SMC: Hessian-Aware One-Shot Dataset Distillation

Heng Shu[1,2]
[1]Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
shuheng2025@iscas.ac.cn

Qiming Yang[1*]
[1]Institute of Software, Chinese Academy of Sciences, Beijing, China
yangqiming@iscas.ac.cn

Yuquan Wu[1*]
[1]Institute of Software, Chinese Academy of Sciences, Beijing, China
yuquan@iscas.ac.cn

***Abstract*—Dataset distillation addresses the resource-intensity of deep learning by compressing a large training set into a small, high-fidelity synthetic dataset, significantly reducing memory footprint and training time while preserving competitive performance. One-shot dataset distillation extends this concept by generating a single, versatile synthetic set adaptable to varying Images-Per-Class (IPC) budgets, eliminating the need for repeated runs in scenarios with diverse requirements. However, existing one-shot methods are hampered by optimization instability, manifested as persistent and spatially unbalanced pixel updates, and tend to yield synthetic samples with blurred contours. To overcome these geometric and visual limitations, we propose Stabilized Multi-dimensional Condensation (SMC), consisting of three synergistic components: a Hessian-aware curvature regularization for landscape smoothing, a dynamic hybrid loss enhanced by a learned sampler for a balanced integration of feature stability and distributional fidelity, and a staged one-shot condensation schedule to coordinate the global optimization process. Extensive experiments across multiple datasets demonstrate that SMC delivers significant and consistent improvements. Notably, on CIFAR-10, SMC achieves an average accuracy of 60.6%, outperforming the state-of-the-art baseline by 2.2% on average. Code will be made available upon publication.**

***Index Terms*—Dataset Distillation, Hessian-Aware Curvature Regularization, Dynamic Loss Reweighting, One-shot Condensation**

I. Introduction

Dataset distillation, also known as dataset condensation, aims to replace a large-scale training corpus $\mathcal{R}$ with a compact synthetic set $\mathcal{S}$. The core objective is to train models on $\mathcal{S}$ that achieve comparable performance to those trained on the original $\mathcal{R}$ [1], thereby largely improving efficiency by reducing computational and storage burdens. Building upon this foundation, one-shot dataset distillation advances the paradigm by generating a single, versatile synthetic dataset (dubbed as *mother set*) through one distillation run. The most important advantage of the one-shot method is its adaptability: from the *mother set*, effective subsets can be extracted to meet different Images-Per-Class (IPC) budgets without the need for repeated, costly distillation procedures. It is particularly useful for privacy-preserving applications and deployments constrained by bandwidth or storage, as well as for edge and federated patterns.

Fig. 1. Visual comparison of instance-level feature matching($\mathcal{L}_\mathrm{P}$) vs. distribution matching($\mathcal{L}_\mathrm{D}$) for one-shot dataset distillation.

Traditional dataset distillation generally falls into two principal paradigms: instance-level (or pair-wise) feature matching and distribution matching. Instance-level approaches minimize distances between corresponding image batches of real and synthetic samples. While computationally efficient, the inherent variability among instances often disrupts the synthesis process, blurring structural contours—a critical shortcoming given that perceptual research highlights shape bias (as opposed to texture bias) as essential for model robustness [2]. In contrast, distribution-matching methods align global statistics (e.g., via Maximum Mean Discrepancy) to better preserve object contours, yielding more visually faithful samples, though they often suffer from low discriminative density under tight budgets. Based on that, one-shot dataset distillation introduces a hierarchical scheduling strategy to generate a single, scalable synthetic set for varying IPC requirements. However, since it inherits these static objectives without adaptation, current one-shot methods fail to reconcile structural fidelity with optimization stability. This limitation manifests in the pronounced visual degradation and geometric instability that we analyze

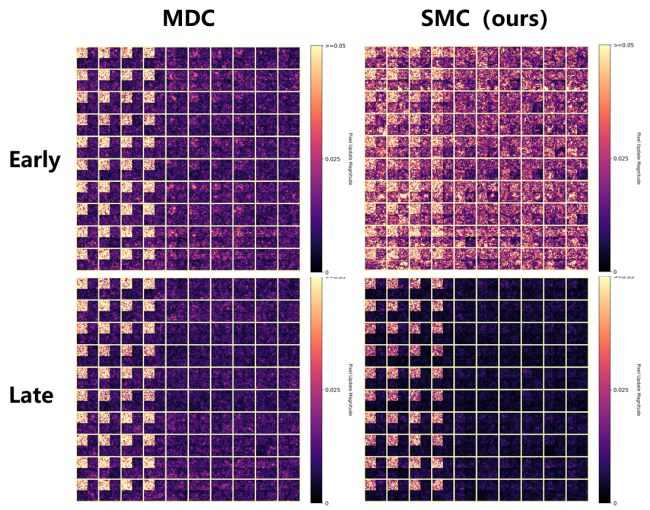

Fig. 2. **Visualizing Optimization Dynamics.** Comparison of pixel updates ($|\mathbf{x}_t - \mathbf{x}_{t-100}|$) at Early and Late stages. MDC (Left) persistently concentrates updates on primary (top-left) patches throughout distillation, leading to imbalance. In contrast, SMC (Right) exhibits more uniform and aggressive exploration initially, yet achieves superior stability and convergence in the late stage via Hessian regularization. Quantitative verification is provided in Sec. IV.

in the following section.

As shown in **Fig. 1**, we find that using either instance-level ($\mathcal{L}_{\mathrm{P}}$) or distribution-wise ($\mathcal{L}_{\mathrm{D}}$) metrics in isolation within the one-shot setting amplifies their respective weaknesses: instance-level feature matching ($\mathcal{L}_{\mathrm{P}}$) tends to produce synthetic samples suffering contour degradation, especially in large-IPC regimes. In contrast, distribution matching ($\mathcal{L}_{\mathrm{D}}$) better preserves contour structure but reduces per-sample informativeness under tight, small-IPC budgets—a limitation we empirically verify in our ablation study (Sec. IV). Beyond these static visual artifacts, we uncover a more fundamental instability during the distillation process. As visualized in the pixel update heatmap (**Fig. 2**), the cutting-edge one-shot method exhibits patch-wise update imbalance. Instead of transitioning from a global exploration phase to stable convergence, the trajectory of the current one-shot method persistently concentrates high-frequency updates on primary (top-left) sub-patches throughout the entire distillation process. This ceaseless oscillation in specific regions indicates that the optimization fails to settle into a stable solution, suggesting the synthesizer is trapped in sharp minima while leaving secondary patches under-optimized. To address both the visual deficiencies and geometric instability, a redesign of the optimization objectives is needed. Conceptually, merging the strengths of instance-level and distribution-level matching offers a promising path to restore contour fidelity, akin to strategies in NRR-DD [3] and ROME [4]. However, simply combining $\mathcal{L}_{\mathrm{P}}$ and $\mathcal{L}_{\mathrm{D}}$ in a one-shot setting is non-trivial: it involves training sets of different scales, where naive data truncation leads to information loss, while independent sampling leads to stochastic optimization conflicts and induces cross-size interference. Furthermore, the patch-wise update imbalance necessitates a mechanism

Algorithm Schematic.png

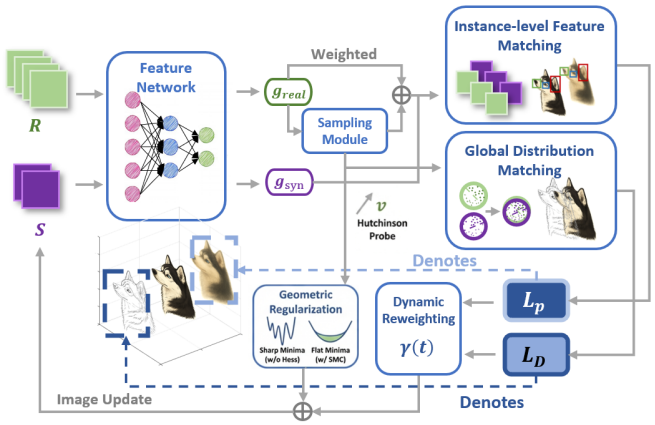

Fig. 3. SMC Framework Schematic

beyond simple loss matching. The persistent high-frequency oscillation indicates that standard objectives fail to constrain the optimization curvature, requiring explicit regularization to guide the synthesis process away from unstable, sharp minima.

Based on these observations, we argue that the key to producing high-quality one-shot distilled datasets lies in multi-scale structural alignment and geometric smoothing of the loss landscape. A robust solution must therefore: (i) move beyond simple instance matching by integrating multi-order distributional constraints to restore contour fidelity, and (ii) explicitly regularize the optimization curvature to converge toward flat minima, thereby rectifying the update imbalance and ensuring optimization stability under stochastic subsampling.

To fulfill these requirements, we present **S**tabilized **M**ulti-dimensional **C**ondensation (SMC), a unified framework constructed upon three strategic pillars: (1) a *Hessian-aware curvature regularization* that utilizes Hutchinson's estimator [5] to explicitly smooth the loss landscape, ensuring robustness against subsampling; (2) a *dynamic hybrid loss integrated with a learned compensation sampler*, which adaptively anneals between instance-level and distributional objectives while correcting gradient bias via adversarial reweighting; and (3) a *staged one-shot condensation schedule* that employs a coarse-to-fine strategy to generate a single, scalable synthetic set adaptable to varying IPC budgets. Fig. 3 provides an overview of the SMC pipeline.

In summary, our contributions are:

- We propose a Hessian-aware curvature regularization technique utilizing Hutchinson's trace estimator. This explicitly rectifies the patch-wise update imbalance, guiding the synthesis towards flat minima and significantly enhancing one-shot stability.
- We design a dynamic multi-order loss reweighting strategy paired with a learned compensation sampler, which jointly reconciles low-order instance convergence with high-order distribution matching to restore structural fidelity.
- We introduce a staged one-shot condensation schedule to

coordinate the global optimization, producing a unified synthetic set which consistently delivers superior performance across a wide range of IPC budgets, especially in challenging low-data regimes.

Extensive experiments demonstrate that SMC outperforms previous state-of-the-art methods on standard benchmarks. On CIFAR-10, SMC improves average downstream accuracy by **2.2%** overall, with a notable gain of **4.6%** in the IPC_3–IPC_6 range. Moreover, on the more tricky CIFAR-100 dataset, SMC achieves an average accuracy of **38.5%**, surpassing other counterparts by **1.0%**, demonstrating superior scalability to complex distributions. Quantitative analysis (Fig. 5) further verifies that SMC effectively mitigates the patch-wise update imbalance.

II. RELATED WORK

A. Gradient/Distribution Matching.

These approaches optimize synthetic data by aligning specific statistical targets between real and synthetic sets, focusing on either feature distributions or parameter gradients. Methods such as DM [6], NCFM [7], and CAFE [8] primarily match feature or model-distribution statistics. Zhao et al. propose DC [9] (Dataset Condensation), which shifts the focus to matching the gradients of network weights relative to real data. DSA [10] extends this with differentiable data augmentation: applying the same random transforms to both real and synthetic images lets the synthetic data inherit augmentation priors from real data. While such methods strike a good balance between efficiency and accuracy [11], an accuracy gap remains when the distilled set is extremely small.

B. Trajectory Matching.

Some works match network behavior on synthetic vs. real data (FTD [12], MAT [13], MKDT [14]). MTT [15] optimizes synthetic data so that models trained on them follow the same parameter trajectory as models trained on real data. MTT significantly outperforms previous methods and can even distill higher-resolution datasets. Nevertheless, it requires storing large trajectory data and extensive computation, making it less efficient and hard to extend to multi-size scenarios.

C. Representative/One-shot Distillation Methods.

These works address sample selection or size flexibility, similar to the **Once-for-All** (OFA) paradigm [16]. DREAM [17] selects representative real samples instead of random ones, reducing gradient bias and accelerating training. IDC [18] generates more synthetic data within a fixed storage budget via multi-formalism frameworks. Multisize Dataset Condensation (MDC) [19] compresses N distillation tasks into one: it adds an adaptive subset loss so the synthetic set works well at multiple sizes. MDC thus needs only one condensation run and reuses images, saving space. As discussed earlier, MDC relies on implicit regularization through subset sampling, which we identify as a source of optimization instability, leading to unbalanced updates and convergence to sharp minima.

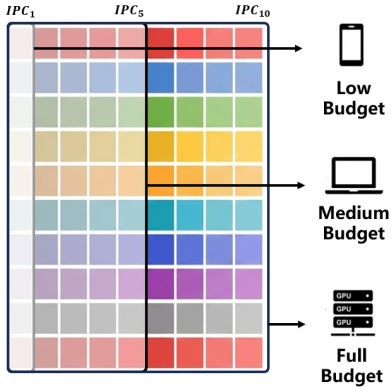

Fig. 4. **Concept of Scalable One-Shot Distillation.** Our framework generates a unified *mother set* that structurally embeds optimal subsets for varying capacities. This enables **budget-aware deployment**: the same condensed source can be flexibly tailored for diverse scenarios, ranging from resource-constrained edge devices (Low Budget) to high-performance clusters (Full Budget)

D. Optimization Landscape and Flat Minima.

The geometry of the loss landscape is closely linked to generalization. Hochreiter et al. [20] and works on Sharpness-Aware Minimization (SAM) [21] demonstrate that "flat minima" (regions with low curvature) generalize better than sharp minima. While SAM minimizes the curvature of model parameters θ, to our knowledge, this work is among the first to identify that the instability of one-shot subset sampling stems from sharp minima, and propose an explicit Hutchinson-based Hessian regularization to resolve this specific dilemma. We demonstrate that encouraging flat minima in the data synthesis space plays a critical role in stabilizing one-shot distillation, an aspect that has not been extensively explored in existing methods.

Although these advances improve specific aspects, existing methods still struggle to jointly capture full distributions, ensure geometric stability against sharp minima, and adapt to multiple target sizes.

III. METHODOLOGY

We propose **S**tabilized **M**ulti-dimensional **C**ondensation (SMC) to address the challenges mentioned before in dataset distillation, particularly in balancing low-order instance matching and high-order structural alignment, and resolving the geometric instability and patch-wise update imbalance in one-shot dataset distillation. As illustrated in Fig. 3, the framework is built upon three strategic pillars: (i) geometric regularization via Hessian-aware smoothing, (ii) adaptive multi-order alignment realized by a dynamic hybrid loss and a learned compensation sampler, and (iii) a staged one-shot schedule.

A. Hessian-Aware Curvature Regularization

We treat the subset selection for different IPC budgets as structured perturbations in the synthetic data space. A critical failure mode in one-shot distillation is that synthetic data often converge to sharp minima, where the loss landscape has high curvature. In such regions, slight perturbations—such as subset sampling for different IPC budgets—cause drastic

spikes in validation error. To address this, we introduce a regularization term that explicitly minimizes the curvature of the optimization landscape. This explicit smoothing eliminates the sharp minima responsible for the persistent pixel-level oscillations observed in the baseline.

The curvature is characterized by the trace of the Hessian matrix of the task loss $\mathcal{L}_{\text{task}}$ with respect to the synthetic data $\mathcal{S}$. Since computing the exact Hessian trace is computationally prohibitive ($O(|\mathcal{S}|^2)$), we employ the efficient *Hutchinson's Trace Estimator* [5]. We define the Hessian regularization term as

$$\mathcal{R}_{\text{Hess}}(\mathcal{S}) = \text{Tr}\big(\nabla^2_{\mathcal{S}}\mathcal{L}_{\text{task}}(\mathcal{S})\big) \approx \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_k^\top \nabla^2_{\mathcal{S}}\mathcal{L}_{\text{task}}(\mathcal{S})\mathbf{v}_k, \tag{1}$$

where K is the number of probe vectors (set to 5), and $\mathbf{v}_k$ are random vectors drawn from a Rademacher distribution. The Hessian-vector product (HVP) is computed efficiently via automatic differentiation: $\nabla^2_{\mathcal{S}}\mathcal{L}_{\text{task}}\,\mathbf{v} = \nabla_{\mathcal{S}}\big(\nabla_{\mathcal{S}}\mathcal{L}_{\text{task}}\cdot\mathbf{v}\big)$.

Per-Channel Selective Pruning. Rather than minimizing the trace globally across all dimensions, we estimate a per-channel trace τ_c for each image channel c and apply a binary mask M_c to regularize only the dimensions with the highest curvature (Top-$k\%$):

$$\mathcal{R}_{\text{Hess}} = \sum_{c=1}^{C} M_c\,\tau_c, \qquad M_c = \mathbb{I}\big(\tau_c \in \text{Top-}k\%\big(\{\tau_j\}_{j=1}^{C}\big)\big), \tag{2}$$

the final objective is

$$\mathcal{L}_{\text{total}}(\mathcal{S}) = \mathcal{L}_{\text{task}}(\mathcal{S}) + \lambda\,\mathcal{R}_{\text{Hess}}(\mathcal{S}). \tag{3}$$

Although Hessian-vector products introduce additional overhead, the use of a small number of Hutchinson probes keeps the cost manageable in practice. In our implementation, the curvature regularization is applied periodically rather than at every iteration to further control computation.

B. Dynamic Hybrid Loss with Learned Compensation

To restore contour fidelity and structural details lost by simple instance matching, we design a dynamic hybrid loss function integrated with a learned compensation sampler.

Dynamic Hybrid Loss. The total loss $\mathcal{L}(t)$ integrates distribution matching $\mathcal{L}_{\text{D}}$ and instance-level matching $\mathcal{L}_{\text{P}}$:

$$\mathcal{L}_{\text{task}} = \gamma(t)\cdot\mathcal{L}_{\text{D}} + (1-\gamma(t))\cdot\mathcal{L}_{\text{P}}, \tag{4}$$

the weight $\gamma(t)$ increases linearly from γ_{start} to γ_{end} over $T_{\max}$ iterations, prioritizing stable convergence in the early stage of the distillation procedure.

Learned Compensation Sampler. We parameterize a lightweight sampling network s_ϕ to define an adversarial empirical distribution. The real examples are assigned weights w_i via a softmax operation on scores $s_\phi(x_i)$. Using a multi-bandwidth Gaussian kernel k, the weighted squared MMD loss is:

$$\begin{aligned}\widehat{\mathcal{L}}^2_{\text{D}} = &\frac{\sum_{i,i'} w_i w_{i'} k(x_i, x_{i'})}{(\sum_i w_i)^2} + \frac{1}{M^2}\sum_{j,j'} k(y_j, y_{j'}) \\ &- \frac{2}{M\sum_i w_i}\sum_{i,j} w_i k(x_i, y_j),\end{aligned} \tag{5}$$

the sampler is optimized via $\max_\phi \widehat{\mathcal{L}}^2_{\text{D}}$ to highlight under-represented modes. Simultaneously, we decompose the pair loss $\mathcal{L}_{\text{P}}$ into global ($\mathcal{R}$) and sampled ($w(\phi)$) components:

$$\mathcal{L}_{\text{P}} = \alpha\mathbb{E}_{x\sim\mathcal{R}}\|f(x)-f(y)\|^2 + (1-\alpha)\mathbb{E}_{x\sim w(\phi)}\|f(x)-f(y)\|^2, \tag{6}$$

this decomposition ensures the model captures both global feature consistency and local robustness against outliers.

C. Staged One-shot Condensation Schedule

The third component is the staged one-shot condensation schedule. This schedule coordinates the global synthesis process by transitioning from holistic structural alignment to fine-grained subset refinement, preventing the "forgetting" of global patterns while ensuring low-IPC distinctiveness.

- **Stage 1:** The synthesis initially targets the maximum IPC budget. In this phase, the optimization focuses on capturing low-frequency, global structures of the distribution to establish a stable geometric foundation.
- **Stage 2:** In the later stage, we iteratively sample sub-datasets corresponding to different IPC budgets from the mother set. The optimization explicitly targets these subsets to refine high-order details, ensuring that even minimal subsets retain high discriminative power.

D. Overall Optimization Procedure

Algorithm 1 outlines the complete SMC pipeline. The optimization alternates between inner-loop network training and outer-loop synthesis, integrating dynamic hybrid loss minimization, periodic Hessian-aware regularization, and adversarial sampler updates.

IV. EXPERIMENTS

A. Experiment Settings

Condensation Training. For CIFAR-10, CIFAR-100 [22], and SVHN [23], we adopt ConvNet-D3 [24] as the backbone for data condensation. The batch size is set to 128 for IPC ≤ 30 and 256 for IPC > 30. The condensed set is optimized following the IDC training protocol, where the synthetic images are updated through iterative matching between real and synthetic batches. The optimization is carried out for a fixed number of outer iterations, and in each iteration the network is re-initialized and trained for a small number of inner-loop steps.

Condensation Evaluation. Following IDC [18], we evaluate the condensed datasets by training networks from scratch using SGD with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. For CIFAR-10, CIFAR-100, and SVHN, the network is randomly initialized 2000 times;

Algorithm 1 SMC Optimization Procedure

Require: Real set $\mathcal{R}$, synthetic set $\mathcal{S}$, sampler s_ϕ, total iter T, Hessian interval r, sampler interval N_s, Hessian weight λ, annealing $\gamma(t)$, mix coeff α, pruning ratio k.

Initialize $\mathcal{S}$ (e.g., via random/mix) and sampler parameters ϕ.
for $t \leftarrow 1$ to T **do**
  if $t \bmod N_s = 0$ **then**
    Update sampler: $\phi \leftarrow \phi + \eta_\phi \nabla_\phi \widehat{\mathcal{L}}_{\mathrm{D}}^2$ (maximize weighted MMD).
  end if
  Sample real batch $\mathcal{B}_{\text{real}} \sim \mathcal{R}$ and weighted batch $\mathcal{B}_{\text{w}} \sim \mathcal{R}$ using s_ϕ.
  Train network θ on $\mathcal{B}_{\text{real}} \cup \mathcal{S}$ to obtain features/gradients.
  Compute decomposed pair loss: $\mathcal{L}_{\mathrm{P}} \leftarrow \alpha \| \ldots \|_{\mathcal{B}_{\text{real}}} + (1-\alpha) \| \ldots \|_{\mathcal{B}_{\text{w}}}$.
  Compute total task loss: $\mathcal{L}_{\text{task}} \leftarrow \gamma(t)\mathcal{L}_{\mathrm{D}} + (1-\gamma(t))\mathcal{L}_{\mathrm{P}}$.
  if $t \bmod r = 0$ **then**
    for $j \leftarrow 1$ to K (Probes) **do**
      Estimate per-channel Hessian trace τ_c via Hutchinson.
    end for
    Apply Top-$k\%$ mask: $\mathcal{R}_{\text{Hess}} \leftarrow \sum_c M_c \tau_c$.
  else
    $\mathcal{R}_{\text{Hess}} \leftarrow 0$.
  end if
  Update $\mathcal{S} \leftarrow \mathcal{S} - \eta_{\mathcal{S}} \nabla_{\mathcal{S}} (\mathcal{L}_{\text{task}} + \lambda \mathcal{R}_{\text{Hess}})$.
  (Implicitly handled by $\gamma(t)$ scheduling and subset sampling logic).
end for
return Optimized synthetic set $\mathcal{S}$.

for each initialization, it is trained for 100 epochs on the condensed set. All reported results are averaged over these independent runs.

Distillation Parameters. The specific hyperparameters used in our experiments are summarized as follows:

- We employ a dynamic weighting strategy between the MSE-based matching loss and the MMD loss. The weight γ is linearly scheduled from 0 to 0.8 throughout the condensation process. The mixing coefficient α in the prototype interpolation is fixed to 0.2.
- For Hessian-aware curvature regularization, we set the regularization strength to $\lambda = 10^{-4}$ and estimate the trace of the input Hessian using $K = 5$ Hutchinson probes. The regularization term is applied every 10 outer iterations to balance computational cost and optimization stability. Unless otherwise specified, the per-channel keep ratio for Top-$k\%$ pruning is fixed to $k = 50$.

These hyperparameters were fixed across datasets for all main experiments; a more exhaustive sensitivity analysis is left for future work.

TABLE I
COMPARISON WITH SOTA METHODS ON CIFAR-10.

	DSA	MTT	IDC	Dream	MDC	**Ours**
IPC_1	16.8	18.8	27.5	32.5	49.7	**50.9**
IPC_2	21.2	24.9	38.5	39.6	54.6	**54.3**
IPC_3	26.8	31.9	45.3	48.2	53.9	**59.2**
IPC_4	30.2	38.1	50.9	53.8	54.6	**60.4**
IPC_5	33.4	43.2	53.6	55.3	55.2	**59.5**
IPC_6	38.2	49.2	58.0	60.5	58.8	**61.6**
IPC_7	41.2	51.6	61.0	**63.3**	61.5	62.7
IPC_8	45.4	56.3	63.6	**65.0**	63.4	64.7
IPC_9	49.2	58.5	65.7	**67.4**	65.4	65.7
IPC_{10}	52.1	62.8	67.5	**69.4**	66.7	66.7
Avg.	35.4	43.5	53.2	55.5	58.4	**60.6**
Diff.	-25.2	-17.1	-7.4	-5.1	-2.2	-

B. Main results

Comparison with State-of-the-art Methods. In Table I, we compare our approach with state-of-the-art (SOTA) condensation techniques, including DSA [10], MTT [15], IDC [18], Dream [17], and MDC [19]. The results show that our method achieves superior average performance and is particularly effective in low-IPC regimes. Our method obtains an average absolute improvement of 2.2 percentage points over the previous best (MDC), demonstrating a clear overall uplift. We attribute this gain to the combined effect of Hessian regularization and the dynamic hybrid loss. Hessian regularization steers synthesis toward flatter minima, suppressing destructive high-frequency updates and providing convergence stability and preservation of initialization priors. The dynamic hybrid loss, by contrast, encourages aggressive exploration during the early stages of synthesis by exposing the model to varied subset compositions, which promotes the fusion of low- and high-order information.

The gains are particularly notable in the IPC_3–IPC_6 range: compared with MDC we improve by $+5.3$, $+5.8$, $+4.2$, and $+2.8$ percentage points at IPC_3, IPC_4, IPC_5, and IPC_6, respectively. At higher IPC budgets, the advantage becomes less pronounced. This pattern reflects how, within low IPC regime, the richer fusion of low- and high-order information encouraged by the dynamic hybrid loss benefits most from the stabilizing effect of Hessian regularization, yielding higher-quality condensed sets and improved downstream performance.

Our method demonstrates a pronounced advantage on the SVHN dataset, particularly under small-IPC settings, achieving an improvement of 2.2%. On the more challenging CIFAR-100 dataset, SMC consistently outperforms MDC across almost all IPC budgets, achieving an average accuracy of 38.5% (+1.0% over MDC). This result suggests that our approach scales more gracefully to complex class distributions: even under such challenging conditions, the geometric regularization ensures that the condensed data retains robust semantic information.

TABLE II
RESULTS OF SVHN AND CIFAR-100 TARGETING IPC_{10}

Dataset	Method	IPC_1	IPC_2	IPC_3	IPC_4	IPC_5	IPC_6	IPC_7	IPC_8	IPC_9	IPC_{10}	Avg.	Diff.
SVHN	IDC	35.5	51.6	60.4	68.0	74.4	77.7	81.7	83.9	86.0	**87.5**	70.7	-8.3
	MDC	63.3	67.9	72.2	74.1	77.5	78.2	80.9	82.8	84.3	86.4	76.8	-2.2
	Ours	**65.4**	**71.7**	**74.2**	**75.2**	**78.9**	**80.6**	**84.0**	**85.5**	**86.7**	**87.5**	**79.0**	-
CIFAR-100	IDC	14.4	21.8	28.0	32.2	35.3	39.1	40.9	42.7	44.3	**45.4**	34.4	-4.1
	MDC	27.6	31.8	33.6	35.4	36.9	39.0	40.7	42.1	43.9	44.3	37.5	-1.0
	Ours	**28.9**	**31.9**	**35.6**	**37.6**	**38.3**	**40.0**	**42.0**	**42.5**	**43.9**	44.6	**38.5**	-

TABLE III
ABLATION RESULTS

	IPC_1	IPC_2	IPC_3	IPC_4	IPC_5	IPC_6	IPC_7	IPC_8	IPC_9	IPC_{10}	Avg.	Diff.
SMC (Full)	**50.9**	**54.3**	**59.2**	**60.4**	**59.5**	**61.6**	**62.7**	**64.7**	**65.7**	**66.7**	**60.6**	-
w/o Hessian Reg.	48.6	53.3	53.5	54.3	54.9	59.7	61.8	63.5	65.3	66.5	58.2	**-2.4**
w/o $\mathcal{L}_{\mathrm{D}}$ Loss	48.6	52.8	54.1	55.3	56.0	59.8	61.2	63.0	64.6	65.8	58.1	-2.5
w/o $\mathcal{L}_{\mathrm{P}}$ Loss	27.4	34.6	40.1	45.6	48.0	51.4	53.1	55.0	56.3	57.7	46.9	-13.7
Fixed $\gamma(0.2)$	49.9	51.6	52.6	54.8	55.6	59.2	61.7	63.4	65.6	66.3	58.1	-2.5

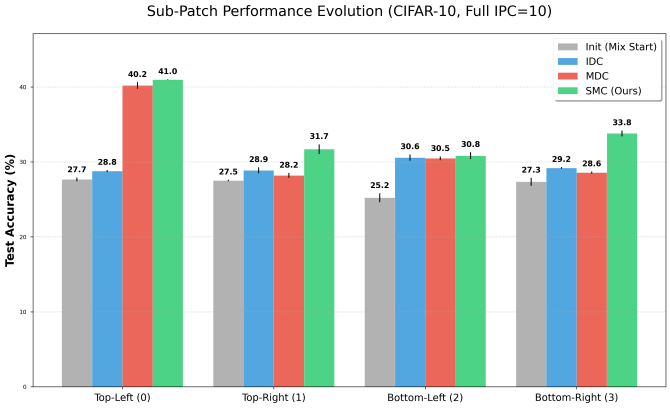

Fig. 5. **Sub-Patch Performance Disparity.** Comparing test accuracy across spatially decomposed sub-patches. While the baseline MDC (red) exhibits a sharp performance decay in secondary patches (e.g., bottom-right drops to near-initialization levels), our SMC (green) effectively rectifies this imbalance, boosting the neglected regions by **+5.2%** without compromising the primary patch.

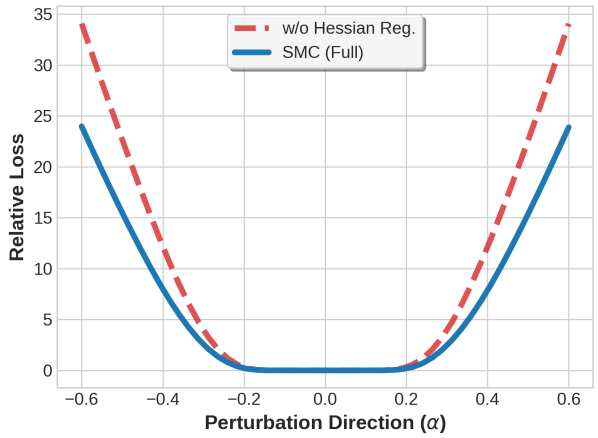

Fig. 6. Ablation on Optimization Landscape.

C. Analysis of Sub-Patch Performance Dynamics.

We analyze the performance of spatially decomposed sub-patches in Fig. 5. The baseline MDC suffers from a sloped decay: while optimizing the primary top-left patch (40.2%), it degrades the secondary bottom-right patch (28.6%) to near-initialization levels. Conversely, SMC rectifies this imbalance, retaining peak primary performance (41.0%) while boosting the bottom-right patch to **33.8%** (+5.2%). This confirms that Hessian regularization acts as a protective barrier against destructive high-frequency updates, successfully reconciling optimization gains with the preservation of initialization priors.

D. Ablation Study

Effect of Hessian-Aware Regularization. The row labeled w/o Hessian Reg. (formerly Global Stabilization) in Table III demonstrates the critical role of geometric smoothing. Removing this term leads to a 2.4% drop in average accuracy. To visualize this geometric impact, we plot the loss landscapes of SMC and the ablation variant in Fig. 6. While the variant without Hessian regularization (red dashed line) exhibits a sharp V-shaped valley indicating sensitivity to perturbations, the full SMC (blue solid line) maintains a wide, flat basin. Notably, the performance gap is widest at low IPCs, confirming that flat minima are essential for preserving representational power when the budget is tight.

Impact of Loss Function Components. The ablation results in Table III show that $\mathcal{L}_{\mathrm{P}}$ and $\mathcal{L}_{\mathrm{D}}$ play complementary roles. $\mathcal{L}_{\mathrm{P}}$ primarily preserves coarse, low-order information: removing the $\mathcal{L}_{\mathrm{P}}$ term causes a large accuracy degradation (e.g., IPC=1 drops from 50.9% to 27.4%). By contrast, removing $\mathcal{L}_{\mathrm{D}}$ produces only a small average decline, but per-IPC gains from adding $\mathcal{L}_{\mathrm{D}}$ are visible across budgets. These patterns indicate

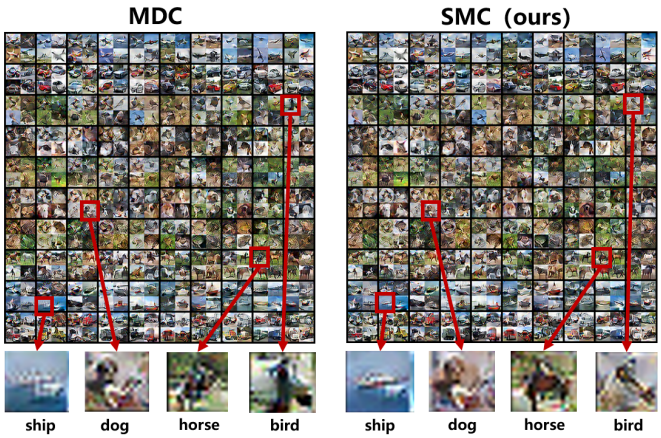

Fig. 7. Visualization of synthetic data generated by SMC (left) and MDC (right) with IPC=10 on CIFAR-10. Each 10×10 grid displays synthetic images for the 10 classes. SMC preserves clear global structures and object contours (e.g., the recognizable silhouette of ships and limbs in animals).

that $\mathcal{L}_{\mathrm{P}}$ maintains the basic, global signal while $\mathcal{L}_{\mathrm{D}}$ supplies complementary, higher-order detail that improves intermediate and larger IPC performance. With γ fixed at 0.2, despite its advantage at low IPC, the capacity of $\mathcal{L}_{\mathrm{D}}$ to provide complementary information is constrained. Consequently, accuracy exhibits only a modest improvement as IPC increases.

E. Visual Quality Analysis

Figure 7 visualizes the synthetic images generated by SMC at IPC=10. Unlike methods that yield blurred boundaries due to texture overfitting, SMC preserves clear global structures and object contours (e.g., recognizable silhouette of ships and limbs in animals). This visual coherence suggests that our Hessian-aware regularization effectively steers the optimization away from sharp, noisy minima, resulting in synthetic data that is not only discriminative but also human-interpretable.

V. CONCLUSION

In this work, we introduced **S**tabilized **M**ulti-dimensional **C**ondensation (SMC), a unified framework that resolves the optimization instability and patch-wise update imbalance in one-shot dataset distillation. Crucially, we identified that the generalization capability of synthetic data is intrinsically linked to the geometry of its optimization landscape. By integrating a Hessian-aware curvature regularization with a dynamic hybrid loss, SMC effectively steers the synthesis towards flat minima, ensuring robustness against stochastic subsampling and preserving structural fidelity across all scales. Extensive experiments demonstrate that SMC not only achieves state-of-the-art accuracy (e.g., **+2.2%** on CIFAR-10) but also yields synthetic images with superior visual coherence. Future work will focus on exploring the theoretical bounds of synthetic data curvature and developing unbiased compression algorithms to further decouple the distillation process from specific real-dataset biases.

REFERENCES

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

[2] R. G. et al., "Imagenet-trained cnns are biased towards texture," in *ICLR*, 2019.

[3] M. Tran, T. Le, X. Le, T. Do, and D. Q. Phung, "Enhancing dataset distillation via non-critical region refinement," in *CVPR*, 2025, pp. 10 015–10 024.

[4] Z. Zhou, W. Feng, Q. Zhang, S. Lyu, Q. Zhao, and G. Cheng, "ROME is forged in adversity: Robust distilled datasets via information bottleneck," in *ICML*, 2025.

[5] M. F. Hutchinson, "A stochastic estimator of the trace," *Communications in Statistics*, vol. 18, no. 3, pp. 1059–1076, 1989.

[6] B. Zhao and H. Bilen, "Dataset condensation with distribution matching," in *WACV*, 2023, pp. 6503–6512.

[7] S. Wang, Y. Yang, Z. Liu, C. Sun, X. Hu, C. He, and L. Zhang, "Dataset distillation with neural characteristic function," in *CVPR*, 2025, pp. 25 570–25 580.

[8] K. Wang, B. Zhao, X. Peng, Z. Zhu, S. Yang, S. Wang, G. Huang, H. Bilen, X. Wang, and Y. You, "CAFE: learning to condense dataset by aligning features," in *CVPR*, 2022, pp. 12 186–12 195.

[9] B. Zhao, K. R. Mopuri, and H. Bilen, "Dataset condensation with gradient matching," in *International Conference on Learning Representations (ICLR)*, 2021.

[10] B. Zhao and H. Bilen, "Dataset condensation with differentiable siamese augmentation," in *International Conference on Machine Learning (ICML)*, 2021, pp. 12 674–12 685.

[11] G. Zhao, G. Li, Y. Qin, and Y. Yu, "Improved distribution matching for dataset condensation," in *CVPR*, 2023, pp. 7856–7865.

[12] J. Du, Y. Jiang, V. Y. F. Tan, J. T. Zhou, and H. Li, "Minimizing the accumulated trajectory error to improve dataset distillation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3749–3758.

[13] W. L. et al., "Robust dataset distillation by matching adversarial trajectories," *arXiv*, 2025.

[14] S. Joshi, J. Ni, and B. Mirzasoleiman, "Dataset distillation via knowledge distillation," in *ICLR*, 2025.

[15] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J. Zhu, "Dataset distillation by matching training trajectories," in *CVPR Workshops*, 2022, pp. 4749–4758.

[16] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," in *ICLR*, 2020.

[17] Y. Liu, J. Gu, K. Wang, Z. Zhu, W. Jiang, and Y. You, "DREAM: efficient dataset distillation by representative matching," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 17 268–17 278.

[18] J. Kim, J. Kim, S. J. Oh, S. Yun, H. Song, J. Jeong, J. Ha, and H. O. Song, "Dataset condensation via efficient synthetic-data parameterization," in *International Conference on Machine Learning (ICML)*, 2022, pp. 11 102–11 118.

[19] Y. He, L. Xiao, J. T. Zhou, and I. W. Tsang, "Multisize dataset condensation," in *International Conference on Learning Representations (ICLR)*, 2024.

[20] S. Hochreiter and J. Schmidhuber, "Flat minima," *Neural Computation*, vol. 9, no. 1, pp. 1–42, 1997.

[21] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *ICLR*, 2021.

[22] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[23] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop*, 2011.

[24] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *CVPR*, 2018, pp. 4367–4375.