

Decomposed Distribution Matching in Dataset Condensation

Sahar Rahimi Malakshan, Mohammad Saeed Ebrahimi Saadabadi,
Ali Dabouei, and Nasser M. Nasrabadi

sr00033, me00018, Ad0046@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu

Abstract

Dataset Condensation (DC) aims to reduce deep neural networks training efforts by synthesizing a small dataset such that it will be as effective as the original large dataset. Conventionally, DC relies on a costly bi-level optimization which prohibits its practicality. Recent research formulates DC as a distribution matching problem which circumvents the costly bi-level optimization. However, this efficiency sacrifices the DC performance. To investigate this performance degradation, we decomposed the dataset distribution into content and style. Our observations indicate two major shortcomings of: 1) style discrepancy between original and condensed data, and 2) limited intra-class diversity of condensed dataset. We present a simple yet effective method to match the style information between original and condensed data, employing statistical moments of feature maps as well-established style indicators. Moreover, we enhance the intra-class diversity by maximizing the Kullback–Leibler divergence within each synthetic class, i.e., content. We demonstrate the efficacy of our method through experiments on diverse datasets of varying size and resolution, achieving improvements of up to 4.1% on CIFAR10, 4.2% on CIFAR100, 4.3% on TinyImageNet, 2.0% on ImageNet-1K, 3.3% on ImageWoof, 2.5% on ImageNette, and 5.5% in continual learning accuracy. [Code](#)

1. Introduction

In response to the challenges imposed by the sheer amount of data in large-scale datasets, e.g., storage and computational burden, the concept of Dataset Condensation (DC) was introduced [6, 15, 57, 58, 67]. Pioneered by Wang *et al.* [58], DC utilizes a nested optimization to synthesize a small dataset that retains the effectiveness of the original dataset. Despite inspiring, their proposal was computationally intensive and infeasible for large-scale setups [6, 15, 57, 67]. Therefore, follow-up studies [6, 37, 57, 66, 68] try to circumvent the nested optimization of [58] by matching the training trajectories [6, 12] or gradients [37, 57, 66, 68] of surrogate models trained on condensed and original datasets. Although promising, relying on computationally extensive

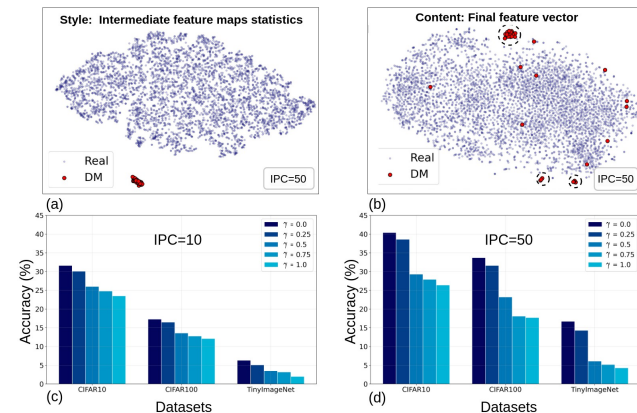


Figure 1. (a, b) 2D t-SNE visualizations of original and condensed images learned by DM [67] for CIFAR10 with IPC=50 in randomly chosen category. (a) Style statistics (concatenation of mean and variance) from the first layer’s feature map, highlighting a significant style discrepancy. (b) Final features of the DNN, showing limited diversity of instances learned by DM. (c, d) Illustrating the negative effect of style discrepancy on performance. During training the style of samples from Herding is [45] drifted toward that of DM [67], with γ representing the drift ratio.

bi-level optimization, i.e., an inner optimization for model updates and an outer one for condensed data updates, limits their practicality [6, 37, 67, 68].

Recently, Zhao *et al.* [67] circumvent the bi-level optimization by leveraging the distance-preserving property of representations obtained from randomly sampled Deep Neural Networks (DNNs), i.e., DNN with random weights [51]. Specifically, they formulate DC as a distribution matching problem between the original and condensed datasets in the embedding of randomly sampled DNNs, dubbed DM [67]. Utilizing random DNNs bypasses the inner optimization of bi-level methods [67]. Consequently, DM severely reduces the computational cost/time of DC since it only updates the condensed data. For instance, for 50 Images Per Class (IPC), DM condenses CIFAR10 $45\times$ faster than bi-level optimization methods like [66, 68].

Despite the efficiency of DM [67], its performance lags behind that of bi-level methods [6, 37, 57, 66, 68]. To study

this deficiency, motivated by the literature on distribution matching [26, 29, 36, 38, 39, 42, 63, 65, 71], we decompose the dataset distribution into two major factors: 1) style, including attributes like texture, and color, and 2) content, which encompasses the semantic information [16, 17, 26, 39]. We trained a Convolutional Neural Network (CNN) using original and condensed data learned by DM [72]. Then, we explored the content and style discrepancy between original and condensed data in the embedding of the trained CNN, shown in Figure 1.

The first and second moments of the intermediate feature maps of CNNs, *i.e.*, mean and variance, capture the style of the input image [28, 41, 65]. As Figure 1a illustrates, there is a significant style discrepancy across original and condensed data. Specifically, despite the similar content (same category), original and condensed data represent distinct styles. Style gap between training and testing results in severe performance degradation due to the DNNs bias toward style [25, 29, 44, 70, 72]. Furthermore, Figure 1c, and d illustrate the effect of this style gap in the performance obtained from condensed datasets. Specifically, during training, the style of samples from Herding coresets selection [45] *i.e.*, original styles, is drifted towards that of DM [67]. Please refer to Section A of Supplementary Materials for more details. As style drifts from the real, the performance decreases, reflecting the importance of style alignment between condensed and real datasets; in line with previous studies on dataset distribution [2, 19, 44, 65, 72].

Content information of the input image is reflected in the final feature embedding of DNNs [22, 26, 43, 47]. Figure 1b compares the t-SNE visualization of feature vectors of original and condensed data, showing no evident content gap between them. However, it reveals a lower intra-class diversity of condensed samples than the original. Specifically, condensed instances form local clusters in the embedding space, reflecting similar information, *i.e.*, low intra-class diversity [55]. DM’s training objective [67] explicitly promotes the content alignment between real and condensed datasets [48, 67] but discards the diversity. Thus, condensed data fails to adequately represent the original dataset’s extensive variability, leading to overfitting when used as a source of training data [1, 61].

Prior works for improving DM [48, 62, 69] either incur significant computational costs to its framework or employ a restricted spatial supervision that reduces the generalization [10, 30]. In this study, our key insight is that condensed data should: 1) express the distribution of original data in both style and content, 2) consist of diverse informative samples, and 3) be synthesized without a bi-level learning regime to be applicable to large scale setups. Concerning style disparity, in addition to the content alignment of DM, we propose a Style Matching (SM) module. SM module leverages well-established style indicators of feature map

moments and correlations to align the style across original and condensed data. Our proposal leverages feature maps from a randomly sampled DNNs, *i.e.* adhering to the computationally efficient framework of DM, to match the style between real and condensed sets.

To encourage intra-class diversity, we employ a criterion based on Kullback–Leibler (KL) divergence [35] to penalize samples that form a local cluster. Our proposal works in the embedding of a random DNN and encourages intra-class diversity while maintaining the plausibility of the samples and the computational efficiency of the DM framework. Our method demonstrates significant improvements across diverse datasets with low, medium, and high resolutions, including CIFAR10, CIFAR100, Tiny ImageNet, and ImageNet-1K, affirming its scalability and generalization from small to large scale datasets. Also, we show the generalization of the proposed method by evaluating on both simple ConvNet and the more sophisticated ResNet architectures. The contributions of the paper are as follows:

- We decompose the distribution matching framework in DC into two major factors: content and style, and reveal the shortcomings of DC in these factors.
- We identify the issue of the style gap between original and condensed data. Then we propose an optimization based on matching statistical moments of feature maps to reduce this style disparity.
- We identify the issue of limited intra-class diversity of the distribution matching process in DC. Then, we propose an optimization method specifically tailored to increase the intra-class diversity by penalizing the synthesized samples that express similar information.

2. Related Work

2.1. Coreset Selection

Coreset or instance selection is a heuristic method that approximates the full dataset by a small subset [14]. For instance, random selection [45] chooses samples arbitrarily; Herding [3, 5] selects samples nearest to each class’s cluster center; and Forgetting [54] identifies samples that are easily forgotten during training. Despite the advances in coreset selection methods, they fail to scale into large-scale setups due to the computational deficiency [23, 60]. Moreover, the heuristic criteria cannot guarantee the optimal solution for down-stream tasks [69]. Dataset condensation offer an alternative approach by synthesizing condensed data that can overcome the limitations of coreset selection methods [60].

2.2. Dataset Condensation

Dataset Condensation (DC) or dataset distillation synthesizes condensed datasets that retain the learning properties of larger originals, enabling efficient model training

with reduced data [58]. This technique has applications in continual learning [46, 50], privacy protection [4, 11], and neural architecture search [7], among others. Wang *et al.* [58] introduced DC, framing it as a meta-learning problem where network parameters are optimized as a function of synthetic data to minimize the training loss on real datasets. Building on this foundation, subsequent studies have leveraged surrogate objectives to address the unrolled optimization challenges inherent in meta-learning framework. Notably, gradient matching methods [13, 31, 37, 66, 68] align DNN gradients between original and condensed datasets, while trajectory matching approaches [6, 8, 12] align the DNNs’ parameter trajectories. Although promising, their reliance on computationally intensive bi-level optimization hinders their applicability to large-scale setups [15, 64].

To address these limitations, Zhao *et al.* [67] formulate DC as a distribution matching problem between the original and condensed datasets within the embeddings of randomly sampled DNNs. Specifically, DM [67] aligns the feature distributions of condensed and original datasets by matching their penultimate layer feature representations. However, DM [67] sacrifices the performance to maintain the efficiency. Thus, strategies such as IDM by Zhao *et al.* [69] and Datadam by Sajedi *et al.* [48] have been introduced to enhance DM. IDM [69] employs semi-trained models, and class-aware regularization, to improve DM performance. However, it diverge from efficient optimization based on randomly sampled DNNs. DataDAM [48], improves DM by using spatial supervision to align the attention maps between real and synthetic datasets. However, such restricted spatial supervision leads to the generalization reduction [10, 30]. Also, CAFE [57] aligns feature distributions of condensed and real datasets across multiple DNN layers using a dynamic bi-level optimization framework. However, it diverges from DM efficient framework, leading to considerable computation cost.

2.3. Style

Style of an image encompasses its visual attributes such as texture and color [17], which are widely represented by the characteristics of intermediate feature maps [26, 41, 42, 56, 72]. Assuming a Gaussian prior for features, the first and second moments, *i.e.*, the mean and variance, of DNN feature maps are well-established style indicators [26, 41, 42, 56, 72]. Furthermore, second order moment between feature activations, captured by the Gram matrix is another widely used style indicator [16, 18, 40]. DNNs show strong bias toward input style, leading to severe performance degradation when the style of the training data is not align with that of test [25, 29, 44, 70, 72]. Style features have been widely used in style transfer [16, 18, 40], domain adaptation and generalization [20, 72, 73], amongst others, showcasing the importance of training and testing

style alignment. However, previous DM-based studies have been overlook the significant role of style, leading to style gap between condensed and original data, as shown in Figure 1a. In this work, we aim to reduce this style gap by employing well-established style indicators in DM-framework.

3. Proposed Method

3.1. Notation

In this paper, we use lowercase letters (*e.g.*, x) to denote scalars, lowercase boldface (*e.g.*, \mathbf{x}) to denote vectors, uppercase letters (*e.g.*, X) to denote functions, uppercase boldface (*e.g.*, \mathbf{X}) to denote matrices and uppercase calligraphic symbols (*e.g.*, \mathcal{X}) to denote sets.

3.2. Preliminary

DC aims to learn a small condensed dataset $\mathcal{S} = \{(\tilde{\mathbf{x}}_1, y_1), \dots, (\tilde{\mathbf{x}}_{|\mathcal{S}|}, y_{|\mathcal{S}|})\}$ from a large-scale real dataset $\mathcal{T} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{|\mathcal{T}|}, y_{|\mathcal{T}|})\}$, such that $|\mathcal{S}| \ll |\mathcal{T}|$ and \mathcal{S} preserve essential information presented in \mathcal{T} [58]. Specifically, DC seeks to learn \mathcal{S} in a way that an arbitrary learning function trained on \mathcal{S} can have similar performance as that trained on \mathcal{T} [58, 67]:

$$\mathcal{S}^* = \arg \min_{\mathcal{S}} \mathbb{E}_{(\mathbf{x}, y) \sim P_D} [|L(\Phi_{\theta_{\mathcal{T}}}(\mathbf{x}), y) - L(\Phi_{\theta_{\mathcal{S}}}(\mathbf{x}), y)|], \quad (1)$$

where \mathbf{x} is a sample from real image distribution P_D , and y is its corresponding label. $\Phi_{\theta} : \mathbb{R}^q \rightarrow \mathbb{R}^d$ denotes a mapping, *i.e.*, DNN, with trainable parameter θ , that maps $\mathbf{x} \in \mathbb{R}^q$ to d -dimensional embedding space. For RGB input q is $3 \times h \times w$. Furthermore, $\theta_{\mathcal{T}}$, and $\theta_{\mathcal{S}}$ are two samples from the distribution of θ trained on \mathcal{T} , and \mathcal{S} , respectively. Finally, L represents the learning objective function, *e.g.*, empirical risk.

Intuitive approach to solving the optimization in Equation 1 is to use a bi-level learning regime by optimizing \mathcal{S} and θ in turn [6, 37, 57, 66, 68]. However, the nested loop of alternately optimizing θ and \mathcal{S} is computationally intensive and scales poorly to large datasets and complex architectures [6, 15, 37, 67, 68]. Inspired by the observation of Giryes *et al.* [21] that a Φ_{θ} with random θ performs a distance-preserving embedding, Zhao *et al.* [67] demonstrate that the validity of \mathcal{S} can also be guaranteed even with random θ . Specifically, [67] reformulates the DC objective as a distribution matching problem in the embedding of random DNNs. [67] enforces the alignment between feature distribution of \mathcal{S} with that of \mathcal{T} in the embedding of random mappings Φ_{θ} :

$$\mathcal{S}^* = \arg \min_{\mathcal{S}} \mathbb{E}_{\theta \sim \Theta} [D(\mathcal{S}, \mathcal{T}; \Phi_{\theta})], \quad (2)$$

where Θ is the distribution of θ , *i.e.*, the distribution used to initialize network parameters, and D is an arbitrary metric measuring the divergence between the two distributions.

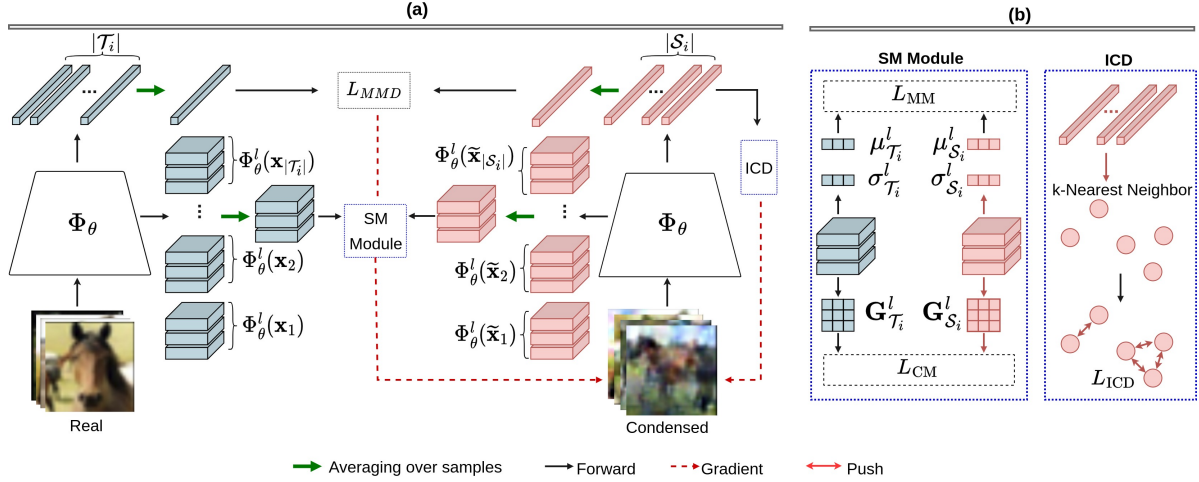


Figure 2. (a) Visualization of the proposed method, which includes a Style Matching (SM) module and Intra-Class Diversity (ICD) components. (b) SM module includes Moments Matching (MM) and Correlation Matching (CM) losses to reduce style discrepancies between real and condensed sets by using the *i.e.*, mean and variance of feature maps as well as correlation among feature maps captured by the Gram matrix in a DNN across different layers. Meanwhile, the ICD component enhances diversity within condensed sets by pushing each condensed sample away from its k nearest intra-class neighbors.

Equation 2 circumvents the nested loop of the bi-level optimization methods by solely updating \mathcal{S} , significantly reducing DC computational cost. This method is known as Distribution Matching (DM) for dataset condensation. Taking Maximum Mean Discrepancy (MMD) [22] as D , the objective function in DM [67] is defined as:

$$L_{MMD} = \mathbb{E}_{\theta \sim \Theta} \left[\sum_{i=0}^{c-1} \left\| \frac{1}{|\mathcal{S}_i|} \sum_{\tilde{\mathbf{x}} \in \mathcal{S}_i} \Phi_\theta(\tilde{\mathbf{x}}) - \frac{1}{|\mathcal{T}_i|} \sum_{\mathbf{x} \in \mathcal{T}_i} \Phi_\theta(\mathbf{x}) \right\|^2 \right], \quad (3)$$

where \mathcal{S}_i and \mathcal{T}_i are the subsets of condensed and real datasets, respectively, for the i -th class, and c is the number of classes.

Despite the efficiency of DM, this expedited learning comes at the cost of reduced performance, *e.g.*, an 8% performance reduction in CIFAR100 ($|\mathcal{T}| = 50000$) compared to DSA [66] (a bi-level optimization method) when condensing all images in a class into 10 instances ($|\mathcal{S}| = 1000$). To study this performance degradation, we decompose the feature distribution into its major factors, *i.e.* style and content components [16, 17, 26, 27, 39, 71, 72]. Style expresses attributes such as texture, color, and smoothness, while the content captures semantic information. Our exploratory experiments, depicted in Figure 1, reveal two major shortcomings in the synthetic dataset \mathcal{S} learned by DM compared to \mathcal{T} : (1) considerable style discrepancy and (2) limited diversity in content information. We address these issues in Sections 3.3 and 3.4, respectively. Our proposed approach is illustrated in Figure 2.

3.3. Style Matching

Experiments in Figure 1a, reveal the failure of DM [67] in capturing the style of the original dataset. Furthermore,

Figure 1c, and d illustrate the disruptive effect of this style gap on performance; in line with recent findings in deep learning community [25, 29, 44, 70, 72]. Hence, we aim to enforce the condensed data to represent the style of the original large dataset. Specifically, we introduce the Style Matching (SM) module to DM [67] framework, comprising two complementary sub-modules: (1) Moments Matching (MM), aligning the first and second moments of feature maps, and (2) Correlation Matching (CM), aligning the correlations among feature maps. We detail them in the following two sections.

3.3.1 Moments Matching

Inspired by the observation in Figure 1a, c, and d, here, we enforce the condensed dataset \mathcal{S} to capture the style of \mathcal{T} in addition to its content. To this end, we utilize the first and second moments, *i.e.*, mean and variance, of the intermediate feature maps to explicitly enforce \mathcal{S} to represent the style of the \mathcal{T} [39, 71, 72]. This is done by minimizing the mean-squared distance of these moments across original and condensed datasets, in the same way as used in the pioneering work of AdaIN [26]:

$$L_{MM} = \sum_{i=0}^{c-1} \frac{1}{2} \left(\sum_{l \in \mathcal{L}} \|\mu_{\mathcal{S}_i}^l - \mu_{\mathcal{T}_i}^l\|^2 + \sum_{l \in \mathcal{L}} \|\sigma_{\mathcal{S}_i}^l - \sigma_{\mathcal{T}_i}^l\|^2 \right), \quad (4)$$

$$\mu_{\mathcal{A}}^l = \frac{1}{|\mathcal{A}|} \sum_{\mathbf{a} \in \mathcal{A}} \mu^l(\mathbf{a}), \quad \sigma_{\mathcal{A}}^l = \frac{1}{|\mathcal{A}|} \sum_{\mathbf{a} \in \mathcal{A}} \sigma^l(\mathbf{a}); \quad \mathcal{A} \in \{\mathcal{S}_i, \mathcal{T}_i\},$$

where the channel-wise mean and variance of l -th layer are denoted by $\mu^l \in \mathbb{R}^{n_l}$ and $\sigma^l \in \mathbb{R}^{n_l}$, respectively. n_l represents the number of channels in the l -th layer of the network Φ_θ . Furthermore, the outer loop over the classes c

Dataset	IPC Ratio% Resolution			Coreset Selection			Training Set Synthesis						Whole Dataset
	IPC	Ratio%	Resolution	Random	Herding [45]	Forgetting [54]	DD [†] [58]	DG [68]	DSA [66]	DM [67]	CAFE [57]	Ours	
CIFAR10	1	0.02	32	14.4±2.0	21.5 ± 1.2	13.5 ± 1.2	-	28.3 ± 0.5	28.8 ± 0.7	26.0±0.8	31.6 ± 0.8	27.9 ± 0.7	84.8 ± 0.1
	10	0.2	32	26.0 ± 1.2	31.6 ± 0.7	23.3 ± 1.0	36.8 ± 1.2	44.9 ± 0.5	52.1 ± 0.5	48.9 ± 0.6	50.9 ± 0.5	53.0 ± 0.2	
	50	1	32	43.4 ± 1.0	40.4 ± 0.6	23.3 ± 1.1	-	53.9 ± 0.5	60.6 ± 0.5	63.0 ± 0.4	62.3 ± 0.4	65.6 ± 0.4	
CIFAR100	1	0.2	32	4.2 ± 0.3	8.3 ± 0.3	4.5 ± 0.2	-	12.8 ± 0.3	13.9 ± 0.3	11.4 ± 0.3	14.0 ± 0.3	13.5 ± 0.2	56.2 ± 0.3
	10	2	32	14.6 ± 0.5	17.3 ± 0.3	15.1 ± 0.3	-	25.2 ± 0.3	32.3 ± 0.3	29.7 ± 0.3	31.5 ± 0.2	33.9 ± 0.2	
	50	10	32	30.0 ± 0.4	33.7 ± 0.5	-	-	30.6 ± 0.6	42.8 ± 0.4	43.6 ± 0.4	42.9 ± 0.2	45.3 ± 0.3	
Tiny ImageNet	1	0.2	64	1.4 ± 0.1	2.8 ± 0.2	1.6 ± 0.1	-	5.3 ± 0.1	5.7 ± 0.1	3.9 ± 0.2	-	4.9 ± 0.1	37.6 ± 0.4
	10	2	64	5.0 ± 0.2	6.3 ± 0.2	5.1 ± 0.2	-	12.9 ± 0.1	16.3 ± 0.2	12.9 ± 0.4	-	17.2 ± 0.3	
	50	10	64	15.0 ± 0.4	16.7 ± 0.3	15.0 ± 0.3	-	12.7 ± 0.4	5.1 ± 0.2	25.3 ± 0.2	-	27.4±0.1	
ImageNet-1K	1	0.2	64	0.5±0.2	-	-	-	-	-	1.3±0.2	-	2.1±0.1	33.8±0.3
	10	2	64	2.9±0.4	-	-	-	-	-	5.5±0.4	-	7.5±1.2	
	50	10	64	7.1±1.5	-	-	-	-	-	11.4±1.2	-	15.6±0.8	

Table 1. The performance (testing accuracy %) comparison with state-of-the-art DC and coreset selection methods. We condense the given number of IPCs using the training set, train a DNN on the condensed set from scratch, and evaluate the network on the original testing data. Whole Dataset: the accuracy of the model trained on the whole original training set. Ratio (%): the ratio of condensed images to the whole training set. DD[†] uses AlexNet [34] for CIFAR10 dataset and all other methods use ConvNet for training and evaluation. Some entries are marked as absent due to unreported values or scalability issues of optimization-based methods.

	CIFAR10			CIFAR100			TinyImageNet		
	1	10	50	1	10	50	1	10	50
Resolution	32 × 32			32 × 32			64 × 64		
Random	10.3±0.8	25.7±0.5	36.8±1.2	2.5±0.5	9.5±0.9	21.2±0.8	0.5±0.6	4.2±0.5	6.5±0.8
DM [67]	19.1±1.9	32.6±0.9	44.9±0.7	4.1±0.2	13.5±0.4	28.3±0.2	1.6±0.2	6.1±0.2	11.5±0.9
Ours	22.3±0.7	40.9±0.6	51.6±0.5	6.3±0.3	21.4±0.4	34.0±0.2	2.0±0.2	8.6±0.4	15.1±0.3
Whole Dataset	93.07±0.1			75.61±0.3			41.45±0.4		

Table 2. The performance (testing accuracy %) comparison with DM [67] for CIFAR10, CIFAR100, and TinyImageNet datasets by employing ResNet-18 architecture for training and evaluation.

is to adapt the style matching loss function in [26] to the DC framework. The μ and σ of the feature maps at a specific layer capture the style information represented in every individual channel of that layer [16, 26]. Matching these statistics across original and condensed data reduces the gap between the style information among them without imposing rigorous spatial constraints that can reduce cross-architecture generalization [10, 30, 62]. We enforce first and second moments matching across multiple layers \mathcal{L} of the DNN to ensure comprehensive style matching [44].

3.3.2 Correlation Matching

Another well-established style indicator is the one introduced by Gatys *et al.* [16] consisting of correlations among feature maps [17, 18, 26]. Specifically, Gatys *et al.* [16] represent the style of the input image to a DNN by the correlation between i -th and j -th filters in layer l . This correlation is captured by the Gram matrix $\mathbf{G}^l \in \mathbb{R}^{n_l \times n_l}$, computed as:

$$\mathbf{G}^l = \Phi^l (\Phi^l)^\top \quad (5)$$

where $\Phi^l \in \mathbb{R}^{n_l \times (h_l \cdot w_l)}$ represents the feature maps from layer l , with n_l being the number of filters and $h_l \cdot w_l$ being the spatial dimensions of the feature maps.

We optimize the mean-squared distance between the entries of \mathbf{G} across the condensed and original datasets over a set of \mathcal{L} layers, providing stationary and multi-scale style

feature representations [18]. Formally, the proposed Correlation Matching (CM) loss, L_{CM} , is formulated as:

$$L_{CM} = \mathbb{E}_{\theta \sim \Theta} \left[\frac{1}{4(h_l w_l)^2 n_l^2} \sum_{i=0}^{c-1} \sum_{t \in \mathcal{L}} \left(\frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{G}^l(\tilde{\mathbf{x}}) - \frac{1}{|\mathcal{T}_i|} \sum_{\mathbf{x} \in \mathcal{T}_i} \mathbf{G}^l(\mathbf{x}) \right)^2 \right], \quad (6)$$

where $\frac{1}{4(h_l w_l)^2 n_l^2}$ is the normalization factor [18, 41]. By minimizing Equation 6, we enforce the condensed set to capture the style statistics unique to the real datasets in each class [16, 18, 65]. It is worth noting that Equation 4 captures style details within each feature map, ignoring their correlations. Equation 6 accounts for the correlations among feature maps, complementing Equation 4. Therefore, to include style information represented in each feature map and correlation among feature maps, we define the style matching loss function as:

$$L_S = \alpha L_{MM} + L_{CM}, \quad (7)$$

where α is a balancing factor between L_{MM} and L_{CM} . Note that L_{MM} and L_{CM} discard the spatial information, desired for cross-architecture generalization [10, 30].

3.4. Intra-Class Diversity

The MMD objective in DM, Equation 3, supports content matching between \mathcal{T} and \mathcal{S} [67]; however, the resulting \mathcal{S} suffers from limited intra-class diversity, as shown in Figure 1b. Specifically, synthesized \mathcal{S} contains similar samples within each class, *i.e.*, samples forming local clusters in the embedding space. It has been shown that the generalization error is bounded by the dataset diversity [30, 49, 52]. In other words, the more diverse the instances within a dataset, the more generalizable the model trained on that dataset will be.

To promote intra-class diversity, we design L_{ICD} as the Kullback–Leibler divergence among latent features of sam-

Algorithm 1 Decomposed Distribution Matching in Dataset Condensation

Input: \mathcal{T} : Real dataset, Φ_θ : DNN, Θ : distribution for initializing θ , $\lambda \geq 0$, $\alpha \geq 0$, $\beta \geq 0$, t : total training iterations
Output: Condensed dataset \mathcal{S}

- 1: Initialize \mathcal{S} with real images from \mathcal{T}
- 2: **for** $iter = 0 \dots t - 1$ **do**
- 3: Initialize Φ_θ with $\theta \sim \Theta$;
- 4: Sample \mathcal{S}_i from $\mathcal{S} \forall i \in \{0, \dots, c - 1\}$
- 5: Sample \mathcal{T}_i from $\mathcal{T} \forall i \in \{0, \dots, c - 1\}$
- 6: Compute $L_S = \alpha L_{MM} + L_{CM}$
- 7: Compute $L_C = \beta L_{ICD} + L_{MMD}$
- 8: Update the synthetic dataset $\mathcal{S} \leftarrow \mathcal{S} - \eta \nabla_{\mathcal{S}} (\lambda L_S + L_C)$
- 9: **end for**

ples in \mathcal{S}_i . To effectively penalize samples from forming local clusters, *i.e.*, representing similar information, and to preserve the correct class semantics while introducing new information beneficial for model training, we enforce k -nearest neighbors constraint on the diversity criterion. Therefore, the proposed diversity criterion is as follows:

$$L_{ICD} = \mathbb{E}_{\theta \sim \Theta} \left[- \sum_{i=0}^{c-1} \sum_{\tilde{\mathbf{x}} \in \mathcal{S}_i} KL(S(\Phi_\theta(\tilde{\mathbf{x}})) \| S(\overline{\mathbf{m}}_{\tilde{\mathbf{x}}})) \right], \quad (8)$$

$$\text{s.t. } \overline{\mathbf{m}}_{\tilde{\mathbf{x}}} = \frac{1}{k} \sum_{\tilde{\mathbf{x}} \in \mathcal{A}_i^k} \Phi(\tilde{\mathbf{x}}),$$

where $KL(a||b)$ denotes the Kullback–Leibler divergence between distributions a , and b , and $S(\cdot)$ is the Softmax function that transforms feature vectors into probability vectors, enabling the measurement of KL divergence between features [59]. $\overline{\mathbf{m}}_{\tilde{\mathbf{x}}}$ represents the mean feature over the set \mathcal{A}_i^k , and \mathcal{A}_i^k denotes k closest intra-class synthetic instances to $\tilde{\mathbf{x}}$:

$$\mathcal{A}_i^k = \{\tilde{\mathbf{x}}_j; \underset{\tilde{\mathbf{x}}_j \in \mathcal{S}_i, \tilde{\mathbf{x}}_j \neq \tilde{\mathbf{x}}}{\operatorname{argmin}_k} \|\Phi(\tilde{\mathbf{x}}_j) - \Phi(\tilde{\mathbf{x}})\|^2\}. \quad (9)$$

This optimization penalizes synthetic samples that cluster in the embedding space of Φ , resulting in more diverse intra-class instances and efficacy in capturing the distribution of original data.

Equation 3 focuses on the content matching between original and condensed data [67], but ignores the diversity among instances of \mathcal{S} . Thus, we propose to regularize Equation 3 with Equation 8 as the content matching loss:

$$L_C = \beta L_{ICD} + L_{MMD}. \quad (10)$$

Finally, we learn the synthetic dataset by solving the following optimization problem:

$$\mathcal{S}^* = \underset{\mathcal{S}}{\operatorname{argmin}} (\lambda L_S + L_C), \quad (11)$$

where λ balances the contributions of the style matching loss L_S in the overall optimization. A summary of the learning algorithm is provided in Algorithm 1.

	ImageWoof		ImageNette	
Img/Cls	1	10	1	10
Resolution	128 × 128		128 × 128	
Random	13.9±1.1	26.9±1.8	23.1±1.5	47.5±2.2
DM [67]	20.9±1.5	31.2±0.6	32.5±0.4	55.6±0.7
Ours	23.8±0.5	34.5±0.3	36.0±0.6	58.1±0.2
Whole Dataset	67.0±1.3		87.4±0.1	

Table 3. The performance (testing accuracy %) comparison with DM [67] for ImageWoof and ImageNette subsets of mageNet-1K, by employing ConvNet architecture for training and evaluation.

4. Experiments

4.1. Datasets

We conduct evaluation on CIFAR10, CIFAR100 [32] (32 × 32 pixels), and TinyImageNet along with ImageNet-1K [9] (resized to 64 × 64 pixels). Also, we evaluate our method on high-resolution (128×128 pixels) subsets of ImageNet-1K, *i.e.*, ImageNette and ImageWoof, containing instances from 10 classes [6].

4.2. Implementation details

We evaluate our method on $IPC \in \{1, 10, 50\}$ using ConvNet and ResNet-18 [24]. Experimental settings and DNN architectures are consistent with DM [67] unless specified. To handle input size of 64×64 and 128×128 pixels, we extend the ConvNet, which has three blocks, by adding a fourth and fifth convolutional block, respectively. We initialize \mathcal{S} with randomly selected images from \mathcal{T} and optimize using SGD optimizer with a fixed learning rate of 1.0. The differentiable augmentation strategy [66] is employed, as used in DM [67]. We train 20 DNNs from scratch on condensed sets with different initialization seeds, evaluating each on real test data. This process is repeated five times, resulting in five condensed datasets and 100 trained DNNs per IPC. We report the mean and variance of accuracy across these networks. The same DNN architecture is used for both training and evaluation unless specified. Hyperparameters α , β and λ are set to 1.0, 10.0 and 5×10^3 , respectively, determined empirically. Also, the number of nearest neighbors for Equation 8 is set to $0.2 * IPC$. Section B of Supplementary Material provides detailed ablation on hyperparameters.

4.3. Comparisons with State-of-the-art Methods

Here we compare our approach with DC baselines of DM [67], CAFE [57], DD [58], DG [68], and DSA [66], as well as coreset selection methods of Random [45], Herding [5, 45], and Forgetting [54], as shown in Table 1. Comparing the results of DC approaches with coreset selection methods highlights the superiority of DC over coreset selection. Our method consistently outperforms DM across

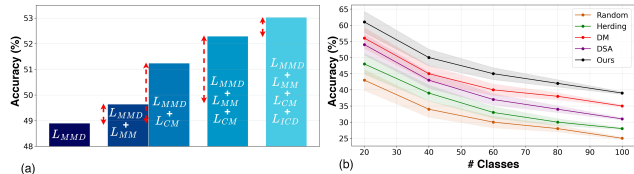


Figure 3. (a) Ablation on loss components on CIFAR10 with IPC=10 by employing ConvNet. b) Evaluation in continual learning for CIFAR100 in five steps, *i.e.*, 20 classes per step. Shaded regions show the performance tolerance.

datasets and IPCs. Particularly, with IPC=10, proposed approach surpasses DM [67] on CIFAR10, CIFAR100, Tiny-ImageNet and ImageNet-1K, with considerable margins of 4.1%, 4.2%, 4.3%, and 2.0%, respectively. Concretely, with IPC=50, our method outperforms DM by 2.6%, 1.7%, 2.1%, and 4.2% in CIFAR10, CIFAR100, TinyImageNet and ImageNet-1K, respectively. These consistent improvements across datasets of varying sizes and resolutions underlines that the proposed method is not confined to the dataset size and resolution.

Moreover, Table 1 demonstrates the consistent superiority of the proposed method over DM at IPC = 1. As intra-class diversity is not applicable to IPC=1, these improvements underscore the effectiveness of the proposed SM module and highlight the role of style alignment between \mathcal{T} and \mathcal{S} . With IPC = 10 our method improves CAFE [57] on CIFAR10, and CIFAR100, with noticeable margin of 2.1%, and 2.9%, respectively. Concretely, with IPC = 50 our proposal outperforms CAFE on CIFAR10, and CIFAR100 with considerable margin of 3.4%, and 2.6%, respectively. Please note that CAFE [57] does not report performance on large-scale datasets such as ImageNet-1K due to its heavy computation load. These improvements across datasets and IPCs showcase the importance of style alignment and intra-class diversity which results to outperforming CAFE with fewere computational burden.

The proposed method surpasses CAFE [57] in all IPCs except IPC=1. Concretely, improvements of our approach over DM are more pronounced at IPC > 1. These results highlight the importance and effectiveness of L_{ICD} , which is relevant only for IPC > 1. Table 2 showcases the efficacy of our method on ResNet-18, as a more sophisticated architecture than ConvNet. Particularly, our approach outperforms DM with IPC=10 by considerable margins of 8.3%, 7.9%, 2.5% in CIFAR10, CIFAR100 and TinyImageNet, respectively. In IPC=50, our method improves DM by 6.6%, 5.7%, and 3.6% in CIFAR10, CIFAR100 and TinyImageNet, respectively. These improvements highlights that our method is not limited to simple architectures like ConvNet. Again, improvements are more pronounced at IPC > 1, emphasizing the importance of L_{ICD} . Comparing Tables 1 and 2, changing ConvNet to ResNet-18, *i.e.*, more sophisticated architecture, results in more consider-

Method	Train Model	Test Model			
		ConvNet	AlexNet	VGG-11	ResNet-18
DSA [66]	ConvNet	51.9±0.4	34.3±1.6	42.3±0.91	41.0±0.4
DM [67]	ConvNet	48.6±0.63	38.3±1.2	40.8±0.4	39.2±1.2
CAFE [57]	ConvNet	50.9 ± 0.5	41.1±0.8	41.9±0.1	40.1±0.2
Ours	ConvNet	53.0±0.3	48.7±0.8	46.2±0.8	42.6±0.8
Ours	AlexNet	36.4±0.9	32.8±1.34	32.5±1.0	33.9±0.9
	VGG-11	41.2±0.4	37.4±0.3	41.7±0.4	38.8±0.8
	ResNet-18	41.9±0.5	34.7±1.9	36.65±1.0	40.93±0.6

Table 4. Cross-architecture (testing accuracy %) performance of our proposed method compared to DM [67], DSA [66] and CAFE [57] methods for CIFAR10 with IPC=10.

able improvements over DM. This suggests better generalization and practicality of the proposed method since it scales well with increased network complexity.

Table 3 compares our method with DM on datasets with higher resolution, *i.e.*, 128×128 pixels, using ConvNet as the backbone. Concretely, our method outperforms DM across datasets and IPCs, showcasing that it is not restricted to low- and medium-resolution datasets. Specifically, the proposed approach improves upon DM by at least 2.9%, and 2.5% in ImageWoof and ImageNette, respectively. Consistent improvements in Tables 1, 2, and 3 showcase that the proposed method is not confined to a specific resolution, dataset scale, or DNN architecture.

4.4. Cross-architecture Evaluations

Here we assess the cross-architecture transferability of our method by learning a condensed dataset with one architecture and evaluating it on different architectures. To this end, we used ConvNet, AlexNet [33], VGG-11 [53], and ResNet-18 [24] architectures, as shown in Table 4. Using ConvNet for condensing dataset, our approach consistently outperforms its competitors across all evaluation architectures, underscoring its transferability across diverse DNN architectures. Specifically, our method outperforms DM by 10.3%, 5.4% and 3.4% when testing with AlexNet, VGG-11 and ResNet-18, respectively. Also, the performance of employing ConvNet for learning condensed set surpasses more complex architectures. This result is in line with the observation of DM [67] that more complex architecture results in convergence issues and noisy features.

4.5. Orthogonality to DM-based Methods

Our proposal aims to improve distribution matching in DC by decomposing the dataset distribution into style, and content. Previous studies based on DM [67] overlook the significance of style alignment between original and condensed data [48, 69]. Here, we evaluate the effectiveness of our proposed SM module on two DM-based methods, as shown in Table 5. Specifically, we modified the training code of DataDM [48] and IDM [69] to include our style alignment loss function, resulting in improved performance. The results indicate that our method is orthogonal to ex-

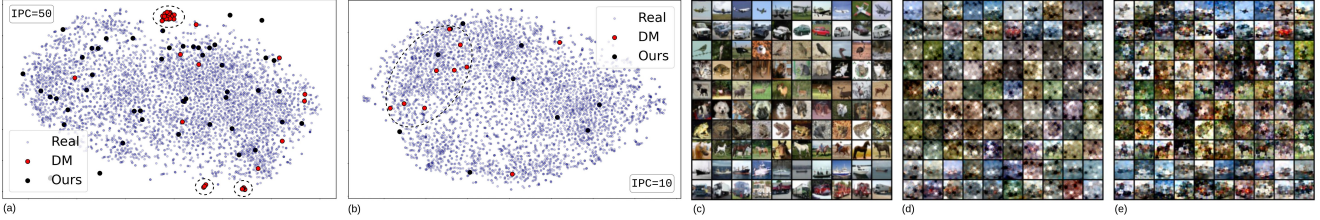


Figure 4. (a, b) Intra-class diversity of two randomly selected classes of CIFAR10 with (a) IPC=50 and (b) IPC=10. Our method enhances diversity across both IPCs, addressing the limited intra-class diversity issue in DM. (c, d, e) Visualizations samples from (c) original and (d) condensed DM [67] and (e) our method for CIFAR10 with IPC=10. Both methods are initialized from real samples. The proposed method improves visual quality and diversity.

Img/Cls	CIFAR10			CIFAR100			TinyImageNet		
	1	10	50	1	10	50	1	10	50
Resolution	32 × 32			32 × 32			64 × 64		
DM [67]	26.0±0.8	48.9±0.6	63.0±0.4	11.4±0.3	29.7±0.3	43.6±0.4	3.9±0.2	12.9±0.4	25.3±0.2
DM [67]+SM	27.8±0.7	52.3±0.4	64.1±0.7	13.5±0.2	33.0±0.1	45.1±0.3	4.91±0.1	16.1±0.2	25.2±0.3
DataDAM [48]	32.0±1.2	54.2±0.8	67.0±0.4	14.5±0.5	34.8±0.5	49.4±0.3	8.3±0.4	18.7±0.3	28.7±0.3
DataDAM [48]+SM	33.2±0.9	56.4±0.6	68.5±0.4	15.6±0.7	35.8±0.6	50.2±0.2	9.6±0.5	20.1±0.4	29.8±0.2
IDM [69]	45.6±0.7	58.6±0.1	67.5±0.1	20.1±0.3	45.1±0.1	50.0±0.2	10.1±0.2	21.9±0.2	27.7±0.3
IDM [69]+SM	46.8±0.4	60.2±0.2	68.8±0.3	21.6±0.4	47.2±0.3	51.3±0.4	11.6±0.4	23.8±0.5	28.9±0.2
Whole Dataset	84.8±0.1			56.2±0.3			37.6±0.4		

Table 5. Performance (testing accuracy %) comparison after integrating our proposed Style Matching (SM) loss with baseline DM, and two recent DM-based methods, DataDam [48] and IDM [69]. Results are for CIFAR-10, CIFAR-100, and TinyImageNet datasets using the ConvNet architecture.

isting DM-based methods, highlighting the importance of style alignment, consistent with the well-established DNN bias toward style information [25, 29, 44, 70, 72].

4.6. Ablation on Loss Components

Here, we assess the contribution of each loss component to the overall performance of our method on CIFAR10 with IPC=10. Results in Figure 3a reveal that both style matching supervisions, L_{MM} , and L_{CM} , improve upon the baseline L_{MMD} , highlighting the importance of style matching between original and condensed datasets. Comparing (DM+MM) and (DM+CM) against (DM+MM+CM) validates the complementary nature of style information captured by mean and variance of feature maps (L_{MM}) and the correlation among feature maps (L_{CM}). Furthermore, incorporation of L_{ICD} leads to an additional improvement, emphasizing the significance of intra-class diversity to effectively capture the real dataset distribution.

Furthermore, Figure 4a, and b show the t-SNE visualizations of the feature distribution for two categories with IPC=10 and IPC=50. L_{ICD} effectively addresses limited diversity of DM. This advantage is consistent across IPCs, demonstrating the generalizability of the proposed L_{ICD} . In addition, Figure 4c, d, and e display 10 samples per class from the real CIFAR10 dataset, and the condensed sets learned by DM and our method. Our method improves visual quality and diversity relative to DM, highlighting the efficacy of the SM module (Section 3.3.2) and ICD (Section 3.4) components, respectively, in reducing the style gap and

improving the intra-class diversity. We provided additional visualizations for CIFAR100 and TinyImageNet in Section C of Supplementary Materials. Also, please refer to Section D and E for ablation on style and the impact of the SM module across different blocks of ConvNet, respectively.

5. Applications: Continual Learning

One primary motivation of DC is to mitigate catastrophic forgetting in Continual Learning (CL) [45], making CL a reliable metric for evaluating condensation methods. To evaluate our proposal on CL, we store samples from a data stream within a predefined memory budget in a class-balanced manner. After each memory update, the model is retrained from scratch using the latest memory, which is replaced by the condensed set while adhering to memory budget and class balance constraints. Figure 3b shows our results against the Random [45], Herding [5, 45], DSA [66], and DM. To ensure reliability and omit the effect of class order, these experiments are repeated five times with different class orders. Our method outperforms its competitors with final test accuracy of 39.9%, compared to 24.8%, 28.1%, 31.7%, and 34.4% for Random, Herding, DSA and DM, respectively. As the number of classes increases, the performance gap between the proposed method and the DM baseline is more evident emphasising the scalability of our method into large-scale setup.

6. Conclusion

In this paper, we decomposed the distribution matching in DC into style and content matching. Specifically, we alleviate two shortcomings of (1) style discrepancy between original and condensed datasets, and (2) limited intra-class diversity in the condensed set, in current DC methods based on distribution matching. Our proposed style matching module reduces style disparity between real and condensed datasets by utilizing the first and second moments of DNN feature maps. We introduce a criterion based on KL-divergence to promote intra-class variability within the condensed dataset. The efficacy of the proposed method is demonstrated through extensive experiments on datasets of varying sizes and resolutions, across diverse architectures, and in the application of continual learning.

References

- [1] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019. 2
- [2] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018. 2
- [3] Eden Belouadah and Adrian Popescu. Scail: Classifier weights scaling for class incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1266–1275, 2020. 2
- [4] Nicholas Carlini, Vitaly Feldman, and Milad Nasr. No free lunch in” privacy for free: How does dataset condensation help privacy”. *arXiv preprint arXiv:2209.14987*, 2022. 3
- [5] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision*, pages 233–248, 2018. 2, 6, 8
- [6] 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*, pages 4750–4759, 2022. 1, 3, 6
- [7] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. *Advances in Neural Information Processing Systems*, 35:810–822, 2022. 3
- [8] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 3
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [10] Wenxiao Deng, Wenbin Li, Tianyu Ding, Lei Wang, Hongguang Zhang, Kuihua Huang, Jing Huo, and Yang Gao. Exploiting inter-sample and inter-feature relations in dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17057–17066, 2024. 2, 3, 5
- [11] Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *International Conference on Machine Learning*, pages 5378–5396. PMLR, 2022. 3
- [12] Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3758, 2023. 1, 3
- [13] Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [14] Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020. 2
- [15] Yunzhen Feng, Shanmukha Ramakrishna Vedantam, and Julia Kempe. Embarrassingly simple dataset distillation. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 3
- [16] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015. 2, 3, 4, 5
- [17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2, 3, 4, 5
- [18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3, 5
- [19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2
- [20] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016. 3
- [21] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, 64(13):3444–3457, 2016. 3
- [22] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 2, 4
- [23] Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coresets selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195. Springer, 2022. 2
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7
- [25] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020. 2, 3, 4, 8
- [26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 3, 4, 5
- [27] Mahesh Joshi, Mark Dredze, William Cohen, and Carolyn Rose. Multi-domain learning: when do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1302–1312, 2012. 4
- [28] Nikolai Kalischek, Jan D Wegner, and Konrad Schindler. In the light of feature distributions: moment matching for

- neural style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9382–9391, 2021. 2
- [29] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140, 2022. 2, 3, 4, 8
- [30] Samir Khaki, Ahmad Sajedi, Kai Wang, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Atom: Attention mixer for efficient dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7692–7702, 2024. 2, 3, 5
- [31] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoon Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. *arXiv preprint arXiv:2205.14959*, 2022. 3
- [32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 7
- [34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5
- [35] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 2
- [36] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4544–4553, 2020. 2
- [37] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoon Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. *arXiv preprint arXiv:2202.02916*, 2022. 1, 3
- [38] Pan Li, Lei Zhao, Duanqing Xu, and Dongming Lu. Optimal transport of deep feature for image style transfer. In *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*, pages 167–171, 2019. 2
- [39] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019. 2, 4
- [40] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 3
- [41] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017. 2, 3, 5
- [42] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 2, 3
- [43] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 3074–3082, 2015. 2
- [44] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 2, 3, 4, 5, 8
- [45] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 1, 2, 5, 6, 8
- [46] Andrea Rosasco, Antonio Carta, Andrea Cossu, Vincenzo Lomonaco, and Davide Bacciu. Distilled replay: Overcoming forgetting through synthetic samples. In *International Workshop on Continual Semi-Supervised Learning*, pages 104–117. Springer, 2021. 3
- [47] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 2
- [48] 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*, pages 17097–17107, 2023. 2, 3, 7, 8
- [49] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011. 5
- [50] Mattia Sangermano, Antonio Carta, Andrea Cossu, and Davide Bacciu. Sample condensation in online continual learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2022. 3
- [51] Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On random weights and unsupervised feature learning. In *icml*, 2011. 1
- [52] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 5
- [53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7
- [54] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *ICLR*, 2019. 2, 5, 6
- [55] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2
- [56] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3
- [57] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and

- Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2022. [1](#), [3](#), [5](#), [6](#), [7](#)
- [58] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. [1](#), [3](#), [5](#), [6](#)
- [59] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016. [6](#)
- [60] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *arXiv preprint arXiv:2301.07014*, 2023. [2](#)
- [61] David Junhao Zhang, Heng Wang, Chuhui Xue, Rui Yan, Wenqing Zhang, Song Bai, and Mike Zheng Shou. Dataset condensation via generative model. *arXiv preprint arXiv:2309.07698*, 2023. [2](#)
- [62] 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*, volume 38, pages 9314–9322, 2024. [2](#), [5](#)
- [63] Junyi Zhang, Ziliang Chen, Junying Huang, Liang Lin, and Dongyu Zhang. Few-shot structured domain adaptation for virtual-to-real scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. [2](#)
- [64] Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Dongkuan Xu. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11950–11959, 2023. [3](#)
- [65] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8035–8045, 2022. [2](#), [5](#)
- [66] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021. [1](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [67] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [68] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. 2021. [1](#), [3](#), [5](#), [6](#)
- [69] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023. [2](#), [3](#), [7](#), [8](#)
- [70] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European Conference on Computer Vision*, pages 535–552. Springer Nature Switzerland Cham, 2022. [2](#), [3](#), [4](#), [8](#)
- [71] Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. Semi-supervised domain generalization with stochastic stylematch. *International Journal of Computer Vision*, pages 1–11, 2023. [2](#), [4](#)
- [72] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. [2](#), [3](#), [4](#), [8](#)
- [73] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3):822–836, 2024. [3](#)