Robust Trajectory Distillation: Hybrid Reweighting Meets Teacher-Inspired Targets

Kaifeng Chen[1], Lechao Cheng[1,2,3,4]✉, Jiyang Li[1], Shengeng Tang[1], Fan Zhang[5], Yantao Pan[4], Yaxiong Wang[1], Tuanrui Hui[1], and Zhun Zhong[1]

[1] Hefei University of Technology
[2] Jianghuai Advance Technology Center
[3] Anhui Provincial Key Laboratory of Humanoid Robots
[4] Kaiyang Laboratory, Chery
[5] Zhejiang University of Technology

Abstract. Dataset distillation (DD) condenses large corpora into compact, information-rich subsets for efficient training and reuse. However, under noisy supervision, DD risks condensing corrupted associations together with useful signals, degrading robustness. Conventional noisy-label remedies (sample selection, loss weighting, label correction) tightly couple noise estimation with model optimization, often require clean anchors, and can amplify confirmation bias—assumptions that are misaligned with DD's goal of compact, plug-and-play supervision. We therefore propose a trajectory-based DD framework that jointly suppresses noise and preserves transferable knowledge without relabeling or clean subsets. It comprises two complementary components: Selective Guidance Reweighting (SGR), which fuses global forgetting patterns (second-split forgetting) with local neighborhood consistency into a progressive reweighting scheme that prioritizes clean supervision along the teacher trajectory; and Teacher-Inspired Auxiliary Targets (TIAT), which inject auxiliary residual guidance distilled from intermediate teacher dynamics to reinforce informative signals while remaining internally consistent. Together, SGR and TIAT produce distilled datasets with cleaner and richer representations under noisy supervision. The framework is robust, label-preserving, computationally lightweight, and broadly applicable, yielding consistent gains over state-of-the-art DD baselines across symmetric, asymmetric, and real-world noise.

Keywords: Dataset Distillation · Learning with Noisy Labels · Trajectory Matching · Robust Learning · Sample Reweighting

1 Introduction

The growing scale and complexity of deep models have spurred interest in dataset distillation (DD)—a technique that synthesizes compact, information-rich subsets capable of approximating the training efficacy of large real datasets [5, 39].

✉ Corresponding Author: chenglc@hfut.edu.cn

By condensing essential knowledge into a small synthetic set, dataset distillation enables efficient model retraining, rapid adaptation, and even privacy-preserving data sharing. Recent studies [4, 12, 42] further highlight DD's robustness potential under imperfect supervision, suggesting that distilled datasets can serve as a natural filter that separates informative clean signals from noisy patterns.

However, most existing DD frameworks implicitly assume clean supervision, which severely limits their reliability in real-world noisy-label environments. Large-scale web-curated datasets often contain substantial annotation noise due to human bias, crowdsourcing inconsistency, and semantic ambiguity [22]. Traditional noisy-label learning techniques—such as sample selection [23, 28, 51], loss weighting [24, 45], and label correction [30, 33]—attempt to identify and mitigate label corruption during training. Yet, these methods tightly couple noise estimation and model optimization, forming a self-referential feedback loop where the model acts as both noise detector and predictor. Such coupling often amplifies confirmation bias, leading to overconfidence in incorrect labels and reduced generalization, while also requiring clean validation anchors or heavy iterative refinement, limiting their scalability.

In this context, dataset distillation under noisy supervision emerges as a compelling alternative paradigm. Rather than continuously re-estimating noisy labels, DD aims to learn a condensed set of synthetic samples that intrinsically encode cleaner supervision and transferable knowledge. Nevertheless, two fundamental challenges remain unresolved:

1. **Noise-Agnostic Distillation.** Existing methods such as DANCE [42] and DATM [12] lack mechanisms to distinguish corrupted from clean signals during condensation, causing performance degradation under noise (e.g., a 2–3% accuracy drop on CIFAR-10 at 20% noise, 50 IPC).
2. **Capacity-Constrained Synthesis.** Distilled datasets are typically parameterized as fixed-size image tensors (e.g., 50 IPC), restricting representational capacity and leading to premature information compression before noise–clean disentanglement completes.

These issues motivate our central question: *How can we improve both the quality and capacity of distilled data under noisy supervision?* We address this from two complementary perspectives:

– **Challenge ❶ (Learning Better):** Improve the fidelity of synthetic data by enhancing teacher-signal reliability and promoting cleaner knowledge retention.
– **Challenge ❷ (Learning More):** Increase the amount of high-quality supervision absorbed by the synthetic dataset through richer teacher–student interactions.

To this end, we propose two synergistic modules. **Selective Guidance Reweighting (SGR)** refines the teacher trajectory with a hybrid KNN–SSFT [20,

27] mechanism that integrates local neighborhood consistency with global forgetting trends, yielding cleaner and more trustworthy guidance signals. **Teacher-Inspired Auxiliary Targets (TIAT)** supplement the main trajectory with auxiliary residual signals derived from the teacher's internal dynamics—such as intermediate predictions or temporal consistency—thus enriching supervision without introducing conflicting objectives. Together, SGR and TIAT strengthen both the *quality* and *capacity* of knowledge transfer, enabling robust and scalable dataset distillation under noisy-label conditions.

In summary, we propose a distillation framework specifically tailored for learning under noisy supervision. Conceptually, the framework mirrors a teacher who both (i) refines their expertise to deliver higher-quality instruction and (ii) assigns *homework-like* auxiliary tasks that further reinforce and consolidate the student's learning. Our contributions are as follows:

- We introduce a progressive mechanism that integrates both dynamic and static sample reweighting, fusing complementary teacher-trajectory signals and yielding robust, consistent suppression of label noise.
- We propose an auxiliary guidance regularization that enforces consistency with clean trajectories during distillation, amplifying the influence of clean signals throughout training and mitigating error amplification.
- The framework is label-preserving and computationally lightweight: it requires neither relabeling nor costly retraining schedules, making it practical for large-scale, real-world noisy datasets.
- Extensive experiments across diverse datasets and noise regimes show consistent improvements over strong baselines, demonstrating the method's generality and robustness.

2 Related Works

2.1 Dataset Distillation

Dataset distillation [5, 31, 39] aims to compress a large dataset into a compact synthetic set while preserving downstream performance. The foundational work [39] introduced the idea of optimizing synthetic data to match training dynamics observed on real data. Subsequent research has evolved along three major paradigms: *meta-learning*, *parameter matching*, and *distribution matching*. Meta-learning methods [25, 29, 49] adopt a bi-level optimization framework to generalize across model initializations but suffer from high computational overhead. Parameter matching, including gradient [47] and trajectory matching [2, 8], directly aligns model updates between real and synthetic data, with trajectory-based methods achieving state-of-the-art results [12]. Variants further explore progressive optimization [3], hybrid data composition [19], and group-wise structures [14]. Distribution matching offers an efficient alternative by aligning real and synthetic data distributions in feature space [43, 46], class relations [6], or image-label correlations [42], without requiring nested optimization. Recent advances even reformulate this as neural feature alignment [37].

Beyond dataset-level condensation, model-level knowledge distillation [40, 44] studies have shown that the formulation of teacher-derived targets is critical to effective knowledge transfer. For example, regularizing the direction and norm of student representations toward teacher-derived class prototypes can substantially improve transfer quality [40]. Unlike these model-compression settings, our goal is to optimize a compact synthetic dataset whose induced training dynamics remain reliable under label noise. However, most methods assume clean supervision and degrade under real-world label noise. To address this limitation, we propose a noise-aware extension of trajectory matching that explicitly models and suppresses corruption during both the distillation and deployment phases.

2.2 Learning with Noisy Labels

Real-world datasets often contain corrupted labels due to annotation errors or automated collection. To address this, noisy label learning has developed three major strategies: *sample selection*, *label correction*, and *sample reweighting*. Sample selection methods aim to identify clean data during training, typically using loss-based filtering. A representative work is Decoupling [28], which introduced the small-loss trick. Follow-up works [13, 16] incorporate external guidance or co-training, while curriculum-based methods [26, 48] and early-learning regularization [23] enhance robustness via dynamic filtering or temporal consistency. Some approaches even detect noisy samples prior to training [38, 51]. The distinction between noisy and intrinsically difficult samples becomes particularly challenging under imbalanced class distributions. Fang et al. [10] addressed this problem by separating corrupted examples from clean tail-class samples through augmentation consistency and leave-noise-out regularization. Label correction methods refine labels using model predictions [30, 50], semi-supervised strategies [21], or meta-learning and neighborhood consistency [20, 36]. Sample reweighting strategies [7, 32, 45] instead adjust loss contributions based on sample reliability. Additionally, K-NN-based methods [1, 15] have proven effective for noise detection by evaluating local label consistency in the feature space. Distillation-based robust learning has also been investigated under imperfect annotations. Reliability-aware sample selection and consistency regularization can be used to ensure that only trustworthy knowledge is transferred between models [11]. Building on this, dataset distillation has recently emerged as a promising alternative. As shown in [4], it addresses limitations such as iterative error amplification and privacy concerns, while offering strong performance under noisy supervision.

3 Preliminaries

3.1 Dataset Distillation with Noisy Labels

Let $\mathcal{D}_{\text{real}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$ denote a real-world training dataset, where $\tilde{y}_i$ is the observed (and potentially noisy) label for input x_i. We assume the existence of label noise such that $\tilde{y}_i \neq y_i^*$ for some i, where y_i^* is the clean but unobserved

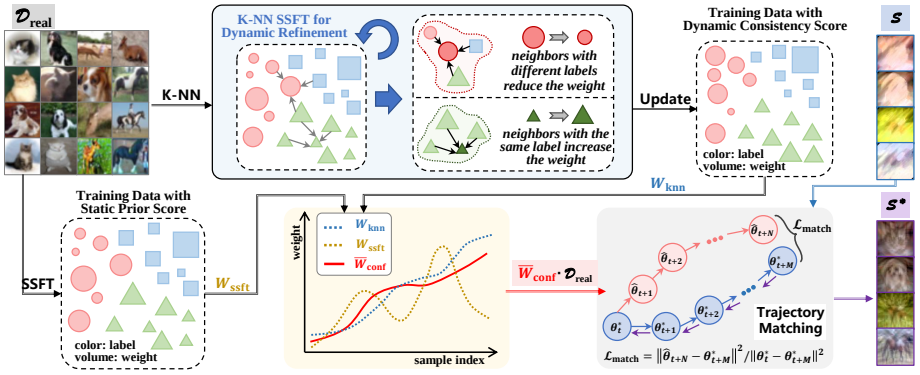

Fig. 1: The overall pipeline. Our proposed pipeline includes two main components: (1) during teacher trajectory training, we apply sample-specific weighting adjustments to modulate the influence of each sample based on its estimated reliability; and (2) in the subsequent distillation phase, we leverage a subset of high-confidence samples to impose additional constraints and regularization, thereby enhancing the quality and generalization of the distilled dataset.

ground-truth label. In contrast, the test set $\mathcal{D}_{\text{test}}$ is assumed to be entirely clean and representative of the true data distribution, and is used to evaluate generalization performance. Our objective is to synthesize a compact dataset $\mathcal{S}$, with $|\mathcal{S}| \ll |\mathcal{D}_{\text{real}}|$, such that a model trained solely on $\mathcal{S}$ achieves lower generalization error on $\mathcal{D}_{\text{test}}$ than one trained on the full noisy dataset $\mathcal{D}_{\text{real}}$. Formally, the distillation task can be formulated as the following optimization problem:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathcal{L}(\mathcal{S}, \mathcal{D}_{\text{real}}), \tag{1}$$

where $\mathcal{L}$ is a general objective function. Following prior trajectory-matching approaches [12], we instantiate $\mathcal{L}$ as a loss that aligns the learning dynamics of a student trained on $\mathcal{S}$ with those of a teacher trained on $\mathcal{D}_{\text{real}}$. Trajectory matching-based dataset distillation methods typically adopt a bilevel optimization scheme comprising an *inner loop* and an *outer loop*. The inner loop simulates the training dynamics of a student model on the current synthetic dataset $\mathcal{S}$, while the outer loop updates $\mathcal{S}$ such that the student's optimization trajectory closely matches that of a teacher trained on the real dataset $\mathcal{D}_{\text{real}}$.

Specifically, a teacher model is first trained on $\mathcal{D}_{\text{real}}$ to produce a reference trajectory $\tau^* = \{\theta_t^*\}_{t=1}^{T}$, where θ_t^* denotes the model parameters at iteration t. Meanwhile, the synthetic dataset $\mathcal{S}$ is initialized—either by sampling from Gaussian noise or selecting real samples—with corresponding soft or hard labels.

In the *inner loop*, we simulate student learning dynamics by iteratively updating parameters using $\mathcal{S}$. Given a starting point $\hat{\theta}_t = \theta_t^*$, the student parameters are updated over N steps according to:

$$\hat{\theta}_{t+n+1} = \hat{\theta}_{t+n} - \alpha \nabla_{\hat{\theta}_{t+n}} \ell(\mathcal{A}_n(\mathcal{S}); \hat{\theta}_{t+n}), \quad n = 0, \ldots, N-1, \tag{2}$$

where ℓ denotes the task loss (e.g., cross-entropy), $\mathcal{A}(\mathcal{S})$ denotes a mini-batch drawn from $\mathcal{S}$ with optional data augmentation, and α is the learning rate. This simulates the student trajectory $\hat{\tau} = \{\hat{\theta}_{t+n}\}_{n=0}^{N}$ induced by $\mathcal{S}$.

In the *outer loop*, we optimize the synthetic dataset $\mathcal{S}$ such that the student's final parameters align with the teacher's future trajectory. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_{max}\}$ be a set of anchor steps. At each $t \in \mathcal{T}$, the teacher parameters after M additional steps are denoted as θ^*_{t+M}. In our method, $\mathcal{L}$ in Eq. 1 is instantiated as the following normalized trajectory alignment loss:

$$\mathcal{L}_{\text{traj}} = \frac{\left\|\hat{\theta}_{t+N} - \theta^*_{t+M}\right\|_2^2}{\left\|\theta^*_t - \theta^*_{t+M}\right\|_2^2}. \tag{3}$$

The normalization term $\left\|\theta^*_t - \theta^*_{t+M}\right\|_2^2$ calibrates the alignment error relative to the teacher's own parameter change, making the loss scale-invariant across steps. The numerator measures how closely the student, trained on $\mathcal{S}$, approximates the teacher's future parameters, while the denominator normalizes for the teacher's update magnitude to ensure scale invariance. This outer-loop objective is used to update $\mathcal{S}$ via gradient-based optimization, enabling the synthetic data to induce faithful learning dynamics that reflect those observed on real data.

4 Methodology

4.1 Overview

Figure 1 illustrates the overall workflow of our proposed framework for noisy dataset distillation, which aims to construct a compact synthetic dataset $\mathcal{S}$ from a noisy real-world dataset $\mathcal{D}_{\text{real}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$. The core idea is to train the student such that it replicates the learning dynamics of a teacher model trained on $\mathcal{D}_{\text{real}}$, while suppressing the adverse effects of label noise. To this end, our framework follows a three-stage pipeline. First, we train a teacher model on $\mathcal{D}_{\text{real}}$ to obtain a reference parameter trajectory $\tau^* = \theta^*_t$; during this stage, we apply *Selective Guidance Reweighting (SGR)* (cf. Sec. 4.2) to assign sample-specific weights based on reliability estimates derived from second-split forgetting and KNN-based local consistency, thereby producing a cleaner supervision signal. Second, we initialize a synthetic dataset $\mathcal{S}$ with soft or hard labels and simulate the inner-loop training dynamics of a student model on $\mathcal{S}$. In this stage, we further incorporate *Teacher-Inspired Auxiliary Targets (TIAT)* (cf. Sec. 4.3) —a set of auxiliary consistency signals extracted from intermediate teacher states over high-confidence samples — to boost supervision beyond trajectory alignment. These high-confidence real samples are used only to refine teacher checkpoints and construct auxiliary targets, they are not paired with, indexed against, or directly mapped to individual synthetic samples. Finally, in the outer loop, we optimize $\mathcal{S}$ via a normalized trajectory matching loss computed across multiple anchor steps, aligning the student's learned trajectory with that of the teacher. This bilevel optimization process jointly improves the quality of supervision (via

SGR) and the effectiveness of trajectory alignment (via TIAT), enabling the distilled dataset to retain cleaner, more generalizable knowledge even under noisy conditions.

4.2 Selective Guidance Reweighting

We propose **Selective Guidance Reweighting (SGR)** to control the fidelity of signals used for synthetic data optimization. SGR introduces a hybrid reliability estimator for each training sample i, composed of:

- a **dynamic consistency score** $W^{(i)}_{\text{knn}}$ based on feature-space neighborhood agreement.
- a **static prior score** $W^{(i)}_{\text{ssft}}$ derived from sample forgetting behavior;

These scores are combined into a unified weight $W^{(i)}$ via time-dependent convex interpolation.

Dynamic Estimation via KNN. Given the predicted label distributions $\hat{y}_j$ of the K nearest neighbors of the observed sample i in the feature space, we define its KNN consistency score as:

$$W^{(i)}_{\text{knn}} = 1 - \frac{1}{K}\sum_{j=1}^{K} \text{JS}(\tilde{y}_i, \hat{y}_j), \tag{4}$$

where $\tilde{y}_i$ is viewed as a one-hot distribution and $\text{JS}(\cdot,\cdot)$ denotes the Jensen–Shannon divergence. This score ranges in $[0, 1]$, with higher values indicating better local agreement.

Static Prior via SSFT. Inspired by [27], we introduce a forgetting-based difficulty score to quantify the reliability of each training sample. For each sample i, we record:

- $t^{(i)}_{\text{learn}}$: the earliest epoch when the sample is first correctly classified;
- $t^{(i)}_{\text{forget}}$: the earliest epoch after which it is misclassified again.

These timestamps reflect the memorization and retention behavior of a sample during training. To assess sample-level difficulty, we define the SSFT score as:

$$s^{(i)} = \lambda \cdot \frac{t^{(i)}_{\text{learn}}}{\max_j t^{(j)}_{\text{learn}}} + (1-\lambda)\cdot\left(1 - \frac{t^{(i)}_{\text{forget}}}{\max_j t^{(j)}_{\text{forget}}}\right), \tag{5}$$

where $\lambda \in [0, 1]$ balances the contributions of learnability and forgettability. A high $s^{(i)}$ implies that sample i was learned late and forgotten early, thus more likely to be noisy or difficult. We convert this difficulty score into a static reliability weight:

$$W^{(i)}_{\text{ssft}} = 1 - s^{(i)}, \tag{6}$$

such that clean and stable samples are assigned higher weights at the beginning of training. This reflects the global memorization difficulty of sample i: higher $W^{(i)}_{\text{ssft}}$ values favor clean and stable samples, while lower values tend to downweight potentially noisy or hard-to-learn examples.

Hybrid Weighting via Curriculum Fusion. To balance global priors and evolving local consistency, we define a convex combination:

$$W_t^{(i)} = (1 - \alpha_t) \cdot W_{\text{ssft}}^{(i)} + \alpha_t \cdot W_{\text{knn}}^{(i)}, \tag{7}$$

where $\alpha_t \in [0, 1]$ is a time-dependent blending coefficient that increases linearly over training epochs. Specifically, we set:

$$\alpha_t = \min\left(\frac{t}{T_{\text{warmup}}} \cdot \alpha_{\max}, \alpha_{\max}\right), \tag{8}$$

where T_{warmup} is a predefined transition period (e.g., 20% of training). Optionally, when training multiple teacher trajectories indexed by $p \in \{1, \ldots, P\}$, we set $\alpha_{\max}^{(p)} = \frac{p-1}{P}$ to diversify the static-dynamic balance across teachers (more details on multi-trajectory aggregation are given in Eq. 9).

Remark. The timestep weight $W_t^{(i)}$ is used to modulate the contribution of sample i when updating the teacher model parameters or computing guidance for $\mathcal{S}$ in outer-loop optimization. This strategy enables the distillation process to prioritize clean, informative signals throughout training.

4.3 Teacher-Inspired Auxiliary Targets (TIAT)

Although the Selective Guidance Reweighting (SGR) module effectively suppresses noisy signals along the teacher trajectory, its role is essentially confined to improving the quality of teacher-side guidance, and it lacks an explicit mechanism to guarantee that the distilled dataset $\mathcal{S}$ remains consistently aligned with clean supervision in the representation space. More concretely, SGR amplifies cleaner and more consistent teacher signals, thereby enhancing the reliability of the teacher trajectory itself, but it does not directly regulate how the student model absorbs and responds to these reweighted signals during trajectory alignment and parameter updates. As a consequence, even after reweighting, the student optimization can still overfit residual noise or locally unstable regions in the teacher trajectory, manifesting as decision boundary shifts and limited generalization, and thus failing to fully realize the theoretical benefits of noise suppression.

To address this issue, we propose **Teacher-Inspired Auxiliary Targets (TIAT)**, a complementary mechanism that operates on the student side. TIAT injects clean-aware regularization into the distillation process by identifying a reliable subset and modeling trajectory-level uncertainty, while further incorporating auxiliary supervision derived from intermediate, high-confidence teacher states. This design effectively regularizes the student's optimization, promoting more stable and consistent learning dynamics. By jointly optimizing trajectory alignment (Eq. 10) and uncertainty-aware auxiliary objectives (Eq. 12–Eq. 13), TIAT bridges the gap between teacher-signal refinement and student optimization stability, ensuring that the synthetic dataset learns predominantly from the most trustworthy stages of the teacher's knowledge evolution.

Probabilistic Confidence Aggregation. Let $W_p^{(i)}$ denote the final-stage reliability weight assigned to sample i by the p-th teacher trajectory, trained with static-dynamic blending coefficient $\alpha_{\max}^{(p)}$. To estimate the sample's overall confidence, we define:

$$\bar{W}_{\text{conf}}^{(i)} := \mathbb{E}_{p\sim\mathcal{U}(\mathcal{P})}\left[W_p^{(i)}\right] \approx \frac{1}{P}\sum_{p=1}^{P} W_p^{(i)}, \tag{9}$$

where $\mathcal{P} = \{1, \ldots, P\}$ is the index set of all trajectories. This aggregated score summarizes the expected reliability of sample i across all teacher views.

We then compute the probabilistic trajectory alignment loss $\mathcal{L}_{\text{match}}$ that encourages student updates to follow the aggregated teacher behavior:

$$\mathcal{L}_{\text{match}} = \frac{\|\hat{\theta}_{t+N}^{(i)} - \theta_{t+M}^*\|_2^2}{\|\theta_{t+M}^* - \theta_t^*\|_2^2}, \tag{10}$$

where $\hat{\theta}_{t+N}^{(i)}$ denotes the student parameters after being updated on synthetic sample i for N optimization steps starting from the teacher state at iteration t.

Uncertainty-Aware Auxiliary Regularization. To quantify confidence variance across trajectories, we compute:

$$\text{Var}_p(W_p^{(i)}) = \mathbb{E}_{p\sim\mathcal{U}(\mathcal{P})}\left[\left(W_p^{(i)} - \bar{W}_{\text{conf}}^{(i)}\right)^2\right]. \tag{11}$$

This variance reflects the stability of confidence scores assigned to sample i; low-variance samples are deemed more consistently reliable.

We then define an approximately reliable subset $\mathcal{D}_{\text{sub}}$ using two complementary criteria: $W_{\text{ssft}}^{(i)} \geq \delta_{\text{sup}}$ and $\text{Var}_p(W_p^{(i)}) \leq \sigma_{\text{inf}}$, where δ_{sup} and σ_{inf} are fixed thresholds. This set captures samples that are both statistically confident and dynamically stable, forming the basis of our auxiliary regularization. Based on $\mathcal{D}_{\text{sub}}$, we fine-tune θ_t^* to produce a cleaner teacher checkpoint θ_{t+M}^{ft} and define the auxiliary loss as:

$$\mathcal{L}_{\text{aux}} = \frac{\|\hat{\theta}_{t+N}^{(i)} - \theta_{t+M}^{\text{ft}}\|_2^2}{\|\theta_{t+M}^{\text{ft}} - \theta_t^*\|_2^2}. \tag{12}$$

To obtain the refined teacher checkpoint θ_{t+M}^{ft}, we perform a short fine-tuning of the teacher model using only the reliable subset $\mathcal{D}_{\text{sub}}$. This fine-tuning process retains the same learning protocol as the original teacher but is restricted to several additional epochs on high-confidence samples, effectively filtering out the influence of noisy labels. The resulting θ_{t+M}^{ft} thus represents a noise-suppressed continuation of the teacher trajectory.

To consolidate the complementary contributions of the original and refined trajectories, we formulate a unified objective function as follows:

$$\mathcal{L}_{\text{total}} = (1-\beta)\cdot\mathcal{L}_{\text{match}} + \beta\cdot\mathcal{L}_{\text{aux}}, \quad \beta \in [0,1], \tag{13}$$

where β balances the influence of the original teacher trajectory and the clean-adjusted auxiliary trajectory.

Discussion. TIAT introduces an uncertainty-aware auxiliary supervision signal into dataset distillation, driven by both trajectory consensus and variance-aware selection. By formulating confidence as a probabilistic mean and explicitly modeling uncertainty, the method enables soft guidance that avoids hard filtering while remaining computationally efficient. This module complements SGR by enforcing trajectory-level regularization from the student side, forming a closed-loop distillation pipeline that is robust to noisy supervision.

5 Experiments

5.1 Experimental Setup

We evaluate our method on two widely used image classification datasets, CIFAR-10/100 [17] and Tiny-ImageNet [18], under noisy-label settings. Comprehensive experiments are conducted across different noise regimes and images-per-class (IPC) configurations to assess the robustness and scalability of our approach in the context of noisy label learning (LNL) and dataset distillation (DD).

Noise Settings. Following prior work [9, 15, 34], we consider both synthetic and real-world label noise:

- **Symmetric Noise.** Labels are uniformly corrupted across all non-target classes. Given a noise rate η and C classes, the label transition is defined as
$$P(y^{\text{noisy}} = c' \mid y^{\text{clean}} = c) = \begin{cases} 1-\eta, & c' = c, \\ \frac{\eta}{C-1}, & c' \neq c, \end{cases}$$
where $c, c' \in \{1, \ldots, C\}$. We set $\eta \in \{20\%, 40\%\}$ in our experiments.
- **Asymmetric Noise.** Labels are flipped according to class-dependent transition rules, which reflect realistic confusions (e.g., *cat* $\leftrightarrow$ *dog*, *truck* $\rightarrow$ *automobile*). We adopt the standard asymmetric transition matrices used in prior LNL studies [15, 34].
- **Real-World Noise.** We additionally evaluate on CIFAR-N [41], a human-annotated variant of CIFAR-10/100 that exhibits natural labeling inconsistencies. Its multi-annotator design captures human ambiguity patterns, providing a realistic benchmark for noisy labels.

Implementation Details. We compare our method against the following baselines: (i) a *noisy baseline*, which trains directly on the corrupted dataset without any distillation; (ii) a *random subset selection* baseline, which constructs condensed datasets by uniformly sampling the same number of images per class from the noisy set under the given IPC budget; and (iii) state-of-the-art dataset distillation approaches, including DATM [12], DANCE [42], RDED [35], and RCIG [25]. To ensure a fair comparison, we follow the key experimental settings in prior work, particularly DATM. Specifically, we adopt a three-layer ConvNet backbone for CIFAR-10/100 and a four-layer ConvNet for Tiny-ImageNet in all methods. For evaluation, we report the mean and standard deviation of test accuracy over 5 independently initialized networks trained on the distilled dataset

Table 1: Test accuracy (%) on CIFAR-10 under (a)symmetric noise (20% and 40%).

Noise Type	Noise Ratio / Method	20% IPC=10	20% 50	20% 500	20% 1000	40% 10	40% 50	40% 500	40% 1000
Symmetric	**Full Dataset**	76.1±0.2				66.2±0.5			
	Subset	25.1±0.3	38.8±0.5	53.2±0.7	57.1±0.6	15.3±0.5	27.6±0.5	38.9±0.7	43.1±1.0
	RCIG [25]	67.1±0.3	71.5±0.3	72.0±0.2	-	65.6±0.3	67.8±0.5	68.9±0.3	-
	RDED [35]	52.3±0.8	67.3±0.9	74.9±0.4	-	52.2±0.8	62.3±0.3	65.9±0.4	-
	DANCE [42]	**68.7±0.4**	71.9±0.3	73.0±0.2	-	**65.8±0.2**	67.6±0.2	68.4±0.3	-
	DATM [12]	62.7±0.6	71.7±0.3	80.0±0.4	81.9±0.1	58.2±0.2	67.2±0.3	74.3±0.3	76.7±0.2
	Ours	64.5±0.3	**73.5±0.4**	**81.5±0.2**	**83.3±0.3**	64.8±0.5	**71.6±0.3**	**80.0±0.2**	**81.4±0.1**
Asymmetric	**Full Dataset**	80.2±0.2				71.1±0.4			
	Subset	30.0±0.4	44.1±0.2	63.2±0.9	68.8±0.3	25.2±0.3	40.0±1.1	59.8±0.6	64.8±1.1
	RCIG [25]	64.6±0.3	69.9±0.3	71.0±0.3	-	58.9±0.6	63.1±0.3	64.2±0.2	-
	RDED [35]	46.1±1.6	64.8±0.5	75.4±0.5	-	43.6±1.1	60.0±1.3	69.3±1.2	-
	DANCE [42]	**69.2±0.3**	73.4±0.1	73.6±0.4	-	**66.1±0.2**	**69.5±0.1**	69.6±0.2	-
	DATM [12]	61.9±0.4	71.3±0.4	77.7±0.3	80.4±0.3	55.5±0.3	64.0±0.3	69.7±0.5	71.4±0.3
	Ours	63.7±0.8	**73.6±0.3**	**79.8±0.3**	**82.1±0.3**	58.7±0.5	67.0±0.5	**73.2±0.4**	**75.0±0.4**

$\mathcal{S}$. Unless otherwise specified, we follow DATM to use $P = 100$ teacher trajectories. For SGR, we set the KNN neighborhood size to $K = 10$, the SSFT balance coefficient to $\lambda = 0.6$, and the warm-up length to $T_{warmup} = 60$. The SSFT timestamps are generated once before distillation and reused across experiments, introducing only one-time preprocessing overhead. For TIAT, we use a 60% high-confidence subset ratio and set $\beta = 0.1$, as validated in Table 7.

5.2 Benchmarking Dataset Distillation Results on Noisy Dataset

As shown in Table 1, across all four noise settings (symmetric/asymmetric at 20% and 40%), our method achieves the strongest overall robustness across noise settings, especially under higher noise rates and larger IPC budgets, while DANCE remains competitive in several low-IPC cases. The advantage is particularly pronounced in high-IPC regimes and under severe noise (40%), where competing methods degrade rapidly, while our method maintains both strong accuracy and smooth performance trends. Even under the challenging asymmetric noise setting, where existing dataset distillation methods typically incur substantial performance drops, our approach remains consistently superior at both 20% and 40% noise rates. These observations indicate that our clean-aware sample reweighting and trajectory-informed distillation substantially improve robustness to label corruption, especially when the training data is simultaneously scarce and noisy.

On real-world noisy labels (CIFAR-N, Table 2), most methods overfit as IPC increases, with accuracy saturating or even degrading. In contrast, trajectory-matching approaches (DATM and ours) remain robust and benefit more from larger IPC, indicating a stronger ability to extract clean supervision. As noise grows from Aggre (9%) to Worst (40%), all methods degrade, but our approach consistently leads, with the margin widening at higher IPC and exceeding the best baseline by over 3% from IPC 50 onward. These results demonstrate the

strong noise resilience and compression robustness of our framework across diverse noise levels and data scales.

Table 3 and Table 4 further compares our method with recent state-of-the-art approaches on **CIFAR-100N** and **CIFAR-100** under symmetric and asymmetric noise at 20% and 40% corruption ratios. Across all IPC levels, our method attains the best accuracy in nearly all settings. Notably, under 40% symmetric noise, we outperform prior methods by substantial margins, achieving **52.8%** at 100 IPC and **43.2%** at 10 IPC. Even in the challenging **Worst** case, our method still yields the highest result (**45.9%** at 50 IPC). These results further underscore the robustness of our clean-aware reweighting and trajectory-guided distillation design under severe label noise.

Extending this analysis, we evaluate our method on the larger-scale **Tiny-ImageNet** dataset under symmetric noise levels of 20% and 40% in Table 5. Even under this challenging setting, our approach consistently outperforms all baseline methods. While DATM already achieves strong performance (30.4% and 28.2%), our method further improves the accuracy to 30.8% and 30.1%, demonstrating enhanced robustness to label noise and more effective extraction of clean supervisory signals under extremely low-data regimes. These results indicate that our proposed clean-aware sample reweighting and trajectory-informed distillation not only generalize to CIFAR datasets but also remain effective on higher-resolution and more challenging datasets.

5.3 Ablation Studies

Impact of Each Component. We conduct an ablation study on CIFAR-10 under symmetric label noise levels of 20% and 40%, and evaluate performance across different data condensation settings (IPC $\in \{10, 50, 500, 1000\}$). Starting from the DATM baseline, we progressively add our two modules: *Selective*

Table 2: Test accuracy (%) on **CIFAR-10N**.

Noise Type	IPC / Method	10	50	500	1000	Full Dataset
Aggre (9.03%)	**Subset**	31.5±0.6	43.4±0.3	63.7±0.9	69.7±0.5	
	RCIG [25]	66.5±0.2	72.9±0.3	-	-	
	RDED [35]	48.4±1.1	65.4±1.9	77.2±0.4	-	
	DANCE [42]	**69.7±0.4**	**74.1±0.4**	74.6±0.3	-	82.5±0.4
	DATM [12]	63.7±0.5	73.2±0.2	80.4±0.2	83.0±0.3	
	Ours	65.2±0.4	73.5±0.3	**80.8±0.1**	**83.4±0.2**	
Rand1 (18%)	**Subset**	28.9±0.8	40.9±0.5	60.0±0.6	63.0±0.5	
	RCIG [25]	66.9±0.2	72.4±0.2	-	-	
	RDED [35]	49.2±1.5	67.0±0.8	77.2±1.0	-	
	DANCE [42]	**69.8±0.3**	73.3±0.4	73.8±0.3	-	79.4±0.1
	DATM [12]	63.3±0.5	72.5±0.3	79.4±0.3	81.8±0.3	
	Ours	64.5±0.5	**73.6±0.3**	**80.6±0.3**	**82.9±0.1**	
Worst (40.21%)	**Subset**	22.2±0.5	35.8±0.4	46.2±0.4	50.0±1.1	
	RCIG [25]	65.3±0.3	67.7±0.4	-	-	
	RDED [35]	50.6±1.3	65.2±1.4	72.5±0.8	-	
	DANCE [42]	**65.8±0.3**	68.5±0.2	68.8±0.1	-	67.1±0.2
	DATM [12]	58.8±0.6	67.8±0.3	73.2±0.4	75.3±0.5	
	Ours	62.4±0.6	**70.8±0.5**	**77.1±0.4**	**79.2±0.4**	

Table 3: Test accuracy (%) on **CIFAR-100** under (a)symmetric noise (20% and 40%).

Noise Type	Noise Ratio / Method	20% IPC=10	20% 50	20% 100	40% IPC=10	40% 50	40% 100
Symmetric	**Full Dataset**		48.9±0.4			39.9±0.2	
	Subset	12.4±0.4	21.1±0.3	26.2±0.3	9.0±0.2	14.6±0.3	17.3±0.3
	RCIG [25]	41.2±0.3	38.4±0.3	-	36.5±0.4	30.7±0.2	-
	DANCE [42]	45.8±0.2	48.0±0.3	47.5±0.2	39.7±0.2	42.0±0.4	42.6±0.3
	DATM [12]	45.1±0.1	49.7±0.3	48.9±0.2	40.6±0.4	45.1±0.4	44.4±0.3
	Ours	**45.8±0.3**	**51.6±0.1**	**55.7±0.2**	**43.2±0.2**	**48.4±0.4**	**52.8±0.2**
Asymmetric	**Full Dataset**		46.2±0.5			33.0±0.1	
	Subset	12.2±0.1	21.9±0.5	26.9±0.3	9.5±0.4	15.0±0.3	18.9±0.3
	RCIG [25]	38.9±0.4	37.5±0.2	-	28.7±0.4	27.9±0.3	-
	DANCE [42]	43.8±0.3	46.7±0.3	48.2±0.4	32.2±0.4	34.2±0.2	35.22±0.3
	DATM [12]	40.3±0.4	45.4±0.3	50.2±0.3	29.4±0.4	32.0±0.3	36.0±0.3
	Ours	**44.6±0.2**	**50.3±0.3**	**54.4±0.1**	**34.7±0.1**	**36.5±0.4**	**40.7±0.4**

Table 4: Test accuracy (%) on **CIFAR-100N**.

IPC	**RCIG** [25]	**DANCE** [42]	**DATM** [12]	Subset	Full Dataset	**Ours**
10	37.0±0.3	42.0±0.3	39.9±0.5	11.2±0.2	**44.4±0.3**	41.0±0.3
50	35.3±0.2	43.6±0.3	43.9±0.2	20.3±0.4	44.4±0.3	**45.9±0.1**

Guidance Reweighting (SGR) and *Teacher-Inspired Auxiliary Targets* (TIAT). As reported in Table 6, incorporating SGR yields consistent gains across all configurations under both symmetric and asymmetric noise. Under symmetric noise, the improvements are particularly notable in severe and low-IPC settings (e.g., +4.7% at 10 IPC with 40% noise), demonstrating the effectiveness of trajectory-level denoising when label corruption is substantial. A similar trend can be observed under asymmetric noise, where SGR consistently improves performance across IPC scales (e.g., +2.1% at 50 IPC with 20% noise and +1.9% at 1000 IPC with 40% noise). Although the absolute gains under asymmetric noise are generally smaller than those under symmetric noise, the improvements remain stable, suggesting that trajectory-level denoising is effective not only for random corruption but also for structured, class-dependent noise patterns. Introducing TIAT on top of SGR further boosts accuracy across both noise types. Under symmetric noise, the full model achieves substantial additional improvements (e.g., +5.7% at 500 IPC under 40% noise). Notably, under asymmetric noise, TIAT delivers even more pronounced relative gains in high-noise regimes (e.g., +3.7% at 1000 IPC with 40% noise), highlighting the benefit of uncertainty-filtered auxiliary supervision when noise exhibits systematic bias. Overall, the full model (**DATM + SGR + TIAT**) consistently outperforms the baseline

Table 5: Test accuracy (%) on **Tiny-ImageNet** under symmetric noise (IPC=10).

Noise	**RCIG** [25]	**DANCE** [42]	**DATM** [12]	Subset	Full Dataset	**Ours**
20%	21.5±0.2	22.0±0.2	30.4±0.3	4.4±0.2	30.0±0.3	**30.8±0.2**
40%	17.9±0.2	17.3±0.3	28.2±0.2	3.0±0.1	22.6±0.5	**30.1±0.4**

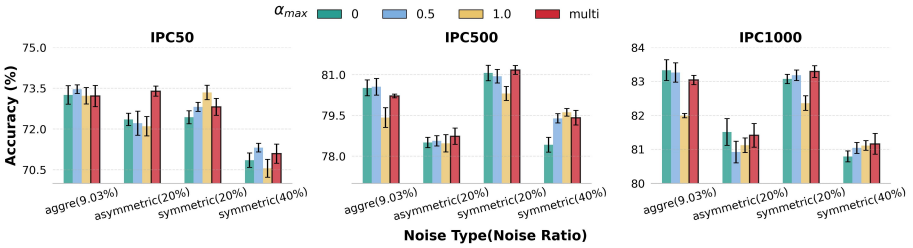

Fig. 2: Performance comparison between *Diverse Sampling* and *Fixed Sampling* under various settings on **CIFAR-10**. **Baseline** refers to DATM without our method. **Diverse Sampling** and **Fixed Sampling** incorporate our **Selective Guidance Reweighting** into DATM, where Fixed Sampling uses fixed $\alpha_{\max}$ values (**0, 0.5, 1.0**) for teacher trajectory training, while Diverse Sampling **multi** follows the strategy outlined in Section4.2.

across all IPC and noise settings. SGR provides strong and stable improvements in both random and structured corruption scenarios, while TIAT tends to contribute larger relative gains under higher noise ratios and more challenging regimes. These results confirm the complementary roles of trajectory-level denoising and teacher-side auxiliary supervision, and demonstrate the robustness of our framework to varying noise intensities, corruption structures, and condensation scales.

Effectiveness of Diverse Sampling. To evaluate the robustness of progressive trajectory diversity, we compare *Diverse Sampling*—where P teacher trajectories are trained with $\alpha_{\max}$ values uniformly distributed in $[0, 1]$—against a *Fixed Sampling* baseline with a constant $\alpha_{\max}$. As shown in Fig. 2, under increasing noise levels, the Diverse Sampling strategy consistently outperforms all fixed-α configurations and maintains stable performance without noticeable degradation. In contrast, fixed-α trajectories exhibit varying sensitivity to the noise rate, revealing limited robustness and poor generalization. These results indicate that injecting diversity into teacher signal strength substantially enhances resilience to label corruption and yields more robust, adaptive supervision for dataset distillation.

Effect of High-Confidence Subset Ratio and β Coefficient. We examine how the proportion of high-confidence samples used to construct the subset $\mathcal{D}_{\text{sub}}$ and the auxiliary loss weight β influence distillation performance under 20% and 40% symmetric noise on CIFAR-10 (IPC = 50). As shown in Table 7, our method remains robust across a broad range of subset ratios (50%–70%)

Table 6: Ablation on CIFAR-10 under (a)symmetric noise (20%, 40%) across IPCs.

Noise Type	Noise Ratio / Method	20%				40%			
		IPC=10	50	500	1000	IPC=10	50	500	1000
Symmetric	**DATM**	62.7	71.7	80.0	81.9	58.2	67.2	74.3	76.7
	+ **SGR**	63.8 (↑1.1)	72.8 (↑1.1)	81.2 (↑1.2)	83.3 (↑1.4)	62.9 (↑4.7)	71.1 (↑3.9)	79.4 (↑5.1)	81.2 (↑4.5)
	+ **SGR + TIAT**	**64.5** (↑1.8)	**73.5** (↑1.8)	**81.5** (↑1.5)	**83.3** (↑1.4)	**64.8** (↑6.6)	**71.6** (↑4.4)	**80.0** (↑5.7)	**81.4** (↑4.7)
Asymmetric	**DATM**	61.9	71.3	77.7	80.4	55.9	63.9	69.7	71.4
	+ **SGR**	63.1 (↑1.2)	73.4 (↑2.1)	78.7 (↑1.0)	81.4 (↑1.0)	56.7 (↑0.8)	65.7 (↑1.8)	71.5 (↑1.8)	73.3 (↑1.9)
	+ **SGR + TIAT**	**63.8** (↑1.9)	**73.6** (↑2.3)	**79.8** (↑2.1)	**82.1** (↑1.7)	**58.7** (↑2.8)	**67.1** (↑3.2)	**73.2** (↑3.5)	**75.1** (↑3.7)

Table 7: Ablation of β under symmetric noise (20%, 40%) on CIFAR-10. Top two accuracies are in **bold**, and the default β is highlighted.

β \ Subset	Symmetric (20%)			Symmetric (40%)		
	50%	60%	70%	50%	60%	70%
$\beta = 0.1$	73.3±0.11	**73.5±0.38**	73.3±0.4	71.9±0.2	71.5±0.4	**71.8±0.1**
$\beta = 0.5$	72.5±0.29	73.4±0.12	**74.0±0.3**	71.7±0.3	**72.2±0.3**	71.4±0.2
$\beta = 1.0$	67.4±0.12	69.6±0.17	71.6±0.3	70.0±0.3	71.7±0.2	71.5±0.1

and β values. While $\beta = 0.5$ attains the highest accuracy in some cases, $\beta = 0.1$ consistently delivers stable performance across both noise levels, especially when combined with a 60% high-confidence subset. In contrast, setting $\beta = 1.0$ tends to degrade performance, as it overemphasizes the auxiliary objective at the expense of trajectory alignment. Based on these observations, we adopt $\beta = 0.1$ and a 60% subset ratio as default hyperparameters in all main experiments.

Efficiency Analysis. Table 8 reports runtime and peak GPU memory on CIFAR-10 using one RTX 3090. Since both DATM and our method use the same TESLA-style memory-saving implementation, SGR/TIAT introduce negligible additional peak memory; the main cost is training-time overhead from reliability estimation and auxiliary target construction.

Table 8: Runtime and peak GPU memory analysis.

IPC	Peak Mem.	DATM Time	Ours Time	Overhead
10	2.5GB	6.5h	7.0h	1.08×
50	6.2GB	17h	25h	1.47×
500	10.4GB	24h	36h	1.50×
1000	11.6GB	60h	70h	1.17×

6 Conclusion

We investigate the underexplored problem of dataset distillation under noisy-label settings and identify two key challenges: overfitting to label noise and limited capacity of the synthetic set to retain clean signals. To address these, we introduce Selective Guidance Reweighting and Teacher-Inspired Auxiliary Targets. Experiments on benchmark datasets validate the robustness and effectiveness of our approach, paving the way for future research in noise-resilient dataset distillation.

Acknowledgment

This work has been supported in part by the National Natural Science Foundation of China (Grant No. 62472139, 62502142, 62502144), the Natural Science Foundation of Anhui Province (Grant No. 2508085QF226), the Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2403), Zhejiang University. The computation is completed on the HPC Platform of Hefei University of Technology.

References

[1] Bahri, D., Jiang, H., Gupta, M.: Deep k-nn for noisy labels. In: International Conference on Machine Learning. pp. 540–550. PMLR (2020)

[2] Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories (2022), https://arxiv.org/abs/2203.11932

[3] Chen, X., Yang, Y., Wang, Z., Mirzasoleiman, B.: Data distillation can be like vodka: Distilling more times for better quality (2023), https://arxiv.org/abs/2310.06982

[4] Cheng, L., Chen, K., Li, J., Tang, S., Zhang, S., Wang, M.: Dataset distillers are good label denoisers in the wild. arXiv preprint arXiv:2411.11924 (2024)

[5] Cui, J., Wang, R., Si, S., Hsieh, C.J.: Dc-bench: Dataset condensation benchmark (2022), https://arxiv.org/abs/2207.09639

[6] Deng, W., Li, W., Ding, T., Wang, L., Zhang, H., Huang, K., Huo, J., Gao, Y.: Exploiting inter-sample and inter-feature relations in dataset distillation (2024), https://arxiv.org/abs/2404.00563

[7] Di Salvo, F., Doerrich, S., Rieger, I., Ledig, C.: An embedding is worth a thousand noisy labels. arXiv preprint arXiv:2408.14358 (2024)

[8] Du, J., Jiang, Y., Tan, V.Y.F., Zhou, J.T., Li, H.: Minimizing the accumulated trajectory error to improve dataset distillation (2023), https://arxiv.org/abs/2211.11004

[9] Englesson, E., Azizpour, H.: Robust classification via regression for learning with noisy labels. In: ICLR 2024-The Twelfth International Conference on Learning Representations, Messe Wien Exhibition and Congress Center, Vienna, Austria, May 7-11t, 2024 (2024)

[10] Fang, C., Cheng, L., Mao, Y., Zhang, D., Fang, Y., Li, G., Qi, H., Jiao, L.: Separating noisy samples from tail classes for long-tailed image classification with label noise. IEEE Transactions on Neural Networks and Learning Systems **35**(11), 16036–16048 (2023)

[11] Fang, C., Wang, Q., Cheng, L., Gao, Z., Pan, C., Cao, Z., Zheng, Z., Zhang, D.: Reliable mutual distillation for medical image segmentation under imperfect annotations. IEEE Transactions on Medical Imaging **42**(6), 1720–1734 (2023)

[12] Guo, Z., Wang, K., Cazenavette, G., Li, H., Zhang, K., You, Y.: Towards lossless dataset distillation via difficulty-aligned trajectory matching (2024), https://arxiv.org/abs/2310.05773

[13] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)

[14] He, Y., Xiao, L., Zhou, J.T., Tsang, I.: Multisize dataset condensation (2024), https://arxiv.org/abs/2403.06075

[15] Iscen, A., Valmadre, J., Arnab, A., Schmid, C.: Learning with neighbor consistency for noisy labels. 2022 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4662–4671 (2022)

[16] Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International conference on machine learning. pp. 2304–2313. PMLR (2018)
[17] Krizhevsky, A.: Learning multiple layers of features from tiny images (2009), https://api.semanticscholar.org/CorpusID:18268744
[18] Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
[19] Lee, Y., Chung, H.W.: Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching (2024), https://arxiv.org/abs/2406.18561
[20] Li, J., Li, G., Liu, F., Yu, Y.: Neighborhood collective estimation for noisy label identification and correction. In: European Conference on Computer Vision. pp. 128–145. Springer (2022)
[21] Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
[22] Li, W., Wang, L., Li, W., Agustsson, E., Gool, L.V.: Webvision database: Visual learning and understanding from web data (2017), https://arxiv.org/abs/1708.02862
[23] Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems **33**, 20331–20342 (2020)
[24] Liu, Y., Guo, H.: Peer loss functions: Learning from noisy labels without knowing noise rates. In: International conference on machine learning. pp. 6226–6236. PMLR (2020)
[25] Loo, N., Hasani, R., Lechner, M., Rus, D.: Dataset distillation with convexified implicit gradients (2023), https://arxiv.org/abs/2302.06755
[26] Lyu, Y., Tsang, I.W.: Curriculum loss: Robust learning and generalization against label corruption. arXiv preprint arXiv:1905.10045 (2019)
[27] Maini, P., Garg, S., Lipton, Z., Kolter, J.Z.: Characterizing datapoints via second-split forgetting. Advances in Neural Information Processing Systems **35**, 30044–30057 (2022)
[28] Malach, E., Shalev-Shwartz, S.: Decoupling" when to update" from" how to update". Advances in neural information processing systems **30** (2017)
[29] Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. CoRR **abs/2107.13034** (2021), https://arxiv.org/abs/2107.13034
[30] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014)
[31] Sachdeva, N., McAuley, J.: Data distillation: A survey (2023), https://arxiv.org/abs/2301.04272
[32] Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learning an explicit mapping for sample weighting. Advances in neural information processing systems **32** (2019)

[33] Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International conference on machine learning. pp. 5907–5915. PMLR (2019)
[34] Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems **34**(11), 8135–8153 (2022)
[35] Sun, P., Shi, B., Yu, D., Lin, T.: On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
[36] Tu, Y., Zhang, B., Li, Y., Liu, L., Li, J., Wang, Y., Wang, C., Zhao, C.R.: Learning from noisy labels with decoupled meta label purifier. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19934–19943 (2023)
[37] Wang, S., Yang, Y., Liu, Z., Sun, C., Hu, X., He, C., Zhang, L.: Dataset distillation with neural characteristic function: A minmax perspective (2025), https://arxiv.org/abs/2502.20653
[38] Wang, T., Huan, J., Li, B.: Data dropout: Optimizing training data for convolutional neural networks. In: 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI). pp. 39–46. IEEE (2018)
[39] Wang, T., Zhu, J., Torralba, A., Efros, A.A.: Dataset distillation. CoRR **abs/1811.10959** (2018), http://arxiv.org/abs/1811.10959
[40] Wang, Y., Cheng, L., Duan, M., Wang, Y., Feng, Z., Kong, S.: Improving knowledge distillation via regularizing feature direction and norm. In: European Conference on Computer Vision. pp. 20–37. Springer (2024)
[41] Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with noisy labels revisited: A study using real-world human annotations. arXiv preprint arXiv:2110.12088 (2021)
[42] Zhang, H., Li, S., Lin, F., Wang, W., Qian, Z., Ge, S.: Dance: Dual-view distribution alignment for dataset condensation (2024), https://arxiv.org/abs/2406.01063
[43] Zhang, H., Li, S., Wang, P., Zeng, D., Ge, S.: M3d: Dataset condensation by minimizing maximum mean discrepancy (2024), https://arxiv.org/abs/2312.15927
[44] Zhang, T., Xue, M., Zhang, J., Zhang, H., Wang, Y., Cheng, L., Song, J., Song, M.: Generalization matters: Loss minima flattening via parameter hybridization for efficient online knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20176–20185 (2023)
[45] Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)
[46] Zhao, B., Bilen, H.: Dataset condensation with distribution matching (2022), https://arxiv.org/abs/2110.04181
[47] Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching (2021), https://arxiv.org/abs/2006.05929

[48] Zhou, T., Wang, S., Bilmes, J.: Robust curriculum learning: from clean label detection to noisy label self-correction. In: International Conference on Learning Representations (2021)
[49] Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression (2022), https://arxiv.org/abs/2206.00719
[50] Zhou, Y., Li, X., Liu, F., Wei, Q., Chen, X., Yu, L., Xie, C., Lungren, M.P., Xing, L.: L2b: Learning to bootstrap robust models for combating label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23523–23533 (2024)
[51] Zhu, Z., Dong, Z., Liu, Y.: Detecting corrupted labels without training a model to predict. In: International conference on machine learning. pp. 27412–27427. PMLR (2022)