathcal{F}_i, \hat{\tau})}{\hat{\mathcal{P}}_j(\hat{\mathbf{z}}_j, \mathcal{F}_j, \hat{\tau})} = \frac{\mathcal{P}_i(\mathbf{z}_i^k, \mathcal{F}_i, \tau_t)}{\mathcal{P}_j(\mathbf{z}_j^k, \mathcal{F}_j, \tau_t)} \qquad (12)$$

Expressing this in terms of softmax and taking logarithms:

$$\frac{\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j}{\hat{\tau}} = \log\left(\frac{\mathcal{P}_i(\mathbf{z}_i^k, \mathcal{F}_i, \tau_t)}{\mathcal{P}_j(\mathbf{z}_j^k, \mathcal{F}_j, \tau_t)}\right) \qquad (13)$$

Quantization produces a sharper distribution by concentrating probability mass on fewer classes. To match this sharper distribution under practical logit constraints, the temperature $\hat{\tau}$ must be below 1 to amplify existing logit differences (see Fig. 9). We determine the optimal temperature by minimizing KL divergence:

$$\arg\min_{\hat{\tau}^*} \sum_{i=1}^{n\in\mathcal{B}} \mathcal{P}_i(\mathbf{z}_i^k, \mathcal{F}_i, \tau_t) \log\left(\frac{\mathcal{P}_i(\mathbf{z}_i^k, \mathcal{F}_i, \tau_t)}{\hat{\mathcal{P}}_i(\hat{\mathbf{z}}_i, \mathcal{F}_i, \hat{\tau})}\right). \qquad (14)$$

Optimization via Grid Search: We solve this optimization using grid search over $\hat{\tau} \in [0, 1]$ because: (1) computing gradients through nested softmax operations is expensive, and (2) the KL divergence landscape is well-behaved with respect to τ (Fig. 10). We calculate the optimal temperature using the last batch of the tth epoch and apply it to the $t+1$ epoch, as outlined in Algorithm 1.

Computational Cost: The grid search incurs minimal overhead as it only recomputes softmax values and evaluates KL divergence under different temperatures. Importantly, it operates on a single batch (the last batch of each epoch) rather than the entire dataset, making the cost independent of dataset size. The total additional computational cost across all 300 training epochs is negligible (a few seconds in total).

Algorithm 1: Dynamic Student Temperature via Grid Search.

Input: Teacher logits $\mathcal{P}$, Student logits $\hat{\mathbf{z}}$
Initialize $\mathbb{T} \leftarrow \{0.01, 0.02, \ldots, 1.00\}$, $\mathcal{L}_{\min} \leftarrow \infty$
for $\hat{\tau} \in \mathbb{T}$ **do**
  $\hat{\mathcal{P}} \leftarrow \text{softmax}(\hat{\mathbf{z}}/\hat{\tau})$
  $\mathcal{L}_{\hat{\tau}} \leftarrow \text{KL}(\mathcal{P} \| \hat{\mathcal{P}})$
  if $\mathcal{L}_{\hat{\tau}} < \mathcal{L}_{\min}$ **then**
    $\mathcal{L}_{\min} \leftarrow \mathcal{L}_{\hat{\tau}}$, $\hat{\tau}^* \leftarrow \hat{\tau}$
Output: Optimal student temperature $\hat{\tau}^*$ for next epoch

IV. Experiments

A. Experiment Settings

Dataset details and training settings are provided in Appendix B, available online. Computing resources used for experiments can be found in Appendix C-B, available online.

Dataset: Our experiment results are evaluated on ImageNet-1K [12], and ImageNet-21K-P [43]. We follow the data preprocessing procedure of SRe^{2}L [6] and CDA [7].

Squeeze: We modify the pretrained model by adding class-wise BN running mean and running variance; since they are not involved in computing the BN statistics, they do not affect performance. As mentioned in Section III-B3, we compute class-wise BN statistics by training for one epoch with model parameters kept frozen.

Recover: We perform data synthesis following (7). The batch size for the recovery phase is the same as the IPC. Besides, we adhere to the original settings of SRe^{2}L while changing the pretrained teacher model to our squeezed models with class-wise BN statistics.

Relabel: We use pretrained ResNet18 [44] for all experiments as the relabel model except otherwise stated. For Tiny-ImageNet and ImageNet-1 K, we use the PyTorch pretrained model. For ImageNet-21K-P, we use the timm pretrained model. For DELT [20], due to the missing implementation of the relabeling process, we implement the Random Augmentation [45] under Timm's [46] implementation.

Label Pruning Setting: For label pruning, we exclude the last batch (usually with an incomplete batch size) of each epoch from the label pool. To not introduce any advantages from discarding the incomplete batch, we reduce the training of one batch from every epoch. To attain a $p\%$ pruning ratio, we generate and store the first $(1-p)\%$ of total batches. To implement label pruning effectively, we store additional image indices for every batch in a single file. The additional storage is accounted for within the total auxiliary information $\{\mathcal{F}\}$.

Label Quantization Setting: We store both indices and values of the Top-k output logits. During reconstruction of the full output logits, we set non-Top-k values to zero.

Validation: For LPLD and LPQLD, we adhere to the hyperparameter settings of CDA [7]. Unless otherwise stated, we follow their original hyperparameter settings for all methods.

B. Primary Results

ImageNet-1 K: Table IV compares the ImageNet-1 K pruning results with SOTA methods on ResNet18. Our method outperforms other SOTA methods at various pruning ratios and different IPCs. More importantly, our method consistently exceeds the unpruned version of SRe^{2}L with 78× less storage. This may not appear impressive at first glance; however, in terms of actual storage, the size is reduced from 29.7 G to 0.38 G for IPC50. In addition, we notice the performance at 10× (or 90%) pruning ratio degrades slightly, especially for large IPCs. For example, there is merely a 0.2% performance degradation on IPC200 using ResNet18. Based on this observation, we apply quantization at a 10× pruning rate, which means to achieve $\sim 80\times$ storage reduction, we take the sweet spot of 10× from pruning, and reduce the output logits by 8×. Following this approach, we notice a huge improvement over solely applying pruning. For instance, 7.2% (27.3% vs. 20.1%) accuracy improvements on IPC10 at 79× label compression rate. The improvement is consistent and is more prominent at small IPCs.

ImageNet-21K-P: ImageNet-21K-P has 10,450 classes, significantly increasing the disk storage as each soft label stores probabilities for 10,450 classes. Compared to ImageNet-1 K, ImageNet-21 K has a larger label-to-image ratio due to a larger number of classes. The IPC20 dataset leads to a 1.2 TB (i.e., 1285 GiB) label storage, making the existing framework less practical. However, with the help of our method, it can surpass SRe^{2}L [6] using 500× less label storage.

In addition, exhibited in ImageNet-1 K (Table IV), the Pareto fronts for optimal pruning and quantization ratio combinations change with the dataset size. With a larger IPC, label pruning delivers an improved performance. Such a behavior is more distinguished in ImageNet-21 K (Table V), and we have set 3 different settings at high compression rates. For IPC20, using a base pruning rate of 40× outperforms smaller base pruning rates in terms of both performance and storage.

Label Storage Breakdown: Table VI shows a storage gap between the theoretical and actual compression rates, highlighting the roles of label quantization and pruning. The theoretical rate only considers the output logits $\{\mathbf{z}\}$, while the actual rate must also include auxiliary data, particularly the augmentation information $\{\mathcal{F}\}$. For label quantization, even though augmentation data occupies only up to 1.06% of total storage, its impact grows with higher compression. For example, at IPC10, the theoretical 500× compression rate is reduced to an actual compression rate of 198×, since $\{\mathcal{F}\}$ uses 20 MB of a 30 MB storage budget. Label pruning removes both $\{\mathbf{z}\}$ and $\{\mathcal{F}\}$, but the actual compression rate is slightly lower than the theoretical rate since we store image indices of every batch, trading a small overhead for implementation flexibility.

C. Ablation Study

Table VII presents a comprehensive ablation study that incrementally adds each component to isolate their individual effects and demonstrate their complementary benefits. We analyze the

TABLE IV

IMAGENET-1K LABEL PRUNING RESULT. OUR METHOD CONSISTENTLY SHOWS A BETTER PERFORMANCE UNDER VARIOUS PRUNING LPQLD USES LABEL PRUNING, AND LPQLD USES BOTH LABEL PRUNING AND LABEL QUANTIZATION. THE BASE PRUNING RATE FOR LPQLD IS 10×. $\{\mathbf{z}, \mathcal{F}\}$ DENOTES TOTAL STORAGE OF BOTH AUGMENTATION INFORMATION AND SOFT LABELS. THE VALIDATION MODEL IS RESNET-18. † DENOTES THE REPORTED RESULTS.

ResNet-18	∼ 1× SRe^{2}L	CDA	LPLD	∼ 10× SRe^{2}L	CDA	LPLD	LPQLD	∼ 20× SRe^{2}L	CDA	LPLD	LPQLD	∼ 40× SRe^{2}L	CDA	LPLD	LPQLD	∼ 80× LPQLD	∼ 200× LPQLD
IPC10 (168 M)																	
Acc (%)	20.1	33.3	34.6	18.9	28.4	32.7	**32.8**$\pm$0.2	16.0	21.9	28.6	**30.2**$\pm$0.8	**32.2**$\pm$0.2	11.4	13.2	20.2	**23.1**$\pm$0.5	**29.6**$\pm$0.2	**27.3**$\pm$0.6	**20.0**$\pm$0.3
$\{\mathbf{z}, \mathcal{F}\}$	5,862 M			580 M			570 M	310 M			310 M	300 M	145 M			145 M	129 M	74 M	29 M
$\{\mathbf{z}, \mathcal{F}\}_\uparrow$	1×			10×			10×	19×			19×	20×	40×			40×	45×	79×	202×
IPC20 (331 M)																	
Acc (%)	33.6	44.0	47.2	31.1	39.7	44.7	**45.0**$\pm$0.1	29.2	34.1	41.0	**42.3**$\pm$0.2	**43.2**$\pm$0.3	21.7	24.0	33.0	**35.7**$\pm$0.4	**41.2**$\pm$0.2	**38.6**$\pm$0.2	**30.5**$\pm$0.5
$\{\mathbf{z}, \mathcal{F}\}$	11,722 M			1167 M			1140 M	623 M			623 M	604 M	292 M			292 M	259 M	150 M	59 M
$\{\mathbf{z}, \mathcal{F}\}_\uparrow$	1×			10×			10×	19×			19×	19×	40×			40×	45×	78×	199×
IPC50 (819 M)																	
Acc (%)	46.8†	53.5†	55.4	44.1	50.3	54.4	**54.5**$\pm$0.2	41.5	46.1	51.8	**52.8**$\pm$0.1	**53.1**$\pm$0.1	35.5	38.0	46.7	**48.4**$\pm$0.0	**51.3**$\pm$0.1	**49.6**$\pm$0.1	**43.0**$\pm$0.1
$\{\mathbf{z}, \mathcal{F}\}$	29,302 M			2,928 M			2,850 M	1,562 M			1,562 M	1,515 M	733 M			733 M	649 M	375 M	147 M
$\{\mathbf{z}, \mathcal{F}\}_\uparrow$	1×			10×			10×	19×			19×	19×	40×			40×	45×	78×	199×
IPC100 (1,630 M)																	
Acc (%)	52.8†	58.0†	59.4	51.1	55.1	58.8	**59.1**$\pm$0.1	49.5	53.3	57.4	**58.0**$\pm$0.2	**57.5**$\pm$0.1	44.4	47.2	54.0	**54.9**$\pm$0.1	**56.2**$\pm$0.1	**54.9**$\pm$0.2	**50.1**$\pm$0.4
$\{\mathbf{z}, \mathcal{F}\}$	58,606 M			5,871 M			5,700 M	3,132 M			3,132 M	3,038 M	1,468 M			1,468 M	1,301 M	752 M	295 M
$\{\mathbf{z}, \mathcal{F}\}_\uparrow$	1×			10×			10×	19×			19×	19×	40×			40×	45×	78×	199×
IPC200 (3,256 M)																	
Acc (%)	57.0†	**63.3**†	62.6	56.5	59.4	62.4	**62.4**$\pm$0.1	55.1	58.3	61.7	**61.9**$\pm$0.0	**60.8**$\pm$0.2	51.9	54.4	59.6	**60.1**$\pm$0.1	**59.9**$\pm$0.2	**58.8**$\pm$0.2	**55.4**$\pm$0.2
$\{\mathbf{z}, \mathcal{F}\}$	117,208 M			11,753 M			11,400 M	6,269 M			6,269 M	6,082 M	2,939 M			2,939 M	2,606 M	1,509 M	594 M
$\{\mathbf{z}, \mathcal{F}\}_\uparrow$	1×			10×			10×	19×			19×	19×	40×			40×	45×	78×	197×

TABLE V

LABEL PRUNING RESULT ON IMAGENET-21K-P, USING RESNET-18. $\mathbb{I}$ DENOTES IMAGE STORAGE. $\{\mathbf{z}, \mathcal{F}\}$ DENOTES TOTAL STORAGE OF BOTH AUGMENTATION INFORMATION AND SOFT LABELS. SETTING A, B, C DENOTE USING DIFFERENT BASE LABEL PRUNING RATES OF 10×, 20×, AND 40×, RESPECTIVELY. BEST SETTING IS HIGHLIGHTED IN **BLUE BOLD** TEXT. † DENOTES REPORTED RESULTS.

IPC	$\mathbb{I}$	1× $\{\mathbf{z}, \mathcal{F}\}$	SRe^{2}L	CDA	D3S	LPLD	10× $\{\mathbf{z}, \mathcal{F}\}$	LPLD	LPQLD	40× $\{\mathbf{z}, \mathcal{F}\}$	LPLD	LPQLD	Setting LPQLD	∼100× $\{\mathbf{z}, \mathcal{F}\}$	LPQLD	∼400× $\{\mathbf{z}, \mathcal{F}\}$	LPQLD	∼500× $\{\mathbf{z}, \mathcal{F}\}$	LPQLD
IPC10	3G	643G	18.5†	22.6†	**26.9**†	25.4	65G	24.1	**28.2**	16G	21.3	**25.6**	10×	6.4G	23.5	1.6G	20.5	1.2G	18.9
													20×	6.6G	**23.9**	1.6G	**20.9**	0.8G	**19.2**
													40×	**6.0G**	23.3	**1.3G**	19.8	**0.8G**	18.1
IPC20	5G	1285G	20.5†	26.4†	28.5†	**30.3**	129G	31.3	**34.4**	32G	29.4	**33.8**	10×	13G	26.1	3.1G	23.0	2.3G	21.9
													20×	14G	27.1	3.3G	23.9	1.7G	22.7
													40×	**12G**	**27.7**	**2.7G**	**24.1**	**1.6G**	**23.3**

TABLE VI

STORAGE BREAKDOWN OF SOFT LABELS OF A SINGLE BATCH OF SIZE 128. Q DENOTES LABEL QUANTIZATION AND P DENOTES LABEL PRUNING. NOTE THAT THE BYTES HERE INCLUDE THE OVERHEAD FOR PYTORCH'S SERIALIZATION FORMAT. IMAGENET-1 K.

Component	Shape	Bytes	% of Total	Type	Q	P
1. coordinates of crops	[128, 4]	2224	0.86%	$\mathcal{F}$	×	✓
2. flip status	[128]	1328	0.51%	$\mathcal{F}$	×	✓
3. index of cutmix images	[128]	1328	0.51%	$\mathcal{F}$	×	✓
4. strength of cutmix	[1]	880	0.34%	$\mathcal{F}$	×	✓
5. cutmix bounding box	[4]	1904	0.73%	$\mathcal{F}$	×	✓
6. prediction logits	[128, 1000]	257200	**98.94%**	$\mathbf{z}$	✓	✓

TABLE VII

ABLATION STUDY OF LPQLD. □ DENOTES IMAGE GENERATION COMPONENTS, △ DENOTES SOFT LABEL COMPRESSION COMPONENTS. C: CLASS-WISE MATCHING. CS: CLASS-WISE SUPERVISION. BP: BATCH-LEVEL PRUNING. DKR: DYNAMIC KNOWLEDGE REUSE. Q: LABEL QUANTIZATION (TOP-k). CA: CALIBRATED STUDENT-TEACHER ALIGNMENT. RESULTS ARE ON RESNET-18, IMAGENET-1K, IPC50.

	C	CS	BP	DKR	Q	CA	10×	50×	100×	Full
□	-	-	-	-	-	-	49.4	34.8	25.4	52.0
□	✓	-	-	-	-	-	51.9$_{\uparrow 2.5}$	37.7$_{\uparrow 2.9}$	22.6$_{\downarrow 2.8}$	54.7$_{\uparrow 2.7}$
□	✓	✓	-	-	-	-	53.2$_{\uparrow 3.8}$	39.7$_{\uparrow 4.9}$	29.1$_{\uparrow 3.7}$	55.4$_{\uparrow 3.4}$
△	✓	✓	✓	-	-	-	54.4$_{\uparrow 5.0}$	43.1$_{\uparrow 8.3}$	33.7$_{\uparrow 8.3}$	55.4$_{\uparrow 3.4}$
△	✓	✓	✓	✓	-	-	54.4$_{\uparrow 5.0}$	47.8$_{\uparrow 13.0}$	41.4$_{\uparrow 16.0}$	55.4$_{\uparrow 3.4}$
△	✓	✓	✓	✓	✓	-	-	49.7$_{\uparrow 14.9}$	47.9$_{\uparrow 22.5}$	55.4$_{\uparrow 3.4}$
△	✓	✓	✓	✓	✓	✓	-	51.3$_{\uparrow 16.5}$	49.6$_{\uparrow 24.2}$	55.4$_{\uparrow 3.4}$

results across different compression rates (10×, 50×, 100×) and full label settings. Rows are organized into two groups: □ **Image Generation** (Rows 1-3) focuses on improving synthetic image diversity, while △ **Label Compression** (Rows 4-7) addresses soft label storage efficiency.

□ *Row 1 (Baseline):* SRe^{2}L implementation under CDA's hyperparameter settings. Label compression is achieved by random epoch-level label pruning.

TABLE VIII
LPQLD ON DIFFERENT DISTILLED DATASETS. † DENOTES THE REPORTED RESULTS. RESNET-18, IMAGENET-1K.

$\{\mathbf{z}\}$	DWA [15]				DELT [20]				Minimax [47]			
	1×	10×	50×	100×	1×	10×	50×	100×	1×	10×	50×	100×
10	37.9†	33.4	29.2	26.4	45.8†	41.3	35.6	35.0	44.3†	39.3	37.7	34.9
50	55.2†	51.7	47.8	46.0	59.2†	58.0	55.5	54.4	58.6†	56.1	55.0	53.7
100	59.2†	57.3	54.6	52.9	62.4†	61.7	58.6	58.5	-	59.0	58.4	57.4

TABLE IX
CROSS-ARCHITECTURE RESULT. RESNET-18 IS USED AS THE TEACHER MODEL FOR ALL STUDENT MODELS. IPC50, IMAGENET-1 K.

Model	Size	Full Acc	1×	∼30×	∼40×	∼80×
ResNet-18 [44]	11.7M	69.76	55.4	50.2	51.3	49.6
ResNet-50 [44]	25.6M	76.13	62.2	57.8	56.7	54.8
EfficientNet-B0 [48]	5.3M	77.69	55.5	53.2	54.0	52.1
MobileNet-V2 [49]	3.5M	71.88	49.1	47.8	45.9	45.8
Swin-V2-Tiny [50]	28.4M	82.07	40.6	42.0	44.6	38.7

TABLE X
COMPARISON WITH PRIOR LABEL QUANTIZATION METHODS, MS (MARGINAL SMOOTHING) AND MR (MARGINAL RENORM) [34]. * DENOTES MS AND MR HAVING APPROXIMATELY THE SAME STORAGE. IMAGENET-1K, RESNET-18.

Method	IPC10				IPC50			
Top-k	k=50	k=25	k=5	k=1	k=50	k=25	k=5	k=1
MS [34]	16.5 ±0.2	12.4 ±0.4	7.5 ±0.2	4.0 ±0.1	38.8 ±0.3	33.6 ±0.2	19.5 ±0.3	7.7 ±0.6
MR [34]	17.7 ±0.4	16.6 ±0.2	12.6 ±0.4	5.1 ±0.1	28.6 ±0.2	31.7 ±0.2	34.0 ±0.2	16.7 ±0.2
$\{\mathbf{z}, \mathcal{F}\}$*	753M	478M	203M	203M	3,768M	2,393M	1,020M	1,020M
Ours $\{\mathbf{z}, \mathcal{F}\}$	**29.6**±0.2 **129M**				**51.3**±0.1 **649M**			

□ *Row 2 (+ Class-wise matching):* Re-batches images within the same class during image "recover" phase. This modification increases within-class diversity by exposing the model to varied intra-class patterns.

□ *Row 3 (+ Class-wise supervision):* Replaces global BN statistics with class-wise BN statistics computed separately for each class in the "squeeze" phase, which are then used as supervision in the "recover" phase. Since Row 2 already uses class-wise matching, this provides more precise and reasonable supervision that aligns with the class-specific batching, leading to stronger diversity constraints and larger improvements at higher compression rates.

△ *Row 4 (+ Batch-level pruning):* Switches from epoch-level to batch-level label pruning, enabling finer-grained control over which augmentation-label pairs to retain. This structured selection mechanism consistently improves performance across all compression rates.

△ *Row 5 (+ Dynamic knowledge reuse):* Introduces temperature annealing that gradually decreases teacher temperature during training. This allows the model to extract maximum information from retained labels by starting with softer distributions (learning broad class relationships) and transitioning to sharper distributions (refining specific class boundaries). The gains are most pronounced at high compression rates where label scarcity makes efficient knowledge extraction critical.

△ *Row 6 (+ Label quantization):* Applies Top-k selection to compress soft labels by storing only the top-k *pre-softmax* logits. These stored logits can be later recomputed with different temperatures during training, enabling flexible temperature scheduling. Combined with label pruning, this achieves additional compression. Quantization is not applied at 10× since the base pruning already operates at this rate.

△ *Row 7 (+ Calibrated alignment):* Adds grid-search optimization to dynamically adjust student temperature, aligning the student distribution with the sharper quantized teacher distribution. This calibration is essential for teacher label quantization, without it (as would occur in Row 5), temperature adjustment creates distribution mismatch. Here, it properly compensates for the information loss from quantization, enabling effective dual compression.

D. Additional Analysis

Cross-Architecture Performance: Table IX demonstrates how our compression methods generalize across diverse network architectures, with ResNet-18 serving as the teacher model throughout. The results reveal consistent patterns across different model families, from convolutional networks (ResNet-18, ResNet-50) to efficient mobile architectures (MobileNet-V2, EfficientNet-B0) and transformer-based models (Swin-V2-Tiny). At 10× compression using LPLD, all architectures maintain performance nearly equivalent to their uncompressed (1×) counterparts, with minimal accuracy drops ranging from 0 to 2 percentage points. More aggressive compression (30×, 50×, and 100×) shows expected performance degradation, though the hybrid LPQLD method helps mitigate losses at higher compression rates. Notably, the Swin-V2-Tiny transformer model shows unique behavior, actually **improving** by 2.2 percentage points at 10× compression and maintaining strong performance at 50× compression with LPQLD. These cross-architecture results highlight our findings that when dataset diversity is sufficient, the need for extensive augmentations and soft labels can be significantly reduced regardless of the target architecture.

Label Compression Performance on Different Distilled Datasets: Since our label pruning and quantization techniques are explicit and transferable, we evaluate the effectiveness across various distilled datasets in Table VIII. We apply our label pruning approach to three additional dataset distillation methods: the diversity-driven DWA [15], multisized-pattern DELT [20], and diffusion-based Minimax [47]. Results demonstrate strong compatibility with all methods, with particularly significant improvements observed for IPC settings.

Comparison with Prior Label Quantization Methods: Table X compares LPQLD with FKD's Marginal Smoothing (MS) and Marginal Re-Norm (MR) [34]. Both MS and MR suffer severe performance degradation as Top-k decreases, with **Top-5 and Top-1 having identical storage** (203M for IPC10, 1,020M for IPC50) because auxiliary data dominates, representing the **fundamental limit of quantization-only approaches**. In contrast, LPQLD achieves superior performance (29.6% for

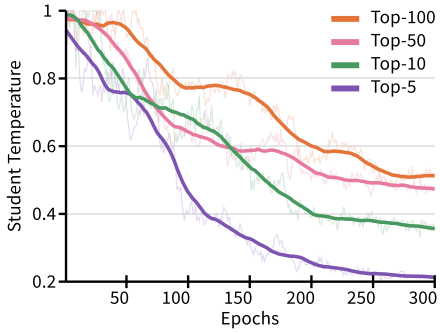


Fig. 11. Optimal student temperature $\hat{\tau}^*$ over training epochs.

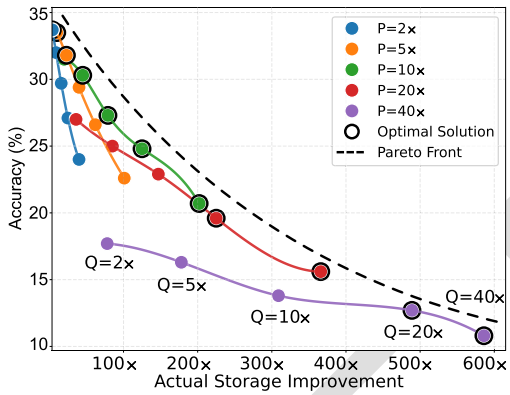


Fig. 12. Pareto front of accuracy vs. storage for different label pruning (P) and label quantization (Q) combinations. IPC10, ResNet-18, ImageNet-1 K.

IPC10, 51.3% for IPC50) with significantly lower storage (129M and 649M) by combining pruning and quantization, breaking through this bottleneck.

Optimal Student Temperature: Fig. 11 showcases when the student model is randomly initialized and predictions are not accurate, a relatively high temperature is assigned. When models can make more accurate predictions, the optimal temperature for students goes down to better align with the concentrated probability mass of the teacher distribution. In addition, we observe that a higher quantization ratio (smaller Top-k) has a lower average student temperature, which is consistent with our hypothesis that student prediction distribution should be scaled more to properly align with a more concentrated teacher distribution.

Pareto Front Analysis in a Fixed IPC: From Fig. 12, we present two critical observations. First, there exists a Pareto front where specific configurations simultaneously excel in both accuracy and storage efficiency compared to alternatives. Second, even within a single IPC (i.e., IPC10), we observe significant configuration variations along the Pareto front. Specifically, when actual storage improvements ($\{\mathbf{z}, \mathcal{F}\}$) remain below $50\times$, combining a $5\times$ label pruning rate with label quantization outperforms configurations using larger pruning ratios as the base compression rate. However, as the actual compression rate increases to $50\times$-$200\times$, configurations with a $10\times$ label pruning rate clearly dominate the Pareto front. Beyond $200\times$ compression, adopting a $20\times$ label pruning rate emerges as the optimal approach.

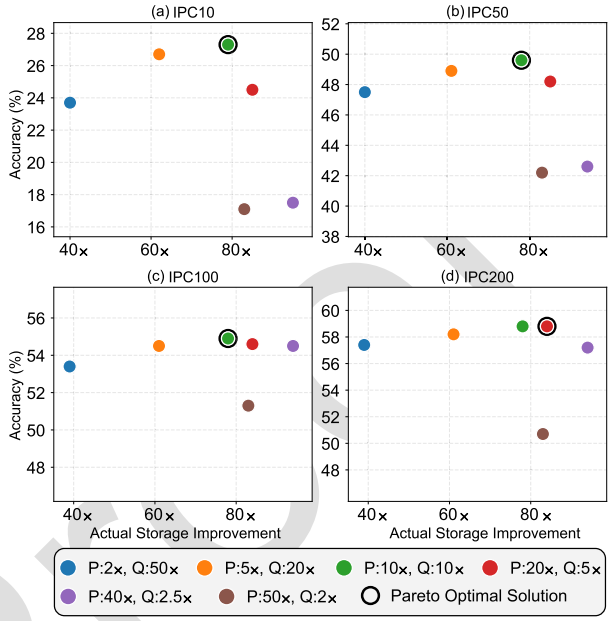


Fig. 13. Illustration of Pareto front changes for different IPCs under the same theoretical compression rate (i.e., $100\times$). `P:50x, Q:2x` denotes theoretically compression of label pruning is $50\times$ and that of label quantization is $2\times$.

Pareto Front Analysis in Different IPCs: Fig. 13 demonstrates that **larger dataset sizes favor higher base compression rates at equivalent compression ratios**. Given a fixed theoretical compression ratio $\{\mathbf{z}\}$, we observe that Pareto optimal solutions shift across dataset scales. At IPC10, configurations with a $10\times$ pruning rate (green dot) achieve higher performance despite lower actual storage improvement compared to configurations using a $20\times$ pruning rate as base. However, as the dataset scales from IPC10 to IPC200, this performance gap consistently diminishes. Eventually, at IPC200, the $20\times$ configuration (red dot) achieves dominance by maintaining comparable performance levels while delivering higher actual storage improvements.

Visualization: Fig. 2(b) visualizes our method on three classes. More visualizations are provided in Appendix D, available online.

V. Conclusion

This work addressed the critical storage bottleneck in dataset distillation by developing Label Pruning and Quantization for Large-scale Distillation (LPQLD). Our analysis identified two fundamental issues: (1) *insufficient image diversity*, where high within-class similarity requires extensive augmentation, and (2) *insufficient supervision diversity*, where limited variety in supervisory signals leads to performance degradation at high compression rates. We systematically addressed both challenges by enhancing *image diversity* via class-wise BN supervision during synthesis, and introducing Label Pruning with Dynamic Knowledge Reuse for *label-per-augmentation diversity* and Label Quantization with Calibrated Student-Teacher Alignment for *augmentation-per-image diversity*. Our extensive

experiments demonstrate that LPQLD achieves 78× compression on ImageNet-1K and 500× on ImageNet-21 K while maintaining or exceeding prior performance. By introducing the supervision diversity framework, we enable practical deployment in resource-constrained environments and pave the path for soft-label empowered dataset distillation. Limitations and future work are in Appendix E, available online. Ethical considerations are in Appendix F, available online.

REFERENCES

[1] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," 2018, *arXiv:1811.10959*.

[2] K. Wang et al., "CAFE: Learning to condense dataset by aligning features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12196–12205.

[3] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu, "Dataset distillation by matching training trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10708–10717.

[4] B. Zhao, K. R. Mopuri, and H. Bilen, "Dataset condensation with gradient matching," in *Proc. Int. Conf. Learn. Representations*, 2021.

[5] B. Zhao and H. Bilen, "Dataset condensation with distribution matching," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 6514–6523.

[6] Z. Yin, E. Xing, and Z. Shen, "Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective," in *Proc. Adv. Neural Inform. Process. Syst.*, 2023, pp. 73582–73603.

[7] Z. Yin and Z. Shen, "Dataset distillation via curriculum data synthesis in large data ERA," *Trans. Mach. Learn. Res.*, Nov. 2024. [Online]. Available: https://openreview.net/forum?id=PlaZD2nGCl

[8] P. Sun, B. Shi, D. Yu, and T. Lin, "On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9390–9399.

[9] S. Shao, Z. Yin, M. Zhou, X. Zhang, and Z. Shen, "Generalized large-scale data condensation via various backbone and statistical matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16709–16718.

[10] M. Zhou, Z. Yin, S. Shao, and Z. Shen, "Self-supervised dataset distillation: A good compression is all you need," 2024, *arXiv:2404.07976*.

[11] L. Xiao and Y. He, "Are large-scale soft labels necessary for large-scale dataset distillation?," in *Proc. 38th Annu. Conf. Neural Inf. Process. Syst.*, 2024, pp. 16406–16437.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[13] J.-H. Kim et al., "Dataset condensation via efficient synthetic-data parameterization," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 11102–11118.

[14] J. Cui, R. Wang, S. Si, and C.-J. Hsieh, "Scaling up dataset distillation to ImageNet-1 k with constant memory," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 6565–6590.

[15] J. Du, X. Zhang, J. Hu, W. Huang, and J. T. Zhou, "Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 119443–119465.

[16] X. Zhong, B. Chen, H. Fang, X. Gu, S.-T. Xia, and E.-H. Yang, "Going beyond feature similarity: Effective dataset distillation based on class-aware conditional mutual information," in *Proc. 13th Int. Conf. Learn. Representations*, 2025.

[17] D. Bo, S. Liu, and X. Wang, "Understanding dataset distillation via spectral filtering," 2025, *arXiv:2503.01212*.

[18] J. Cui, Z. Li, X. Ma, X. Bi, Y. Luo, and Z. Shen, "Dataset distillation via committee voting," 2025, *arXiv:2501.07575*.

[19] Y. M. Asano and A. Saeed, "The augmented image prior: Distilling 1000 classes by extrapolating from a single image," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.

[20] Z. Shen, A. Sherif, Z. Yin, and S. Shao, "DELT: A simple diversity-driven earlylate training for dataset distillation," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2024, pp. 4797–4806.

[21] H. Liu et al., "Dataset distillation via the Wasserstein metric," in *Proc. Int. Conf. Comput. Vis.*, 2025, pp. 1205–1215.

[22] Y. He, L. Xiao, and J. T. Zhou, "You only condense once: Two rules for pruning condensed datasets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 39382–39394.

[23] N. Loo, A. Maalouf, R. Hasani, M. Lechner, A. Amini, and D. Rus, "Large scale dataset distillation with domain shift," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 32759–32780.

[24] K. Wang et al., "Emphasizing discriminative features for dataset distillation in complex scenarios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 30451–30461.

[25] D. Su, J. Hou, W. Gao, Y. Tian, and B. Tang, "D ^4: Dataset distillation via disentangled diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 5809–5818.

[26] L. Li, G. Li, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "Generative dataset distillation: Balancing global structure and local details," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 7664–7671.

[27] M. Chen, J. Du, B. Huang, Y. Wang, X. Zhang, and W. Wang, "Influence-guided diffusion for dataset distillation," in *Proc. 13th Int. Conf. Learn. Representations*, 2025.

[28] X. Zhang, J. Du, P. Liu, and J. T. Zhou, "Breaking class barriers: Efficient dataset distillation via inter-class feature compensator," in *Proc. Int. Conf. Learn. Representations*, 2025.

[29] L. Xiao, S. Liu, Y. He, and X. Wang, "Rethinking large-scale dataset compression: Shifting focus from labels to images," 2025, *arXiv:2502.06434*.

[30] T. Qin, Z. Deng, and D. Alvarez-Melis, "A label is worth a thousand images in dataset distillation," in *Proc. Adv. Neural Inform. Process. Syst.*, 2024, pp. 131946–131971.

[31] Z. Shen, "Ferkd: Surgical label adaptation for efficient distillation," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 1666–1675.

[32] S. Kang, Y. Lim, and H. Shim, "Label-augmented dataset distillation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2025, pp. 1457–1466.

[33] R. Yu, S. Liu, J. Ye, and X. Wang, "Teddy: Efficient large-scale dataset distillation via taylor-approximated matching," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 1–17.

[34] Z. Shen and E. Xing, "A fast knowledge distillation framework for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 673–690.

[35] H. Yin et al., "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8715–8724.

[36] H. Zhang, S. Li, P. Wang, D. Zeng, and S. Ge, "M3D: Dataset condensation by minimizing maximum mean discrepancy," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 9314–9322.

[37] T. Qin, Z. Deng, and D. Alvarez-Melis, "Distributional dataset distillation with subtask decomposition," in *Proc. Int. Conf. Learn. Representations 2024 Workshop Navigating Addressing Data Problems Found. Models*, 2024.

[38] Y. He, L. Xiao, J. T. Zhou, and I. Tsang, "Multisize dataset condensation," in *Proc. Int. Conf. Learn. Representations*, 2024.

[39] A. Jafari, M. Rezagholizadeh, P. Sharma, and A. Ghodsi, "Annealing knowledge distillation," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main Volume*, 2021, pp. 2493–2504.

[40] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[41] S.I. Mirzadeh, M.A. Farajtabar, N. Li, A. L. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5191–5198.

[42] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," in *Proc. Int. Conf. Learn. Representations*, 2016.

[43] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," in *Proc. Int. Conf. Neural Inf. Process. Syst. Benchmarks Track (Round 1)*, 2021.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 3008–3017.

[46] R. Wightman, "PyTorch image models," 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models

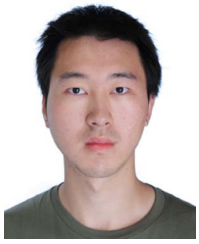
[47] J. Gu et al., "Efficient dataset distillation via minimax diffusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 15793–15803.

[48] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[50] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12009–12019.

Lingao Xiao is currently a student with the National University of Singapore and is an intern student with the Centre for Frontier AI Research (CFAR), Agency for Science Technology and Research (A*STAR), Singapore, and also with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore. He received his bachelor's degree in computer engineering at Nanyang Technological University, Singapore. His research interests include efficient deep learning.



Yang He (Member, IEEE) received the BS and MSc degrees from the University of Science and Technology of China, Hefei, China, in 2014 and 2017, respectively, and the PhD degree from the University of Technology Sydney, in 2022. He is currently a scientist with the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore, and also with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore. He is also an adjunct assistant professor with the National University of Singapore (NUS). His research interests include deep learning, computer vision, model compression, and dataset compression.