SPECTRAL GRADIENT SURGERY FOR DOMAIN-GENERALIZABLE DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset Distillation (DD) synthesizes a compact synthetic dataset that preserves the training utility of a full dataset. However, its standard formulation assumes that test data follow the same distribution as training data, which rarely holds in practice. A straightforward extension—applying post-hoc Domain Generalization (DG) techniques to distilled data—is ill-suited because existing DG methods rely on the natural diversity of real datasets, which compact synthetic sets inherently lack, and their augmentation overhead conflicts with the efficiency goal of DD. We therefore study Domain Generalizable Dataset Distillation (DGDD), a setting in which distilled datasets are evaluated on target domains unseen during distillation, through the lens of Distribution Matching (DM). We hypothesize that the OOD vulnerability of DM stems from compressing domain-consistent and domain-specific information indiscriminately, and propose Spectral Gradient Surgery (SGS). SGS computes domain-wise DM gradients with respect to the synthetic images, transforms them into the Fourier domain, and measures per-frequency phase agreement across source domains. It then augments the standard DM update with two terms: a class-relevant component that emphasizes frequency components consistent across source domains, and a domain-relevant component that preserves source-domain variation to diversify the synthetic set. Across three benchmarks of increasing resolution, SGS consistently improves the average OOD accuracy of two DM methods without modifying their objectives or downstream training cost, at the price of additional offline distillation cost.

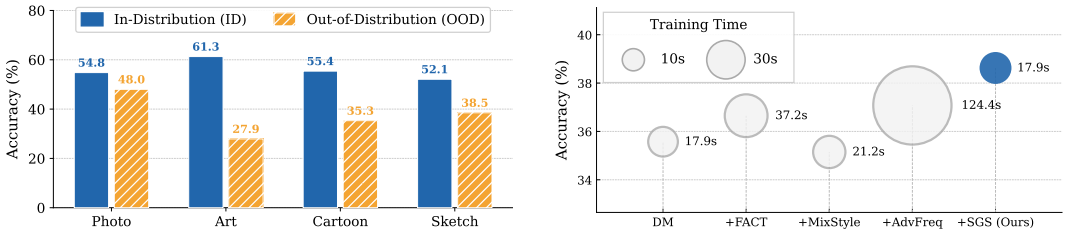
1 INTRODUCTION

In recent years, deep neural networks have achieved remarkable progress across diverse fields, driven by the availability of large-scale training datasets. However, the increasing scale of training data introduces substantial storage and computational costs. To address this, Dataset Distillation (DD) (Wang et al., 2018) synthesizes a compact dataset such that a model trained on it achieves performance comparable to one trained on the full dataset, enabling data storage reduction and efficient training under resource constraints.

DD methods typically update the synthetic dataset by matching the training dynamics between the synthetic and real datasets. Depending on what is matched, DD methods are categorized into (1) Distribution Matching (DM) (Zhao & Bilen, 2023), which aligns feature distributions; (2) Gradient Matching (GM) (Zhao & Bilen, 2021b), which aligns per-step gradients; and (3) Trajectory Matching (TM) (Cazenavette et al., 2022), which aligns parameter trajectories over training. However, existing DD research predominantly assumes that the training and test data are independent and identically distributed (i.i.d.), which rarely holds in practice. Under domain shift, models trained on distilled data often exhibit significant performance degradation, limiting the practical applicability of DD. We investigate this problem through the lens of DM, which has been widely adopted as a baseline due to its strong performance and computational efficiency.

We assess the robustness of a DM-distilled dataset with 10 images per class on PACS (Li et al., 2017) under two evaluation settings: **in-distribution (ID)**, evaluated on the source domains, and **out-of-distribution (OOD)**, evaluated on an unseen target domain via leave-one-domain-out. As shown in Figure 1a, DM suffers from severe OOD performance degradation. Applying Domain Generalization (DG) methods (Xu et al., 2021; Zhou et al., 2021) during downstream training is a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) ID vs. OOD accuracy of DM

(b) Post-hoc DG on DM-distilled data

Figure 1: Motivation on PACS (IPC=10). (a) In-distribution (ID) and out-of-distribution (OOD) accuracy of DM for each target domain. (b) Average OOD accuracy and downstream training time when DG methods are applied post-hoc to DM-distilled datasets.

straightforward remedy, but Figure 1b shows that such post-hoc approaches provide only marginal gains and can even harm performance. We attribute this to distilled datasets differing visually from real images, which violates the assumptions underlying existing DG methods; moreover, their heavy augmentation substantially increases training time, contradicting DD's objective of rapid training. Motivated by this, we study **Domain Generalizable Dataset Distillation (DGDD)**, in which the distilled dataset itself is optimized for generalization to unseen target domains.

We hypothesize that the OOD degradation of DM arises because it compresses the entire multi-domain source into a compact set without distinguishing information that is consistent across source domains from information specific to individual domains. To mitigate this, we propose **Spectral Gradient Surgery (SGS)**, which augments the standard DM update with two gradient terms derived from domain-wise gradients in the spectral domain. SGS computes domain-wise DM gradients with respect to the synthetic images, transforms them with a 2D Fourier transform, and measures cross-domain phase agreement at each frequency via the resultant length. Frequency components with high agreement are aggregated into a *class-relevant* component, while per-domain deviations from the consensus form a *domain-relevant* component that preserves source-domain variation and diversifies the synthetic set. SGS leaves the DM objective unchanged and can be attached to existing DM methods; its additional cost is incurred only during offline distillation. In effect, SGS steers the synthetic images toward patterns that are shared across the observed source domains, and are thus more likely to persist in unseen ones, while retaining the variation of individual domains to keep the compact set diverse.

- We study DGDD, where distilled datasets are evaluated on target domains excluded from distillation, and show that post-hoc DG methods yield limited gains at substantial downstream training cost.
- We propose SGS, which decomposes image-space DM gradients into class- and domain-relevant components by using per-frequency cross-domain phase agreement, while leaving the DM objective unchanged.
- On Digits-DG, PACS, and CORe50-Hard, SGS consistently improves the average OOD accuracy of DM and HDD-DM without downstream training overhead.

2 Related Work

2.1 Dataset Distillation

Existing dataset distillation methods are typically categorized into Gradient Matching (GM) (Zhao & Bilen, 2021b;a), which aligns per-step gradients between real and synthetic data; Trajectory Matching (TM) (Cazenavette et al., 2022; Cui et al., 2023), which aligns parameter trajectories over training; and Distribution Matching (DM) (Zhao & Bilen, 2023), which is the focus of this paper. While GM and TM achieve strong performance, their reliance on costly bi-level optimization limits scalability. DM instead directly aligns feature distributions between real and synthetic data using randomly initialized encoders, offering a more efficient alternative. Sajedi et al. (2023) proposes spatial attention matching across multiple network layers. Zhang et al. (2024) minimizes

Maximum Mean Discrepancy in a reproducing kernel Hilbert space to align higher-order moments. Wang et al. (2025) reformulates distribution matching as a minmax optimization over neural characteristic functions. Li et al. (2025) embeds feature distributions into hyperbolic space to capture hierarchical sample structure. These methods match the feature distribution of the source dataset and are evaluated in-distribution. Oh & Sim (2026) distills from streaming multi-domain datasets, but restricts its scope to seen domains. Most closely related to our setting, Choi et al. (2025) proposes a domain-aware module for multi-domain dataset condensation that preserves heterogeneous domain information within the synthetic set. Its main evaluation, however, is conducted on the same domains used for distillation. In contrast, we treat generalization to domains excluded from distillation as the primary objective and evaluate it across multiple IPC settings and benchmarks.

2.2 Domain Generalization

Domain generalization (DG) aims to mitigate performance degradation under distribution shifts. One line of work enhances training diversity through data augmentation: Xu et al. (2021) mixes the amplitude spectrum of source images while preserving the phase, Zhou et al. (2021) mixes channel-wise feature statistics across instances, Zheng et al. (2024) adversarially optimizes parameterized augmentations, and He et al. (2024) perturbs the amplitude spectrum to reduce reliance on shortcut frequency bands. Another line enforces invariance through cross-domain gradient or predictor consistency: Arjovsky et al. (2019) enforces a shared optimal classifier across domains, Shi et al. (2022) aligns per-domain gradient means, and Rame et al. (2022) matches their variances. Invariance and gradient agreement are thus established ideas in DG. These methods, however, regularize model parameters or representations through parameter-space gradients, whereas SGS applies cross-domain agreement to the image-space gradients of synthetic data, which directly determine the pixel updates during distillation. This enables a spectral decomposition on the 2D Fourier basis, which has no natural counterpart in parameter space. Applying existing DG methods post-hoc to distilled datasets is also insufficient in our setting: they assume diverse data drawn from a natural image distribution, which compact synthetic datasets violate, and their augmentation pipelines add downstream training overhead. This motivates integrating generalization into the distillation process itself.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \bigcup_{s=1}^{S} \mathcal{D}^s$ be the union of S source datasets representing distinct domains over a shared class label space $\mathcal{Y}$, with $\mathbb{P}^{\mathcal{D}}$ denoting its mixture distribution. We define $\mathcal{T}$ as the test dataset drawn from a target distribution $\mathbb{P}^{\mathcal{T}}$. Given $\mathcal{D}$, dataset distillation synthesizes a compact synthetic dataset $\hat{\mathcal{D}} = \{(\hat{\boldsymbol{x}}_j, \hat{y}_j)\}$ that serves as a faithful proxy for $\mathcal{D}$. The synthetic dataset $\hat{\mathcal{D}}$ is typically much smaller than the full dataset $\mathcal{D}$ ($|\hat{\mathcal{D}}| \ll |\mathcal{D}|$), where $|\hat{\mathcal{D}}| = \text{IPC} \times C$, with IPC denoting the number of images per class and C the number of classes in $\mathcal{Y}$. The goal of conventional DD is to find $\hat{\mathcal{D}}$ that minimizes the performance gap between networks trained on $\hat{\mathcal{D}}$ and on $\mathcal{D}$, measured by a loss function $\mathcal{L}$ on $\mathcal{T}$ based on the i.i.d. assumption:

$$\hat{\mathcal{D}}^* = \arg\min_{\hat{\mathcal{D}}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{T}} \left[\mathcal{L}(\phi_{\hat{\mathcal{D}}}(\boldsymbol{x}), y) - \mathcal{L}(\phi_{\mathcal{D}}(\boldsymbol{x}), y)\right], \quad \text{s.t.} \quad \mathbb{P}^{\mathcal{T}} = \mathbb{P}^{\mathcal{D}}, \tag{1}$$

where $\phi_{\hat{\mathcal{D}}}$ and $\phi_{\mathcal{D}}$ denote networks trained on $\hat{\mathcal{D}}$ and $\mathcal{D}$, respectively.

However, the strict i.i.d. assumption rarely holds in the presence of domain shift. To address this limitation, we introduce a more general problem setting of **Domain Generalizable Dataset Distillation (DGDD)**:

$$\hat{\mathcal{D}}^* = \arg\min_{\hat{\mathcal{D}}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{T}} \left[\mathcal{L}(\phi_{\hat{\mathcal{D}}}(\boldsymbol{x}), y) - \mathcal{L}(\phi_{\mathcal{D}}(\boldsymbol{x}), y)\right], \quad \text{s.t.} \quad \text{supp}(\mathbb{P}^{\mathcal{T}}) \setminus \text{supp}(\mathbb{P}^{\mathcal{D}}) \neq \varnothing. \tag{2}$$

Unlike conventional DD settings that rely on the strict i.i.d. assumption, DGDD explicitly targets robustness to domain shift by evaluating on a target distribution $\mathbb{P}^{\mathcal{T}}$ that is not fully covered by the source distribution $\mathbb{P}^{\mathcal{D}}$. A model trained on $\hat{\mathcal{D}}$ should therefore generalize to target domains that are not represented in $\mathcal{D}$, while remaining accurate on the source domains.

3.2 Dissecting Distribution Matching

Distribution Matching (DM) (Zhao & Bilen, 2023) updates a synthetic dataset $\hat{\mathcal{D}}$ by minimizing the discrepancy between the feature distributions of synthetic and real data, computed per class:

$$\mathcal{L}_{\mathrm{DM}} = \sum_{c=1}^{C} \left\| \mathbb{E}_{\hat{\boldsymbol{x}} \in \hat{\mathcal{D}}_c}[\psi(\hat{\boldsymbol{x}})] - \mathbb{E}_{\boldsymbol{x} \in \mathcal{D}_c}[\psi(\boldsymbol{x})] \right\|^2, \tag{3}$$

where ψ is a feature extractor, and $\hat{\mathcal{D}}_c$ and $\mathcal{D}_c$ are the subsets of $\hat{\mathcal{D}}$ and $\mathcal{D}$, respectively, belonging to the c-th class. The i-th synthetic sample $\hat{\boldsymbol{x}}_i$ is updated via the gradient of $\mathcal{L}_{\mathrm{DM}}$ with a learning rate η:

$$\hat{\boldsymbol{x}}_i \leftarrow \hat{\boldsymbol{x}}_i - \eta \cdot \boldsymbol{g}_i, \quad \text{where} \quad \boldsymbol{g}_i = \frac{\partial \mathcal{L}_{\mathrm{DM}}}{\partial \hat{\boldsymbol{x}}_i}. \tag{4}$$

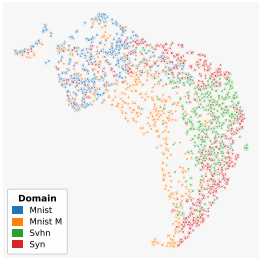


Figure 2: UMAP visualization of channel-wise feature statistics on Digits-DG. Colors indicate a distinct domain.

Because $\boldsymbol{g}_i$ is computed against the union of all source domains, it mixes signals originating from every source domain. The features of source data carry substantial domain information: channel-wise feature statistics of Digits-DG samples form clear domain-wise clusters (Figure 2). This observation motivates the hypothesis that $\boldsymbol{g}_i$ contains, conceptually, two kinds of components:

$$\boldsymbol{g}_i \approx \boldsymbol{g}_i^{\text{class}} + \boldsymbol{g}_i^{\text{domain}}, \tag{5}$$

where $\boldsymbol{g}_i^{\text{class}}$ collects directions that are consistent across source domains, and $\boldsymbol{g}_i^{\text{domain}}$ collects deviations particular to individual source domains. We use cross-domain consistency as a proxy for transferability to unseen domains.

3.3 Spectral Gradient Surgery

SGS operationalizes Eq. 5 by measuring the cross-domain agreement of domain-wise gradients in the spectral domain. It augments the standard DM update with two terms: a *class-relevant component*, which emphasizes spectral components with strong cross-domain consensus, and a *domain-relevant component*, which captures the deviation of each source domain from this consensus and thereby promotes diversity within the synthetic dataset.

Domain-wise Gradient Computation. Given source datasets $\{\mathcal{D}^s\}_{s=1}^{S}$ from S domains, we compute a domain-wise gradient for each source domain $\mathcal{D}^s$:

$$\boldsymbol{g}_i^s = \frac{\partial \mathcal{L}_{\mathrm{DM}}^s}{\partial \hat{\boldsymbol{x}}_i}, \quad \mathcal{L}_{\mathrm{DM}}^s = \sum_{c=1}^{C} \left\| \mathbb{E}_{\hat{\boldsymbol{x}} \in \hat{\mathcal{D}}_c}[\psi(\hat{\boldsymbol{x}})] - \mathbb{E}_{\boldsymbol{x} \in \mathcal{D}_c^s}[\psi(\boldsymbol{x})] \right\|^2, \tag{6}$$

where $\mathcal{D}_c^s$ denotes the subset of $\mathcal{D}^s$ belonging to the c-th class. Intuitively, $\boldsymbol{g}_i^s$ indicates how each pixel of $\hat{\boldsymbol{x}}_i$ should change to better represent domain $\mathcal{D}^s$.

Spectral Gradient Decomposition. Since each $\boldsymbol{g}_i^s$ is an image-space offset for $\hat{\boldsymbol{x}}_i$, one could measure agreement directly in pixel space; we instead measure it in the spectral domain for three reasons. First, pixel-wise agreement compares values at individual spatial coordinates and is therefore sensitive to spatial misalignment of patterns across domains. Second, a single frequency component represents a global spatial pattern over the entire image, so spectral agreement reflects whether such a pattern is shared across domains. Third, complex Fourier coefficients admit a continuous, phase-based measure of agreement, whereas real-valued pixel gradients only allow a binary comparison of their signs. We transform each domain-wise gradient via the 2D Fast Fourier Transform (FFT) $\mathcal{F}$ and compute the complex mean $\bar{G}_i$ across source domains:

$$G_i^s = \mathcal{F}(\boldsymbol{g}_i^s), \qquad \bar{G}_i = \frac{1}{S} \sum_{s=1}^{S} G_i^s. \tag{7}$$

Since not all frequency components agree equally across domains, we quantify the agreement at each frequency by the element-wise resultant length from circular statistics:

$$r_i = \frac{\left|\sum_{s=1}^{S} G_i^s\right|}{\sum_{s=1}^{S} |G_i^s| + \epsilon} \in [0, 1), \tag{8}$$

where $|\cdot|$ denotes the element-wise complex magnitude and ϵ is a small constant for numerical stability. Ignoring ϵ, $r_i = 1$ at a frequency if and only if all non-zero domain-wise coefficients share the same phase, and r_i approaches 0 as they cancel each other out. We then define

$$\boldsymbol{g}_i^{\text{class}} = \mathcal{F}^{-1}(r_i \cdot \bar{G}_i), \qquad \boldsymbol{g}_i^{\text{domain},s} = \mathcal{F}^{-1}(G_i^s - \bar{G}_i), \tag{9}$$

where $\boldsymbol{g}_i^{\text{class}}$ retains phase-consistent components and attenuates phase-inconsistent ones, and $\boldsymbol{g}_i^{\text{domain},s}$ is the deviation of domain s from the consensus. In Appendix C, we show that weighting by r_i attenuates a frequency component with random phases across domains at rate $\mathcal{O}(1/S)$.

Synthetic Data Update. The standard DM gradient $\boldsymbol{g}_i$, computed against the union of all source domains, remains the primary update direction. The class-relevant component $\boldsymbol{g}_i^{\text{class}}$ reinforces patterns consistent across source domains, and the domain-relevant component $\boldsymbol{g}_i^{\text{domain},s(i)}$ preserves the variation of one source domain, where $s(i) \in \{1, \ldots, S\}$ is a pre-assigned source-domain index for $\hat{\boldsymbol{x}}_i$. The assignment is balanced so that every source domain is allocated an equal number of synthetic samples. The synthetic sample is updated as

$$\hat{\boldsymbol{x}}_i \leftarrow \hat{\boldsymbol{x}}_i - \eta \cdot \left(\boldsymbol{g}_i + \lambda_c \cdot \boldsymbol{g}_i^{\text{class}} + \lambda_d \cdot \boldsymbol{g}_i^{\text{domain},s(i)}\right). \tag{10}$$

λ_c and λ_d control the strength of each term; setting $\lambda_c = \lambda_d = 0$ recovers the standard DM update.

4 Experiments

4.1 Experimental Setup

Datasets We employ datasets of increasing scale: Digits-DG (Zhou et al., 2020), PACS (Li et al., 2017), and CORe50-Hard (Lomonaco & Maltoni, 2017). Digits-DG contains 10 classes at 32×32 resolution in four domains: MNIST (M) (LeCun et al., 2002), MNIST-M (N) (Ganin & Lempitsky, 2015), SVHN (V) (Netzer et al., 2011), and SYN (Y) (Ganin & Lempitsky, 2015). PACS contains 7 classes at 64×64 resolution in four domains: Photo (P), Art painting (A), Cartoon (C), and Sketch (S). We construct CORe50-Hard from CORe50 at 128×128 resolution by selecting ten object categories and the four sessions (s5, s7, s9, s10) farthest from the overall mean in amplitude-spectrum space, which differ in indoor/outdoor conditions, backgrounds, and illumination. More details are given in Appendix B.

Implementation Details We implement all experiments in PyTorch. For the feature extractor ψ, we employ ConvNets with 3, 4, and 5 layers for Digits-DG, PACS, and CORe50-Hard, respectively, following prior works (Li et al., 2025; Zhang et al., 2024). Synthetic samples are initialized with real images sampled equally from each source domain and optimized with SGD for 20,000 iterations, using differentiable augmentation (DSA) (Zhao & Bilen, 2021a) as in standard DM practice. To evaluate the distilled datasets, we train ConvNet classifiers from scratch for 1,000 epochs using SGD. All results are the mean and standard deviation over five independent runs. All hyperparameters (λ_c, λ_d, and K) are selected using source-domain gradient statistics only, without access to target-domain images, labels, or accuracy; the selection rule and the values for every split are given in Appendix E.2. Experiments on CORe50-Hard use a single A100 GPU, and all others use a single RTX 3090 GPU.

Evaluation Protocols We evaluate the distilled datasets under two distinct domain-shift settings: multi-domain generalization (MDG) and single-domain generalization (SDG, Appendix F.1). For MDG, we adopt a leave-one-domain-out protocol: we select one domain as the strictly unseen target for evaluation and use all remaining domains as sources for distillation, reporting the top-1 classification accuracy (%) on the target domain. For SDG, we distill the dataset from a single source domain and evaluate the trained model on each of the remaining domains as unseen targets. For every source choice, we report the average top-1 accuracy across all target domains.

Table 1: Multi-domain generalization performance on the Digits-DG, PACS, and CORe50-Hard datasets with IPC=10 and IPC=20.

Dataset	Digits-DG									
IPC	10					20				
Target	M	N	V	Y	Avg.	M	N	V	Y	Avg.
Whole	97.8±0.1	66.2±0.7	75.5±0.2	92.7±0.2	83.1±0.3	97.8±0.1	66.2±0.7	75.5±0.2	92.7±0.2	83.1±0.3
Random	42.6±2.2	34.2±1.6	29.0±0.9	39.2±0.4	36.3±1.3	65.9±0.7	38.5±1.1	38.5±0.3	52.6±0.5	48.9±0.7
Herding	69.9±1.5	38.5±0.9	39.2±0.6	51.5±0.7	49.8±0.9	78.3±2.0	42.4±0.7	44.3±0.9	62.0±0.7	56.8±1.1
DataDAM	80.0±1.6	47.4±1.1	33.5±0.7	46.7±1.8	51.9±1.3	88.1±0.7	52.0±1.2	40.6±1.6	62.0±0.6	60.7±1.0
M3D	88.1±0.5	41.5±0.1	46.8±0.2	62.3±1.0	59.7±0.5	90.1±1.0	46.6±0.9	56.4±0.3	71.2±1.0	66.1±0.8
DM	85.3±0.7	51.4±1.5	53.6±0.9	69.9±0.9	65.1±1.0	88.4±0.8	53.1±1.0	56.4±1.1	73.7±0.2	67.9±0.8
+ SGS	**90.3**±0.4	**54.1**±0.9	**55.3**±0.4	**71.0**±0.8	**67.7**±0.6	**90.7**±0.5	54.3±1.2	**58.0**±0.5	**74.3**±0.8	**69.3**±0.8
HDD-DM	86.3±0.6	52.6±1.1	53.5±0.8	68.4±1.0	65.2±0.9	89.5±0.4	54.4±1.7	56.7±1.1	73.2±0.3	68.5±0.9
+ SGS	87.9±1.2	53.7±2.2	54.1±0.9	70.0±0.4	66.4±1.2	**90.7**±0.3	**54.9**±1.3	57.5±1.0	73.4±0.7	69.1±0.8

Dataset	PACS									
IPC	10					20				
Target	P	A	C	S	Avg.	P	A	C	S	Avg.
Whole	59.1±1.4	36.7±0.4	52.9±1.2	50.2±1.0	49.7±1.0	59.1±1.4	36.7±0.4	52.9±1.2	50.2±1.0	49.7±1.0
Random	38.3±1.0	22.5±0.1	25.4±1.1	24.9±2.3	27.8±1.1	45.6±1.8	23.3±0.8	27.7±0.7	28.0±0.7	31.2±1.0
Herding	37.9±3.3	25.0±0.9	30.2±2.1	26.4±4.0	29.9±2.6	43.7±1.8	26.5±0.7	32.2±1.6	24.8±2.6	31.8±1.7
DataDAM	44.3±1.4	27.6±0.9	32.5±0.9	27.2±4.8	32.9±2.0	49.6±1.3	29.7±0.4	38.3±0.5	36.3±2.5	38.5±1.2
M3D	49.0±0.9	27.1±0.6	**35.9**±1.4	25.5±2.7	34.4±1.4	48.6±0.5	**30.3**±1.0	38.6±0.7	39.3±2.7	39.2±1.2
DM	48.0±0.7	27.9±0.5	33.8±1.5	32.6±2.2	35.6±1.2	49.2±0.4	28.8±1.0	35.4±2.2	38.1±2.7	37.9±1.6
+ SGS	**53.8**±0.4	28.4±1.1	35.1±1.5	**37.2**±1.8	**38.6**±1.2	**53.2**±1.0	29.1±0.6	37.7±2.3	**39.8**±2.5	**40.0**±1.6
HDD-DM	47.8±0.4	28.4±0.7	34.1±1.8	29.2±3.0	34.9±1.5	48.9±1.2	29.5±1.2	37.1±1.7	32.5±1.7	37.0±1.5
+ SGS	50.1±2.2	**29.1**±0.9	**35.9**±1.1	33.4±1.9	37.1±1.5	51.8±0.7	29.9±0.8	**39.3**±1.2	37.5±2.5	39.6±1.3

Dataset	CORe50-Hard									
IPC	10					20				
Target	s5	s7	s9	s10	Avg.	s5	s7	s9	s10	Avg.
Whole	54.4±2.0	46.5±1.7	50.6±1.5	26.7±1.2	44.6±1.6	54.4±2.0	46.5±1.7	50.6±1.5	26.7±1.2	44.6±1.6
Random	23.0±0.4	19.8±0.3	20.7±0.9	16.8±0.4	20.1±0.5	24.2±0.6	26.0±1.1	24.4±1.0	20.0±0.9	23.7±0.9
Herding	24.7±1.4	20.9±0.8	24.6±1.9	19.9±1.1	22.5±1.3	28.1±1.6	27.1±0.6	28.9±0.5	19.9±0.7	26.0±0.9
DataDAM	29.3±2.5	31.7±0.7	29.1±1.0	20.0±1.0	27.5±1.3	31.1±1.0	28.9±1.0	28.0±1.0	20.2±1.0	27.1±1.0
M3D	20.4±0.8	30.0±1.4	26.0±0.7	18.3±0.6	23.7±0.9	22.0±1.2	29.1±2.0	28.5±1.4	**21.6**±0.8	25.3±1.4
DM	28.4±1.7	31.4±2.0	30.3±2.2	19.3±1.4	27.4±1.8	30.5±2.2	32.5±2.1	31.9±0.9	20.5±1.1	28.9±1.6
+ SGS	**29.8**±1.9	**32.6**±2.1	**32.3**±1.7	21.0±1.1	**28.9**±1.7	**32.0**±2.0	**34.7**±0.6	**34.0**±1.1	20.9±1.0	**30.4**±1.2
HDD-DM	25.6±1.4	25.4±2.1	24.6±0.9	18.7±1.3	23.6±1.4	27.7±2.1	30.9±0.9	29.3±0.6	19.0±1.4	26.7±1.3
+ SGS	29.0±1.0	28.9±1.6	28.6±0.9	**21.3**±1.8	27.0±1.3	30.6±2.0	32.0±1.4	29.4±1.4	20.1±1.0	28.0±1.5

4.2 Performance Comparison

We compare SGS under two IPC settings of 10 and 20 against whole-dataset training (*Whole*) as an upper bound, coreset selection (*Random*, *Herding* (Welling, 2009)), and DM-based methods (*DM* (Zhao & Bilen, 2023), *DataDAM* (Sajedi et al., 2023), *M3D* (Zhang et al., 2024), *HDD-DM* (Li et al., 2025)), and integrate SGS into both *DM* and *HDD-DM*. For a fair comparison, we disable the multi-formation parameterization (Kim et al., 2022) of *M3D*, so that all methods use the same number of directly optimized synthetic images. The best and second-best results are highlighted in **bold** and underlined, respectively.

Multi-domain Generalization Performance. We evaluate the proposed SGS primarily in the multi-domain generalization (MDG) setting using a leave-one-domain-out protocol. As reported in Table 1, we observe that recent state-of-the-art DD methods, such as *DataDAM* and *M3D*, often struggle in this multi-domain distillation setting with poor generalization performance. In several cases, they even fall behind the vanilla DM baseline. This observation aligns with the findings of Choi et al. (2025) that conventional DD methods are fundamentally designed under the assumption of a single, homogeneous source domain. When distilling a multi-domain source dataset, their matching objectives indiscriminately entangle conflicting domain characteristics, leading to suboptimal distillation. In contrast, the proposed SGS consistently improves the average OOD accuracy of its underlying DM methods. Interestingly, the performance margins vary depending on the synthetic dataset budget. At a highly constrained budget (IPC=10), SGS yields substantial absolute

Table 2: Single-domain generalization performance on the Digits-DG dataset with IPC=10 and IPC=20.

IPC	10					20				
Source	M	N	V	Y	Avg.	M	N	V	Y	Avg.
Whole	45.3±0.7	62.3±0.6	64.6±0.8	70.4±0.3	60.7±0.6	45.3±0.7	62.3±0.6	64.6±0.8	70.4±0.3	60.7±0.6
Random	38.1±1.0	30.1±0.6	29.1±1.0	33.2±0.5	32.6±0.8	39.7±0.7	49.7±1.2	41.4±2.1	41.6±2.0	43.1±1.5
Herding	38.9±1.2	48.5±0.7	39.7±0.7	36.9±0.8	41.0±0.9	40.4±0.5	51.2±1.1	45.4±1.1	43.1±1.2	45.0±1.0
DM	42.6±0.8	54.7±0.6	45.0±1.3	54.8±0.9	49.3±0.9	42.1±1.5	57.2±0.8	50.0±0.8	55.3±0.7	51.2±1.0
+SGS	**44.6**±0.8	**55.6**±0.1	**52.7**±1.0	**55.4**±0.7	**52.1**±0.7	**46.0**±1.5	**58.3**±0.6	**54.0**±0.3	55.8±0.7	**53.5**±0.8
HDD-DM	42.7±0.5	54.7±0.5	46.4±1.6	50.4±1.0	48.6±0.9	42.5±0.8	56.6±0.2	49.4±0.6	53.3±0.4	50.5±0.5
+SGS	43.4±0.8	**55.6**±0.4	50.4±0.8	53.9±0.8	50.8±0.7	44.3±0.8	58.0±0.8	53.2±1.2	**56.7**±0.6	53.1±0.9

Table 3: Performance comparison with post-hoc DG methods on PACS at IPC=10.

Target	P	A	C	S	Overhead
DM	48.0±0.7	27.9±0.5	33.8±1.5	32.6±2.2	1.0×
+FACT†	45.3±1.4	31.6±0.4	**35.5**±1.5	34.3±2.0	2.1×
+MixStyle†	47.1±1.0	27.8±0.5	33.9±1.0	31.8±2.6	1.2×
+AdvFreq†	41.4±2.2	**31.7**±0.5	31.5±0.5	**40.5**±0.6	6.8×
+SGS	**53.8**±0.4	28.4±1.1	35.1±1.5	37.2±1.8	**1.0**×

† Post-hoc methods applied on top of the DM-distilled dataset.

Table 4: Effect of g^{class}, g^{domain}, and g, evaluated on Digits-DG at IPC=10.

g^{class}	g^{domain}	g	M	N	V	Y	Avg.
		✓	85.3±0.7	51.4±1.5	53.6±0.9	69.9±0.9	65.1±1.0
✓			85.5±0.9	52.6±1.2	52.4±1.3	68.9±1.5	64.9±1.2
	✓		65.8±2.5	42.5±1.0	27.2±0.7	43.1±1.2	44.7±1.4
✓	✓		88.6±0.9	51.8±1.0	52.7±0.8	67.1±0.5	65.1±0.8
✓	✓	✓	**90.3**±0.4	**54.1**±0.9	**55.3**±0.4	**71.0**±0.8	**67.7**±0.6

gains (e.g., +2.6% for DM on Digits-DG, +3.4% for HDD-DM on CORe50-Hard), demonstrating its exceptional data efficiency. Furthermore, on datasets with severe visual domain shifts like PACS, SGS yields robust and consistent absolute gains across both IPC settings (+3.0% and +2.2% over DM and HDD-DM at IPC=10, and +2.1% and +2.6% at IPC=20). We attribute this to the fact that simply increasing the data budget is insufficient to overcome severe domain shifts without explicitly separating domain-consistent from domain-specific information, a capability uniquely provided by SGS. Overall, these results indicate that SGS provides stable OOD improvements across diverse dataset scales and resolutions, firmly validating its plug-and-play compatibility across different DM architectures.

Single-domain Generalization Performance. We further evaluate SGS in the single-domain generalization (SDG) setting, where only one source domain is available for distillation. To broaden the applicability of SGS, we partition the source dataset into K pseudo-domains. Directly motivated by the empirical observation in Figure 2 that style statistics effectively cluster by domain, we extract these statistics from intermediate feature maps and apply K-means clustering to identify latent domain-specific patterns (detailed procedures are provided in Appendix F.1). These pseudo-domains then serve as a proxy for ground-truth domain splits, enabling the SGS. Table 2 reports the average top-1 accuracy across the three unseen target domains for each choice of source domain, demonstrating consistent OOD improvements across all source domains for both DM baselines. These results indicate that the pseudo-domains identified through feature-statistic clustering, while inherently noisier than ground-truth domain labels, provide a highly effective proxy for the cross-domain agreement signal, allowing SGS to operate robustly even in the constrained single-source setting. Additional results, including visualizations and analysis of K, are provided in Appendix F.2.

Comparison with Post-hoc Domain Generalization. We also compare the performance by applying existing DG methods (*FACT* (Xu et al., 2021), *MixStyle* (Zhou et al., 2021), *AdvFreq* (He et al., 2024)) post-hoc during downstream training. However, as shown in Table 3, this approach struggles to guarantee consistent improvements. Because conventional DG methods assume the high diversity of large-scale real data, applying them to highly compressed distilled datasets often yields marginal gains or even degrades performance (e.g., *MixStyle*). Furthermore, complex online augmentations in methods like *FACT* and *AdvFreq* drastically increase downstream training time by 2.1× and 6.8×, fundamentally contradicting the efficiency goal of dataset distillation. In contrast, SGS inherently embeds domain-invariant signals during the offline distillation process. Thus, it achieves the best average OOD generalization without incurring any additional computational overhead (1.0×) during downstream training.

Table 5: Distillation cost per 100 iterations.

K	w/o	2	3	4	5
DM (s)	6.08	13.49	19.78	25.49	29.80
DM (MiB)	276	434	437	438	439

Table 6: Near-compute-matched comparison.

Method	Iter.	M	N	V	Y	Avg.
DM	20k	42.6	54.7	45.0	**54.8**	49.3
SGS	10k	**43.4**	**55.0**	**47.9**	54.2	**50.1**

4.3 Ablation Study

To analyze the contribution of each gradient component in Eq. 10, we conduct an ablation study on Digits-DG (IPC=10). The results are reported in Table 4, where each column represents the OOD accuracy for a specific unseen target domain. When using $\boldsymbol{g}$ alone, SGS reduces to standard DM and serves as the baseline. Replacing $\boldsymbol{g}$ with $\boldsymbol{g}^{\text{class}}$ alone yields comparable OOD accuracy on M and N but slightly degrades performance on V and Y, suggesting that the cross-domain agreement signal alone is insufficient as a primary update direction. This is also reflected in Figure 3b, where the distilled images exhibit clear class semantics but lack visual diversity, sharing highly uniform domain characteristics across samples. In contrast, relying solely on $\boldsymbol{g}^{\text{domain}}$ causes severe performance degradation across all target domains. As shown in Figure 3c, while these images capture domain-specific appearances—which inherently include some recognizable digit-like structures—they fail to encode fine-grained class-discriminative features required for accurate classification. Combining only the two SGS signals ($\boldsymbol{g}^{\text{class}} + \boldsymbol{g}^{\text{domain}}$, without $\boldsymbol{g}$) improves OOD accuracy on M and N, but falls short of the vanilla DM baseline on V and Y, indicating that the standard DM gradient $\boldsymbol{g}$ is indispensable as the primary optimization direction. Finally, combining all three components yields the best performance across every target domain. This demonstrates that $\boldsymbol{g}^{\text{class}}$ and $\boldsymbol{g}^{\text{domain}}$ provide essential and complementary refinements to the primary signal $\boldsymbol{g}$. As visualized in Figure 3a, the resulting distilled images successfully preserve faithful class-discriminative information while exhibiting rich cross-domain visual diversity.



(a) (b) (c)

Figure 3: Visualization of distilled Digits-DG images (IPC=10) under different gradient update strategies: (a) SGS, (b) $\boldsymbol{g}^{\text{class}}$ only, and (c) $\boldsymbol{g}^{\text{domain}}$ only.

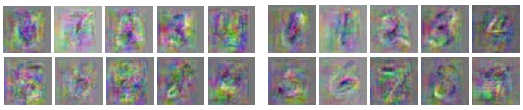
4.4 Further Analysis

Offline Distillation Cost. SGS computes one gradient per domain, so its distillation time grows roughly linearly with the number of domains K. Table 5 reports distillation time in seconds and peak GPU memory in MiB, where w/o denotes the base method. At $K = 3$, as in MDG, SGS increases distillation time by 3.25× over DM, while downstream training cost remains unchanged. To test whether the gain merely reflects extra compute, we run SGS with $K = 2$ for half the iterations under SDG on Digits-DG (IPC=10), which nearly matches the total distillation time of DM. As shown in Table 6, SGS still outperforms DM on average (50.1% vs. 49.3%), suggesting that its benefit is not solely due to additional compute.

Alignment with Unseen Domains. To examine whether $\boldsymbol{g}^{\text{class}}$ transfers beyond the source domains, we additionally compute the DM gradient of the held-out target domain, $\boldsymbol{g}^{\text{tgt}}$, during distillation on PACS under the leave-one-domain-out protocol; $\boldsymbol{g}^{\text{tgt}}$ is used only for measurement and never for optimization. Table 7 reports the cosine similarity between $\boldsymbol{g}^{\text{tgt}}$ and each update together with the OOD accuracy gain of SGS over DM from Table 1. $\boldsymbol{g}^{\text{class}}$ is more aligned with $\boldsymbol{g}^{\text{tgt}}$ than $\boldsymbol{g}$ on every target (+5.5 points on average), whereas $\boldsymbol{g}^{\text{domain}}$ is nearly orthogonal to it, consistent with its role as a diversity term rather than a transferable update. The alignment gain also tracks the accuracy gain: target A shows the smallest improvement in both (+0.9 and +0.5).

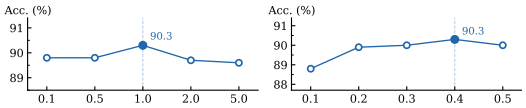
Table 7: Cosine similarity (%) with the held-out target gradient on PACS.

Target	P	A	C	S	Avg.
$\cos(\boldsymbol{g}, \boldsymbol{g}^{\text{tgt}})$	24.8	17.9	20.1	21.6	21.1
$\cos(\boldsymbol{g}^{\text{class}}, \boldsymbol{g}^{\text{tgt}})$	**31.8**	**18.8**	**27.8**	**27.8**	**26.6**
$\cos(\boldsymbol{g}^{\text{domain}}, \boldsymbol{g}^{\text{tgt}})$	0.0	0.0	0.0	0.0	0.0
Acc. gain	+5.8	+0.5	+1.3	+4.6	+3.0



(a) $\boldsymbol{g}$ (b) $\boldsymbol{g}^{\text{class}}$

Figure 4: Visualization of $\boldsymbol{g}$ and $\boldsymbol{g}^{\text{class}}$ on Digits-DG (Upper row: 0 to 4, Lower row: 5 to 9).



(a) λ_c (b) λ_d

Figure 5: Performance changes according to (a) λ_c and (b) λ_d on Digits-DG.

Extension to Gradient Matching. Although we focus on DM, SGS only requires image-space gradients of a matching loss with respect to the synthetic images. We therefore apply it to DC (Zhao & Bilen, 2021b), a gradient-matching method, by computing the outer gradient of its matching loss with respect to the synthetic pixels separately for each source domain and applying the same spectral decomposition. As shown in Table 8, SGS improves DC on all four target domains under MDG, with an average gain of 0.7 points. We regard this as preliminary evidence that SGS extends beyond DM; extending it to trajectory matching, which requires domain-wise gradients through unrolled model updates, is left for future work.

Table 8: SGS on DC (Digits-DG, IPC=10).

Method	M	N	V	Y	Avg.
DC	87.8	55.8	51.2	71.2	66.5
DC + SGS	**89.0**	**56.4**	**51.7**	**71.5**	**67.2**

Visualization of $\boldsymbol{g}$ and $\boldsymbol{g}^{\text{class}}$. To verify whether the proposed spectral decomposition effectively isolates class-discriminative information, we visualize the standard DM gradient $\boldsymbol{g}$ and the class-discriminative signal $\boldsymbol{g}^{\text{class}}$ extracted by SGS in Figure 4. While $\boldsymbol{g}^{\text{class}}$ exhibits sharper, highly recognizable digit structures, $\boldsymbol{g}$ displays visually noisier patterns, where the underlying digit shapes are partially obscured by textural variations caused by domain-specific signals. This qualitative contrast strongly supports our core motivation discussed in Section 3.2: while the standard DM gradient inevitably entangles class-discriminative and domain-specific information, $\boldsymbol{g}^{\text{class}}$—derived from cross-domain agreement in the spectral domain—yields a clearer class-relevant signal.

In-distribution Performance. Table 9 reports the ID performance on Digits-DG at IPC=10, where the test domains coincide with the source domains used for distillation. We observe that SGS yields ID performance comparable to that of the underlying base methods on average, despite substantially improving OOD generalization performance.

Table 9: In-distribution performance on Digits-DG (IPC=10) under the MDG setting.

Target	M	N	V	Y	Avg.
DM	73.6±0.5	82.8±0.6	82.3±0.4	77.7±0.4	79.1±0.5
+ SGS	73.7±0.9	82.3±0.8	81.7±0.4	77.9±0.5	78.9±0.7
HDD-DM	74.1±0.6	83.2±0.4	82.1±0.3	77.5±0.2	79.2±0.4
+ SGS	74.4±0.4	82.4±0.2	83.4±0.2	77.5±0.6	79.4±0.4

Hyperparameter Sensitivity. Figure 5 shows the OOD accuracy on Digits-DG (IPC=10, target M) with respect to λ_c and λ_d. SGS consistently outperforms DM across the entire range, and the trends reflect a trade-off between the two components. As λ_c increases, the update relies more on cross-domain consensus, which tends to reduce the visual diversity of synthetic samples (Figure 3b). Conversely, a larger λ_d emphasizes domain-specific characteristics, which can eventually weaken the class-discriminative information needed for accurate classification, indicating that a balance between the two is essential.

5 Conclusion

We studied dataset distillation under domain shift, where the distilled dataset is evaluated on target domains excluded from distillation. We proposed Spectral Gradient Surgery (SGS), which augments the standard DM update with a class-relevant component, obtained by weighting the per-frequency consensus of domain-wise gradients by their phase agreement, and a domain-relevant component that diversifies the synthetic set. Without modifying the DM objective, SGS consistently improves the average OOD accuracy of DM and HDD-DM on Digits-DG, PACS, and CORe50-Hard, at the cost of additional offline distillation time. Our analysis further shows that the class-relevant component is better aligned with gradients of unseen domains than the original DM gradient. Limitations are discussed in Appendix A.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10718–10727, 2022.

Jaehyun Choi, Gyojin Han, Dong-Jae Lee, Sunghyun Baek, and Junmo Kim. Dam: Domain-aware module for multi-domain dataset condensation. *arXiv preprint arXiv:2505.22387*, 2025.

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6565–6590, 2023.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

Xilin He, Jingyu Hu, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Muhammad Haris Khan, and Linlin Shen. Towards combating frequency simplicity-biased learning for domain generalization. *Advances in Neural Information Processing Systems*, 37:31078–31102, 2024.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Wenyuan Li, Guang Li, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Hyperbolic dataset distillation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on robot learning*, pp. 17–26. PMLR, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Minyoung Oh and Jae-Young Sim. Asymmetric synthetic data update for domain incremental dataset distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 18347–18377, 2022.

Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z. Liu, Yuri A. Lawryshyn, and Konstantinos N. Plataniotis. DataDAM: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17097–17107, 2023.

Yuge Shi, Jeffrey Seely, Philip HS Torr, Narayanaswamy Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1121–1128, 2009.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14383–14392, 2021.

Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. M3d: Dataset condensation by minimizing maximum mean discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 12674–12685, 2021a.

Bo Zhao and Hakan Bilen. Dataset condensation with gradient matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6514–6523, 2023.

Guangtao Zheng, Mengdi Huai, and Aidong Zhang. Advst: Revisiting data augmentations for single domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21832–21840, 2024.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pp. 561–578, 2020.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

APPENDIX

LLM Usage Declaration Large language models (LLMs) were used solely for language refinement, and all research content was generated entirely by the authors.

A LIMITATIONS

Computational cost. SGS computes a separate gradient for each of the K domains at every distillation iteration, so its distillation time grows approximately linearly with K. In our MDG setting with three source domains, SGS increases the distillation time by 3.25× over DM (Table 5). Although this overhead is confined to offline distillation and the near-compute-matched comparison (Table 6) suggests that the gain is not solely due to additional compute, the cost may become significant when the number of source domains is large. Reducing this overhead, for example by sharing computation across domains or by sampling a subset of domains at each iteration, is an important direction for future work.

Extension to other DD paradigms. SGS is designed and validated primarily within the DM framework. Since it only requires image-space gradients of a matching loss with respect to the synthetic images, the same decomposition is conceptually applicable to other DD paradigms, but this has been examined only in a preliminary manner. For gradient matching, applying SGS to DC improves all four target domains on Digits-DG (Table 8), but this result is limited to a single dataset and IPC. For trajectory matching, domain-wise gradients would need to be computed through multiple unrolled model updates, which may mix domain-specific signals across steps and multiply computation and memory by the number of domains; this extension remains untested. A systematic study of SGS across DD paradigms is left for future work.

Interpretation of the decomposition. The class-relevant component is defined by phase agreement among the *source* domains, and is therefore not a strictly semantic filter. A component that is consistently shared by all source domains passes through the filter regardless of whether it is related to class identity; for example, a nuisance pattern common to every source domain would be retained. Our evidence that the class-relevant component captures transferable information is thus empirical, based on its improved alignment with held-out target-domain gradients (Table 7) and on qualitative visualizations (Figure 4), rather than a formal guarantee of semantic disentanglement.

B SPECIFICATION OF DATASETS

Table 10 summarizes the detailed statistics of the three benchmark datasets evaluated in our experiments. Furthermore, to illustrate the visual domain shifts present in these datasets, we provide representative examples in Figure 6, where rows and columns correspond to distinct domains and classes, respectively.

Table 10: Detailed statistics of the benchmark datasets evaluated in our experiments.

	Digits-DG (Zhou et al., 2020)	PACS (Li et al., 2017)	CORe50-Hard (Lomonaco & Maltoni, 2017)
# Domains	4	4	4
# Classes	10	7	10
Domains	MNIST (LeCun et al., 2002) MNIST-M (Ganin & Lempitsky, 2015) SVHN (Netzer et al., 2011) SYN (Ganin & Lempitsky, 2015)	Photo Art painting Cartoon Sketch	s5 s7 s9 s10
# Images per domain	6,000 / 6,000 / 6,000 / 6,000	1,670 / 2,048 / 2,344 / 3,929	14,966 / 14,994 / 14,994 / 14,986
Resolution	32×32	64×64	128×128

CORe50-Hard. CORe50-Hard consists of ten CORe50 object categories: plug adapters, mobile phones, scissors, light bulbs, cans, glasses, balls, markers, cups, and remote controls. Among the CORe50 sessions, we select the four sessions farthest from the overall mean in amplitude-spectrum space (s5, s7, s9, and s10), which differ in indoor/outdoor conditions, backgrounds, and illumination. No additional sample-level selection is applied: under MDG, all samples of the three source sessions are used for distillation, and all samples of the held-out session are used for evaluation.

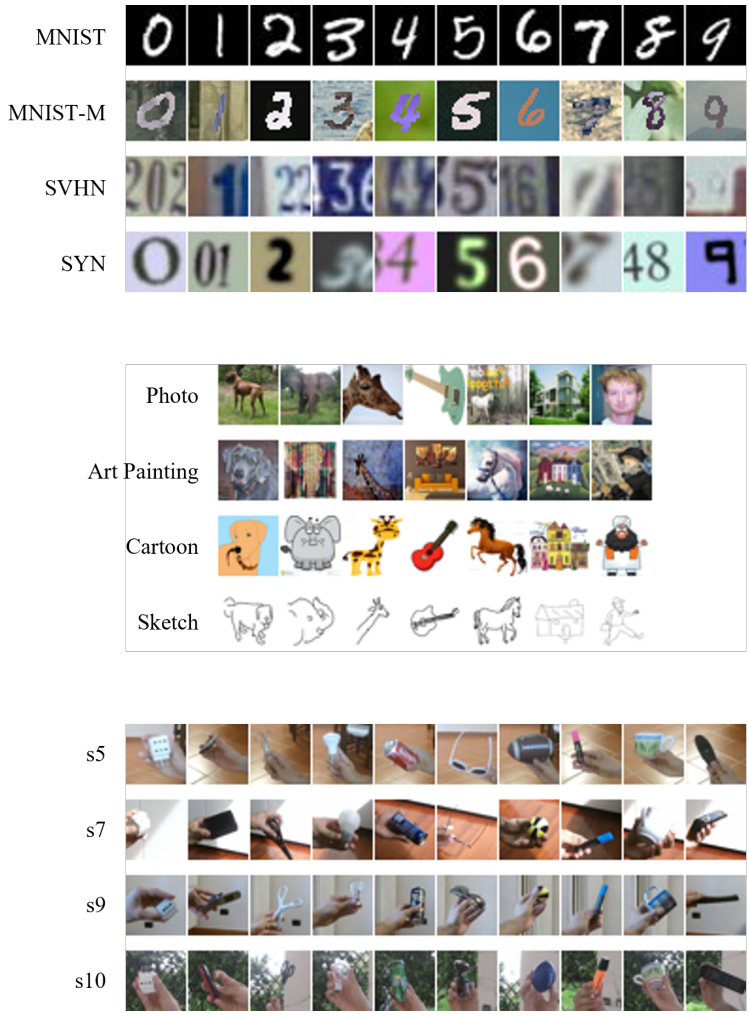


Figure 6: Visual examples from the three benchmark datasets. For each dataset, rows correspond to distinct domains, while columns represent different classes. Note the severe visual domain shifts across the source domains.

C Theoretical Analysis of SGS

We provide a theoretical analysis of how the resultant length r_i in Eq. 9 preserves spectral components agreed across source domains while attenuating those that are not.

C.1 Setup

For a synthetic sample $\hat{\boldsymbol{x}}_i$ and a specific frequency component, we drop both indices for brevity. Let $G^s = A^s e^{j\phi^s}$ denote the spectral gradient of source domain $s \in \{1, \ldots, S\}$, where A^s and ϕ^s are its magnitude and phase. Following Eq. 7–9, we have

$$\bar{G} = \frac{1}{S}\sum_{s=1}^{S} G^s, \qquad r = \frac{\left|\sum_{s=1}^{S} G^s\right|}{\sum_{s=1}^{S} A^s + \epsilon}, \qquad G^{\text{class}} = r \cdot \bar{G}. \tag{11}$$

C.2 Main Result

(i) Preservation of agreed components. If all domain-wise gradients share the same phase, then $|\sum_s G^s| = \sum_s A^s$, and thus $r = 1$ when ignoring ϵ. Consequently, $G^{\text{class}} = \bar{G}$, i.e., components with full cross-domain agreement are preserved without attenuation.

(ii) Attenuation of non-agreed components. Suppose that the magnitudes are bounded as $A^s \in [m, M]$ with $0 < m \leq M$, and that the phases ϕ^s are independent across domains, uniformly distributed on $[0, 2\pi)$, and independent of the magnitudes. Then

$$\mathbb{E}\big[|G^{\text{class}}|\big] \leq \frac{M^2}{m} \cdot \frac{1}{S}. \tag{12}$$

Proof of (ii). Let $Z = \sum_{s=1}^{S} G^s$. By definition,

$$|G^{\text{class}}| = \frac{|Z|}{\sum_s A^s + \epsilon} \cdot \frac{|Z|}{S} \leq \frac{|Z|^2}{S^2 m}, \tag{13}$$

where the inequality uses $\sum_s A^s \geq Sm$. Expanding $|Z|^2 = \sum_s \sum_t G^s \overline{G^t}$, every cross term with $s \neq t$ has zero expectation, since $\mathbb{E}[e^{j\phi^s}] = 0$ for a uniform phase and the domains are independent. Therefore,

$$\mathbb{E}|Z|^2 = \sum_{s=1}^{S} \mathbb{E}\big[(A^s)^2\big] \leq SM^2. \tag{14}$$

Combining the two gives $\mathbb{E}|G^{\text{class}}| \leq SM^2/(S^2 m) = M^2/(mS)$. □

Interpretation. In the same setting, the plain mean satisfies $\mathbb{E}|\bar{G}|^2 = \mathbb{E}|Z|^2/S^2 \geq m^2/S$, so its magnitude decays only at rate $1/\sqrt{S}$. Weighting by r thus further suppresses components without cross-domain agreement, reducing the decay rate to $\mathcal{O}(1/S)$, while leaving fully agreed components intact.

D Additional analysis

Table 11 further compares SGS with controls on Digits-DG (IPC=10), each differing from DM or SGS in a single factor. Rescaling $\boldsymbol{g}$ to the norm of the SGS update at every iteration (norm-matched) improves DM by only 0.6 points, indicating that the gain of SGS is not explained by a larger step size. Adding $\boldsymbol{g}^{\text{class}}$ or $\boldsymbol{g}^{\text{domain}}$ alone to $\boldsymbol{g}$ improves the average by 0.6 and 1.2 points, respectively, and combining both yields the best accuracy on every target. Measuring agreement by the sign of pixel-space gradients improves DM by 1.3 points, while spectral phase agreement adds a further 1.3 points.

Table 11: Controlled comparison on Digits-DG (IPC=10).

Update	M	N	V	Y	Avg.
$\boldsymbol{g}$ (DM)	85.3	51.4	53.6	69.9	65.1
$\boldsymbol{g}$, norm-matched	85.1	53.7	54.0	70.1	65.7
$\boldsymbol{g} + \lambda_c \boldsymbol{g}^{\text{class}}$	86.9	52.9	53.4	69.7	65.7
$\boldsymbol{g} + \lambda_d \boldsymbol{g}^{\text{domain}}$	89.5	52.3	54.8	68.6	66.3
SGS, pixel-space	90.0	51.7	54.0	70.0	66.4
SGS	**90.3**	**54.1**	**55.3**	**71.0**	**67.7**

E Additional Implementation Details

E.1 Baselines

M3D. For a fair comparison, we disable the multi-formation parameterization (Kim et al., 2022) of M3D, which generates multiple training images from each stored image, so that all methods are compared with the same number of directly optimized synthetic images.

DAM. We do not include DAM (Choi et al., 2025) in our quantitative comparison, as its official implementation is not publicly available.

Table 12: Hyperparameters for each split. MDG: λ_c/λ_d for each held-out target. SDG (Digits-DG): $\lambda_c/\lambda_d/K$ for each source.

Dataset	Target	IPC=10	IPC=20
Digits-DG	M	1/0.4	1/0.4
	N, V	2/0.1	2/0.1
	Y	2/0.1	3/0.1
PACS	P, A	1/0.1	2/0.1
	C	0.5/0.1	1/0.1
	S	1.5/0.1	2/0.1
CORe50-Hard	s5	4/0.1	4/0.1
	s7	3/0.1	4/0.1
	s9	3/0.1	3/0.1
	s10	4/0.1	5/0.1

Source	IPC=10	IPC=20
M	1/0.1/3	1/0.1/3
N	1/0.2/3	1/0.05/3
V	1/0.1/3	1/0.1/3
Y	1/0.1/3	1/0.2/3

E.2 Hyperparameter Selection

All hyperparameters are selected using source-domain gradient statistics only, without access to target-domain images, labels, or accuracy. For each split, we monitor the relative magnitudes of the two SGS components with respect to the original DM gradient,

$$R_c = \lambda_c \frac{\|\boldsymbol{g}^{\text{class}}\|}{\|\boldsymbol{g}\|}, \qquad R_d = \lambda_d \frac{\|\boldsymbol{g}^{\text{domain}}\|}{\|\boldsymbol{g}\|}. \tag{15}$$

We choose λ_c such that $R_c \approx 0.5$, 1.0, and 2.0 for Digits-DG, PACS, and CORe50-Hard, respectively, following their increasing resolutions, and λ_d such that $R_d \approx 0.2$ for all datasets. In MDG, the number of source domains is always three, as determined by the protocol. In SDG, we use $K = 3$ pseudo-domains for all source domains and IPCs. Table 12 lists the resulting values.

F Extension to Single Domain Generalization

F.1 Implementation Details

In practice, $\mathcal{D}$ may not provide explicit domain splits, which corresponds to the single domain generalization (SDG) setting. To extend SGS to this setting, we estimate pseudo-domain labels $\{1, \ldots, K\}$ to substitute for ground-truth domain labels. Specifically, motivated by the observation in Figure 2 that feature statistics encode domain-specific information, we extract channel-wise mean and standard deviation of intermediate feature maps from ψ for each sample, and apply K-means clustering (Hartigan & Wong, 1979) over these style statistics to partition $\mathcal{D}$ into K pseudo-domains, which are then used in place of ground-truth domain splits throughout SGS.

F.2 More Results in SDG Setting

Sensitivity of K. We analyze the sensitivity of the hyperparameter K, which determines the number of pseudo-domains, on the Digits-DG (IPC=10, Source: M).

Figure 7 illustrates the average OOD accuracy across different values of K. We observe that the OOD performance generally improves as K increases. This empirical trend is consistent with our theoretical finding in Appendix C, which states stronger suppression of domain-specific signals with a larger number of source domains. However, a larger K linearly scales up the distillation time overhead, as gradients must be computed independently for each pseudo-domain. Furthermore, an excessively large K risks isolating outliers into overly sparse clusters, which injects noisy gradient signals and ultimately degrades the distillation performance.

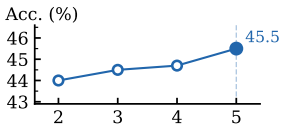
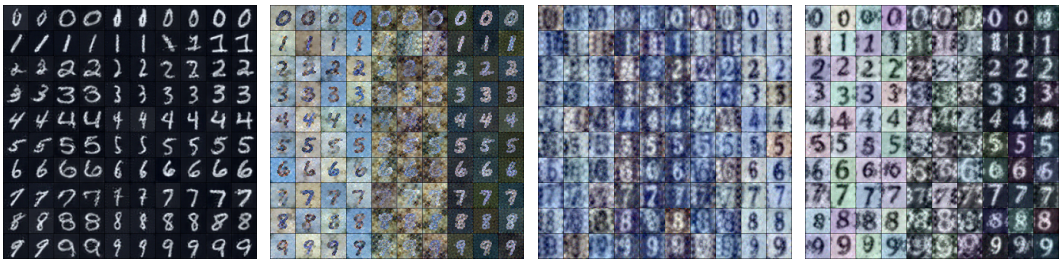


Figure 7: Performance changes according to K.

Visualization. Figure 8 visualizes the distilled images under the Single Domain Generalization (SDG) setting for each source domain, using an IPC of 10 and $K = 3$ pseudo-domains. For the M source domain, the generated images reveal that each pseudo-domain captures distinct intra-domain variations, such as different writing styles, and digit angles. For the other source domains, the pseudo-domains exhibit clear visual distinctions, successfully grouping diverse color palettes and background textures. These visualizations suggest that the pseudo-domain clustering captures meaningful sub-populations within a single source domain.



(a) M (b) N (c) V (d) Y

Figure 8: Visualization of distilled images on the Digits-DG under the SDG setting (IPC=10).

Quantitative Results on PACS. Table 13 presents the SDG results on the PACS dataset (IPC=10, 20). Consistent with our main findings, SGS outperforms baseline methods in most cases, demonstrating its strong generalization capability even when trained on a single source domain.

Table 13: Single-domain generalization performance on PACS with IPC=10 and IPC=20.

IPC	10					20				
Source	P	A	C	S	Avg.	P	A	C	S	Avg.
DM	23.6±0.8	36.7±0.6	32.5±0.8	21.7±1.5	28.6±0.9	27.0±0.9	37.9±1.4	34.5±0.6	19.3±0.8	29.7±0.9
+SGS	**29.0**±0.7	**39.6**±1.4	33.7±1.1	**22.5**±1.0	**31.2**±1.1	**28.2**±1.0	**38.3**±1.0	34.2±0.9	20.4±0.8	30.3±0.9
HDD-DM	25.1±0.4	35.7±0.9	31.5±0.9	21.2±0.2	28.4±0.6	24.3±0.9	36.4±1.4	34.7±0.8	19.5±1.4	28.7±1.1
+SGS	27.2±0.8	37.4±1.5	**34.6**±0.4	**22.5**±1.3	30.4±1.0	26.9±1.3	37.9±1.6	**36.5**±0.6	**23.4**±1.0	**31.2**±1.1

G Initialization Strategy

We study performance variations under different initialization strategies: *Noise*, *Random*, and *Uniform*. *Noise* initializes synthetic images with random noise. *Random* assigns synthetic images by randomly sampling from the entire source dataset. *Uniform* distributes synthetic images evenly across each domain dataset. Table 14 presents the results on Digits-DG and PACS for IPC=10 and IPC=20. *Uniform* achieves the highest average accuracy in all settings. We attribute this to increased diversity: uniform allocation prevents the synthetic set from over-representing any single source domain.

Table 14: Performance of SGS according to different initialization strategies.

Dataset	Digits-DG									
IPC	10					20				
Target	M	N	V	Y	Avg.	M	N	V	Y	Avg.
Noise	89.9±0.5	50.6±1.6	52.5±1.1	**71.4**±0.5	66.1±0.9	89.8±0.6	53.2±1.6	**58.2**±1.0	73.6±0.8	68.7±1.0
Random	89.4±0.8	51.4±1.9	53.9±1.1	70.1±1.2	66.2±1.3	89.5±0.5	52.8±0.7	57.4±1.3	**75.1**±0.2	68.7±0.7
Uniform	**90.3**±0.4	**54.1**±0.9	**55.3**±0.4	71.0±0.8	**67.7**±0.6	**90.7**±0.5	**54.3**±1.2	58.0±0.5	74.3±0.8	**69.3**±0.8

Dataset	PACS									
IPC	10					20				
Target	P	A	C	S	Avg.	P	A	C	S	Avg.
Noise	52.8±0.7	27.6±0.6	33.4±1.7	**37.8**±4.6	37.9±1.9	52.6±1.3	**31.1**±0.7	35.1±1.3	37.7±2.5	39.1±1.5
Random	52.6±3.0	**28.4**±1.1	31.6±1.2	34.1±3.7	36.7±2.3	52.9±0.7	28.6±0.8	34.1±0.8	**40.1**±1.4	38.9±0.9
Uniform	**53.8**±0.4	**28.4**±1.1	**35.1**±1.5	37.2±1.8	**38.6**±1.2	**53.2**±1.0	29.1±0.6	**37.7**±2.3	39.8±2.5	**40.0**±1.6

H More Results of Ablation Study

We provide an extended ablation study on the Digits-DG and PACS to further validate the contribution of each gradient component (Table 15). The results show that the full update ($\boldsymbol{g}+\boldsymbol{g}^{\text{class}}+\boldsymbol{g}^{\text{domain}}$) achieves the highest average OOD accuracy in all settings. Using only $\boldsymbol{g}^{\text{domain}}$ significantly degrades performance, indicating that it mainly carries domain-specific variation rather than class information. Combining the two SGS components without $\boldsymbol{g}$ does not consistently outperform DM, confirming that $\boldsymbol{g}$ remains necessary as the primary update direction.

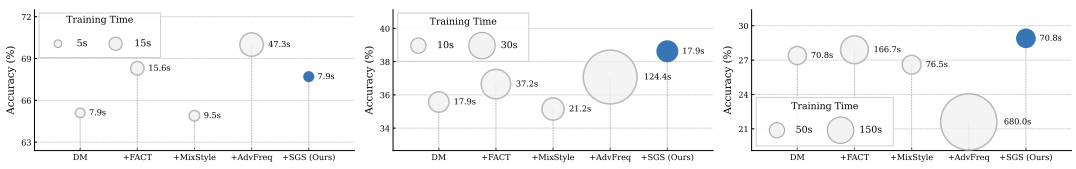
Table 15: Ablation study on Digits-DG and PACS when IPC=10 and IPC=20.

Dataset			Digits-DG									
IPC			10					20				
$\boldsymbol{g}^{\text{class}}$	$\boldsymbol{g}^{\text{domain}}$	$\boldsymbol{g}$	M	N	V	Y	Avg.	M	N	V	Y	Avg.
		✓	85.3±0.7	51.4±1.5	53.6±0.9	69.9±0.9	65.1±1.0	88.4±0.8	53.1±1.0	56.4±1.1	73.7±0.2	67.9±0.8
✓			85.5±0.9	52.6±1.2	52.4±1.3	68.9±1.5	64.9±1.2	88.6±0.9	52.6±1.3	56.5±1.2	74.2±1.5	68.0±1.2
	✓		65.8±2.5	42.5±1.0	27.2±0.7	43.1±1.2	44.7±1.4	68.9±5.3	43.6±1.0	37.7±0.8	52.8±1.2	50.8±2.1
✓	✓		88.6±0.9	51.8±1.0	52.7±0.8	67.1±0.5	65.1±0.8	86.9±0.9	51.2±1.7	57.0±0.8	72.1±0.5	66.8±1.0
✓	✓	✓	**90.3**±0.4	**54.1**±0.9	**55.3**±0.4	**71.0**±0.8	**67.7**±0.6	**90.7**±0.5	**54.3**±1.2	**58.0**±0.5	**74.3**±0.8	**69.3**±0.8

Dataset			PACS									
IPC			10					20				
$\boldsymbol{g}^{\text{class}}$	$\boldsymbol{g}^{\text{domain}}$	$\boldsymbol{g}$	P	A	C	S	Avg.	P	A	C	S	Avg.
		✓	48.0±0.7	27.9±0.5	33.8±1.5	32.6±2.2	35.6±1.2	49.2±0.4	28.8±1.0	35.4±2.2	38.1±2.7	37.9±1.6
✓			49.4±2.6	27.7±0.6	**35.9**±0.8	33.0±2.2	36.5±1.6	52.6±0.7	**29.1**±0.1	37.6±1.4	37.8±1.8	39.3±1.0
	✓		30.6±3.1	24.6±0.7	29.5±0.8	24.5±3.9	27.3±2.1	39.2±2.7	23.4±0.9	30.2±1.1	30.7±3.4	30.9±2.0
✓	✓		**54.7**±1.0	25.6±0.4	35.1±0.9	33.9±3.5	37.3±1.5	51.7±1.2	29.0±0.6	31.9±1.5	38.5±0.7	37.8±1.0
✓	✓	✓	53.8±0.4	**28.4**±1.1	35.1±1.5	**37.2**±1.8	**38.6**±1.2	**53.2**±1.0	**29.1**±0.6	**37.7**±2.3	**39.8**±2.5	**40.0**±1.6

I More Results of Post-hoc Domain Generalization

Figure 9 compares the average OOD accuracy and downstream training time of various post-hoc DG methods applied to DM-distilled datasets when IPC=10. Existing post-hoc methods reveal a strict trade-off between accuracy and efficiency. *MixStyle* is computationally light but consistently underperforms across all benchmarks. Conversely, augmentation-based methods like *FACT* and *AdvFreq* occasionally achieve competitive accuracy, but they suffer from high performance variance across datasets. More critically, their complex augmentation pipelines introduce severe computational overhead, slowing down the downstream training by nearly 2× (*FACT*) and 7× (*AdvFreq*). This directly undermines the core advantage of dataset distillation: accelerating downstream training. SGS achieves the highest average OOD accuracy on PACS and CORe50-Hard without downstream overhead. On Digits-DG, *FACT* and *AdvFreq* achieve higher accuracy than SGS, but at roughly 2× and 6× its downstream training time.

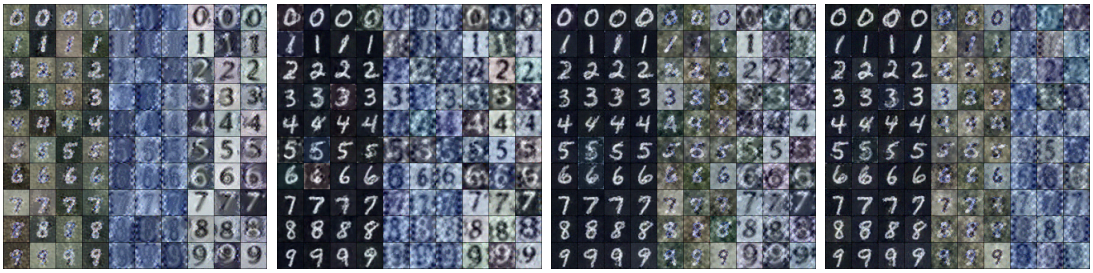


(a) Digits-DG (b) PACS (c) CORe50-Hard

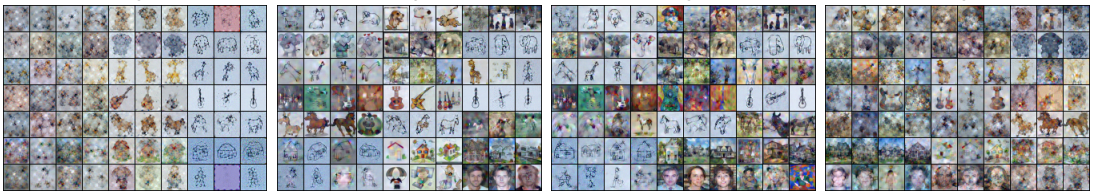
Figure 9: Comparison of average OOD accuracy and downstream training time among various post-hoc DG methods applied to DM-distilled datasets (IPC=10).

J VISUALIZATION OF DISTILLED IMAGES

Figure 10 and Figure 11 visualize distilled images via SGS upon DM on three benchmark datasets when IPC=10 under MDG and SDG settings, respectively.



(a) Target: M (b) Target: N (c) Target: V (d) Target: Y

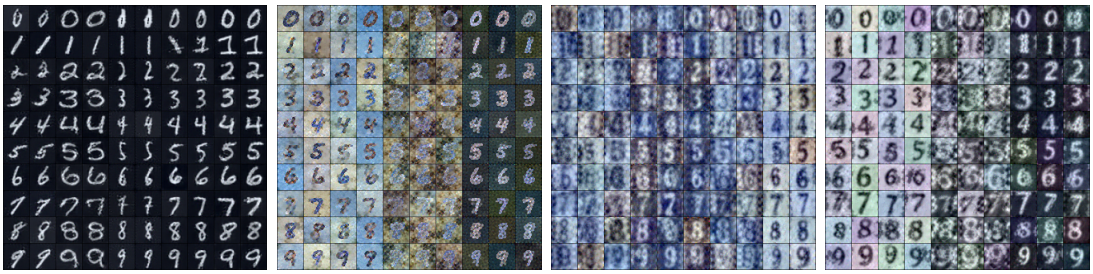


(e) Target: P (f) Target:A (g) Target:C (h) Target:S



(i) Target: s5 (j) Target: s7 (k) Target: s9 (l) Target: s10

Figure 10: Distilled Images of our proposed SGS upon DM when IPC=10 under MDG setting. Each row indicates Digits-DG, PACS, and CORe50-Hard, respectively, on each target domain.



(a) Source: M (b) Source: N (c) Source: V (d) Source: Y

(e) Source: P (f) Source: A (g) Source: C (h) Source: S

Figure 11: Distilled Images of our proposed SGS upon DM when IPC=10 under SDG setting. Each row indicates Digits-DG, and PACS, respectively, on each source domain.