OPTICAL: Leveraging Optimal Transport for Contribution Allocation in Dataset Distillation

Xiao Cui[1,2,3] Yulei Qin[4] Wengang Zhou[1] Hongsheng Li[2,3] Houqiang Li[1]
[1] University of Science and Technology of China [2] CUHK MMLab
[3] CPII under InnoHK [4] Independent Researcher
cuixiao2001@mail.ustc.edu.cn, yuleichin@126.com
{zhwg,lihq}@ustc.edu.cn, hsli@ee.cuhk.edu.hk

Abstract

The demands for increasingly large-scale datasets pose substantial storage and computation challenges to building deep learning models. Dataset distillation methods, especially those via sample generation techniques, rise in response to condensing large original datasets into small synthetic ones while preserving critical information. Existing subset synthesis methods simply minimize the homogeneous distance where uniform contributions from all real instances are allocated to shaping each synthetic sample. We demonstrate that such equal allocation fails to consider the instance-level relationship between each real-synthetic pair and gives rise to insufficient modeling of geometric structural nuances between the distilled and original sets. In this paper, we propose a novel framework named OPTICAL to reformulate the homogeneous distance minimization into a bi-level optimization problem via matching-and-approximating. In the matching step, we leverage optimal transport matrix to dynamically allocate contributions from real instances. Subsequently, we polish the generated samples in accordance with the established allocation scheme for approximating the real ones. Such a strategy better measures intricate geometric characteristics and handles intra-class variations for high fidelity of data distillation. Extensive experiments across seven datasets and three model architectures demonstrate our method's versatility and effectiveness. Its plug-and-play characteristic makes it compatible with a wide range of distillation frameworks.

1. Introduction

Deep learning has achieved remarkable success across computer vision tasks, largely due to the availability of extensive annotated datasets with millions of examples [5, 9, 35, 52].

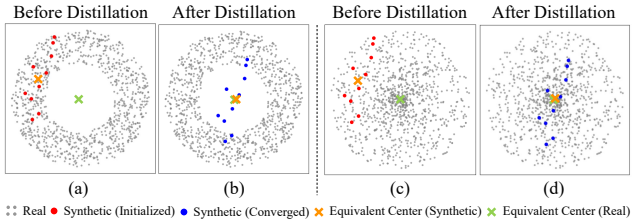

Figure 1. Pitfalls of the existing dataset distillation methods due to homogeneous distance minimization. We visualize two kinds of real data (gray) with the same equivalent center (green) but of distinct geometric structures. The initialized synthetic data (red) all shift towards the same equivalent center (orange) upon convergence (blue). Since contributions from real data are uniformly allocated to synthetic ones, two synthetic sets converge nearly identically, regardless of the structures of the underlying real sets.

However, such reliance on massive datasets brings significant challenges, particularly in storage and computational expenses [29, 46, 48]. To alleviate these burdens, dataset distillation has emerged as a promising solution by condensing large datasets into much smaller versions that still retain essential information [2, 12, 26, 37, 42]. The distilled dataset enables efficient storage and training, making the advanced models more accessible and cost-effective.

Dataset distillation approaches typically fall into two categories: 1) subset selection from real samples and 2) subset synthesis for generated samples. The subset selection methods [16, 38, 39, 45] leverage coreset-based techniques to heuristically select representative real samples from the original dataset for a compact subset. Despite its simplicity, the selected subset fails to capture the characteristics of the entire dataset by nature, resulting in suboptimal compression. Recent studies opt to generate synthetic data for lossless distillation. All these methods measure the difference and divergence between the synthetic and the real datasets for minimization either in terms of gradients or representations. For the former of gradient, optimization-oriented

Corresponding authors: Wengang Zhou and Hongsheng Li.

methods [3, 19, 27, 55, 57] develop bi-level frameworks to align the performance or parameters of models respectively trained on the generated and the original data. Despite their effectiveness, such alternative nested optimization of models and data results in high computation costs until convergence. For the latter of representations, distribution-matching (DM)-based methods [41, 50, 51, 56, 58] directly approximate the full dataset in the embedding space while downplaying the model-level matching.

However, most existing subset synthesis methods merely focus on matching the overall statistics (e.g., the equivalent center of all instances) while neglecting the pairwise relationship between instances from the synthetic and real sets for precise establishment of correspondence. Such limitation arises from the uniform allocation strategy where each real instance contributes equally in shaping the generated ones. Equal weights are assigned to real-synthetic pairs and therefore homogeneous divergence is minimized at the instance-level. Our demonstrative experiments, as shown in Fig. 1, illustrate that the distilled dataset obtained by minimizing the homogeneous distance with uniform contribution (as seen in (b) and (d) of Fig. 1) suffers from capturing the geometric structure of the original full dataset. Although the distilled instances share the equivalent center with the original ones, they converge regardless of the underlying distributions. The homogeneous minimization encourages a shortcut where the averaged representations are employed as a proxy for approximation, failing to take the intra-class variety into consideration.

In this paper, we introduce the **OPTI**mal transport for **C**ontribution **AL**location (**OPTICAL**), a plug-and-play approach grounded in Optimal Transport (OT) that measures pairwise relationship and adaptively adjusts the influence of real instances on each synthetic one, ensuring a fine-grained, context-aware contribution allocation scheme. We reformulate the homogeneous distance minimization problem into a bi-level optimization problem via Matching-and-Approximating. In the matching step, our OPTICAL uses the OT matrix to serve as a cost-efficient mapping tool for divergence measurement. It matches the most relevant real-synthetic pairs by determining an optimal plan to transform the distribution of the synthetic set into that of the real set. In the subsequent approximating step, we follow the contribution allocation scheme defined by OT and perform distance minimization either with optimization-oriented or with DM-based methods. Such mechanism of contribution allocation allows each synthetic sample to learn distinct, customized knowledge from matched real ones. It guarantees that the instance-level correspondence is considered to handle the intra-class variety for a high level of geometric resemblance between the distilled and the original sets.

To further refine the representation capacity for perception of distribution nuances, we project the gradients and representations into a Reproducing kernel Hilbert space that offers a more expressive cost matrix for optimal transport. This transformation leverages kernel evaluations as distances to better capture high-order distribution properties. Given the computation overhead of directly solving the OT matrix, we propose to employ Sinkhorn normalization as an efficient-yet-accurate approximation towards the regularized OT matrix. To sum up, our contributions are:
(1) We offer an in-depth mathematically rigorous analysis on the limitations behind homogeneous distance minimization for dataset distillation. The equal allocation of contributions from real instances is the fundamental problem.
(2) We propose the OPTICAL to reformulate the suboptimal homogeneous distance minimization into a bi-level optimization problem via matching-and-approximating. A novel contribution allocation mechanism is developed to first match real-synthetic pairs via OT and then adjust real instances in shaping synthetic ones during distillation.
(3) OPTICAL is a plug-and-play method that can be generally applied to most of the dataset distillation methods, without introducing significant computation overhead.

We validate the versatility and effectiveness of our OPTICAL through extensive experiments across seven datasets. Notably, our cross-architecture distillation exhibit even greater gains when the network used for evaluation differs from the proxy network used for matching, underscoring the robustness and generalizability of our approach.

2. Related Works

2.1. Dataset Distillation

Dataset distillation, also known as (a.k.a.) dataset condensation, creates a much smaller dataset that delivers similar effect of model training with the full set. Existing approaches can be mainly categorized as follows.

Real Subset Selection. Coreset selection techniques are often leveraged to choose a subset from the full dataset as the distilled dataset. Herding [45] prioritize samples close to class centers. Forgetting [39] prefers samples that exhibit forgetting characteristics as informative ones. K-center methods [16, 38] aim to minimize the maximum distance between the selected samples and their nearest class centers. The effectiveness of these methods depends heavily on the quality of the original dataset and the rationality of the heuristically designed criteria. Consequently, robustness is not promised across domains and tasks.

In this paper, we focus on the synthesis-based dataset distillation where the smaller dataset is newly generated, which permits higher flexibility and generalization potentials than the data selection methods for real subsets.

Synthetic Set Generation. Distance between the real and the generated sets are measured on either gradients or representations, leading to two different types of methods.

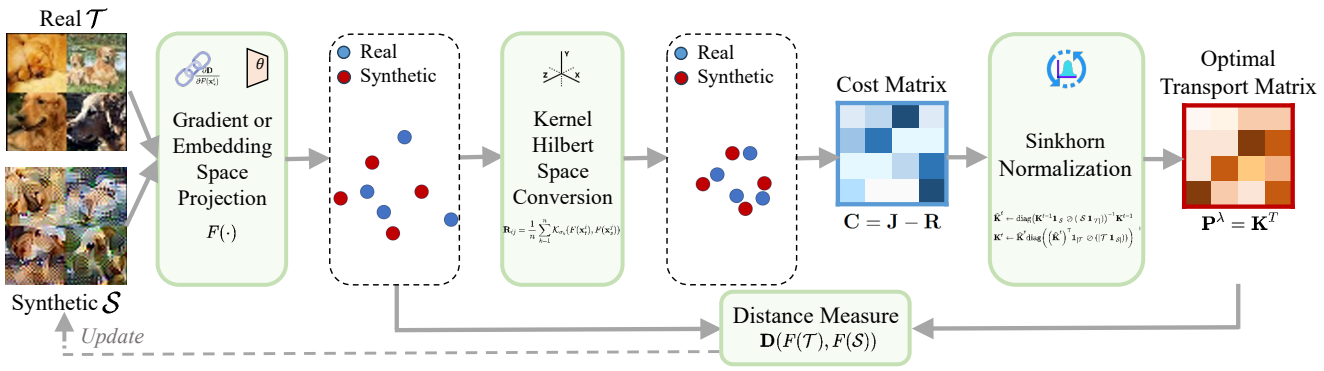

Figure 2. Illustration of the proposed pipeline. We first map the matching objective into the Hilbert space to compute the cost matrix, and then apply the Sinkhorn normalization to derive the optimal transport matrix. Such an optimal transport matrix is subsequently used to re-allocate contributions of the real instances for polishing the synthetic ones in distance measurement and minimization.

Optimization-oriented methods [3, 19, 42, 50, 57] match gradients through a bi-level framework. Models are updated in the outer loop and synthetic instances are refined in the inner loop. DM-based methods [13, 41, 50, 51, 56, 58], on the other hand, focus on feature alignment which bypasses alternative optimization by directly matching the distributions of real and synthetic instances. Despite the broad applicability of these synthetic methods, they all suffer from the homogeneous distance minimization problem where each real instance contributes equally to deliver knowledge to the synthetic ones. Such an equal allocation scheme neglects the fact that instances within the same class can vary greatly (e.g., fine-grained patterns) and the relationship (e.g., resemblance) between real and synthetic instances should be considered. Its insufficient perception of instance-wise distinctive features and class-wise geometric structures leads to inadequate penalty on feature alignment.

This paper targets at the homogeneous distance minimization problem that exists in both optimization-oriented and DM-based methods. We propose a bi-level optimization solution of matching-and-approximating, which adaptively allocates contributions from real to synthetic instances via efficient OT process. It is compatible with any synthesis distillation method in a plug-and-play manner.

2.2. Optimal Transport

The OT theory provides a mathematical framework for comparing probability distributions by calculating the minimal cost required to transform one distribution into the other. It excels at capturing the geometric structure of distributions [40, 53]. Previous methods primarily utilize the OT distance (a.k.a., Wasserstein distance) for measuring distribution differences. Such a metric has been employed across various domains including causal discovery [43, 44], image generation [1, 17, 34], unsupervised learning [4, 18, 23], and reinforcement learning [14, 20, 22, 54]. Due to the computationally intensity of the Wasserstein distance in most scenarios, the Sinkhorn distance has been proposed as a tractable approximation, which introduces an entropy regularization term to the OT problem [10]. Its success in quantifying divergences between any two distributions allows applications ranging from from domain adaptation [33, 49] to classification [28] and distillation [6–8].

In contrast with methods that directly use the OT distance as the minimization objective, we leverage the OT matrix for contribution allocation under our bi-level pipeline. This allows us to simultaneously retain the advantages of OT in divergence measurement and enjoy a high level of compatibility with existing dataset distillation frameworks.

3. Method

3.1. Problem Statement

Dataset distillation aims to create a small synthetic dataset $\mathcal{S} = \{(\boldsymbol{x}_s^i, y_s^i)\}_{i=1}^{|\mathcal{S}|}$ from a large-scale full dataset $\mathcal{T} = \{(\boldsymbol{x}_t^i, y_t^i)\}_{i=1}^{|\mathcal{T}|}, |\mathcal{S}| \ll |\mathcal{T}|$, such that neural models trained on $\mathcal{S}$ perform comparably to those trained on $\mathcal{T}$. The synthetic $\mathcal{S}$ is initialized with randomly chosen real instances from $\mathcal{T}$, and then iteratively refined to simulate $\mathcal{T}$.

3.2. Sub-optimal Dataset Distillation via Homogeneous Distance Minimization

One fundamental limitation behind synthesis-based distillation lies in the homogeneous distance minimization, where equal contributions from real data are devoted to polishing all synthetic samples. Such instance-irrelevant uniform allocation disregards the varying importance and learning necessity of real instances. Without instance-level constraints on matching pairs, the synthetic data merely capture the overall distribution nuances using the averaged representations as proxies. It performs a uniform shift from the equiv-

Table 1. Representative synthesis-based dataset distillation methods with their mapping functions F. The ϕ_θ denotes the encoder of the model f_θ parameterized by θ. The $\mathcal{L}(\cdot)$ refers to the training loss, and the $\mathcal{K}$ is a kernel function. The ϕ_{expert} represents an expert encoder and β denotes a combination coefficient.

Method	$F(\mathcal{S})$
Optimization-oriented Methods	
DC [57], DREAM [27]	$\nabla_\theta \mathcal{L}\left(f_\theta(\mathcal{S})\right)$
DM-based Methods	
DM [56], IDM [58]	$\phi_\theta(\mathcal{S})$
M3D [51]	$\mathcal{K}(\phi_\theta(\mathcal{S}), \cdot)$
DANCE [50]	$\beta\phi_\theta(\mathcal{S}) + (1-\beta)\phi_{\text{expert}}(\mathcal{S})$

alent center of initialized synthetic samples towards that of real ones, failing to harness the rich geometric structures.

Preliminaries. The common objective of dataset distillation is to minimize the distance between the real $\mathcal{T}$ and the synthetic datasets $\mathcal{S}$ via the mapping function $F(\cdot)$:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbf{D}\left(F(\mathcal{S}), F(\mathcal{T})\right), \tag{1}$$

where $\mathbf{D}$ represents the metric used to measure the distributional distance between $\mathcal{T}$ and $\mathcal{S}$. The function $F(\cdot)$ can take various forms, and we can unify both the optimization-oriented methods and the DM-based methods into Eq. (1) by different $F(\cdot)$ (see Table 1). For optimization of $\mathbf{D}$, most existing approaches decompose the overall distribution measurement into two parts, with one part for measuring the internal structure of $\mathcal{S}$ as $\mathbf{D}_s(\cdot)$ and the other for quantifying the interaction between $\mathcal{T}$ and $\mathcal{S}$:

$$\mathbf{D}\left(F(\mathcal{T}), F(\mathcal{S})\right) = \mathbf{D}_s\left(F(\mathcal{S})\right) - \alpha \sum_{i,j} F(\boldsymbol{x}_t^i) F(\boldsymbol{x}_s^j), \tag{2}$$

where α is a normalizing factor. We disregard terms unrelated to $\mathcal{S}$ as they do not contribute to the overall gradients.

Pitfalls. The formulation of Eq. (2) implies that each synthetic point $\boldsymbol{x}_s^j$ is equally affected by all real samples $\boldsymbol{x}_t^i$, which has two noteworthy problems: 1) The lack of importance discrimination among real samples leads to inefficient convergence as those difficult real ones of high intra-class diversity are assigned the same weights with easy ones; 2) The relevance between current synthetic samples and real ones, namely the necessity of mimicking each real sample, is not considered for dynamic adjustment. As a result, there exists a tendency of taking a short-cut path for the synthetic samples to simply approximate the averaged real instances in the embedding space. In this case, the entire optimization of dataset distillation can be viewed as shifting the equivalent class centers of the synthetic set towards those of the real set, neglecting the pairwise instance matching based on individual resemblance. To make it clearer, we present the most commonly used $\mathbf{D}$ [37, 56] for explanations:

$$\mathbf{D}\left(F(\mathcal{T}), F(\mathcal{S})\right) = \| \sum_{j\in\mathcal{S}} \frac{F(\boldsymbol{x}_s^j)}{|\mathcal{S}|} - \sum_{j\in\mathcal{T}} \frac{F(\boldsymbol{x}_t^j)}{|\mathcal{T}|} \|^2. \tag{3}$$

The Eq. (3) encourages the $\mathcal{S}$ to be identical to $\mathcal{T}$ on average. It exerts no explicit constraints on approximation between paired instances. From the perspective of gradients, each synthetic sample receives supervisions as:

$$\frac{\partial \mathbf{D}}{\partial F(\boldsymbol{x}_s^i)} = \frac{2}{|\mathcal{S}|^2} \sum_{j\in\mathcal{S}} F(\boldsymbol{x}_s^j) - \frac{2}{|\mathcal{S}||\mathcal{T}|} \sum_{j\in\mathcal{T}} F(\boldsymbol{x}_t^j). \tag{4}$$

The gradients are identical for polishing each $\boldsymbol{x}_s^i$ in $\mathcal{S}$. Such neglect of instance-wise customization and adjustment triggers off sub-optimal distillation fundamentally.

3.3. Distance Minimization with Dynamic Allocation of Real to Synthetic Instances

To address the drawbacks of homogeneous distance minimization, we propose the OPTICAL (see Fig. 2) to dynamically adjust the contributions from real instances to supervise synthesis. We leverage the OT theory to form a non-uniform contribution matrix in consideration of the relevance between real and synthetic pairs.

Reformulation of the Distance Minimization as the Bi-level Optimization via Matching-and-Approximating. We formulate the distance minimization as a bi-level optimization problem. An instance-wise scheme of contribution assignment is first estimated in the first stage, guiding synthetic samples to learn from real ones in the second stage:

$$\mathbf{D}\left(F(\mathcal{T}), F(\mathcal{S})\right) = \mathbf{D}_s\left(F(\mathcal{S})\right) - \alpha \sum_{i,j} \mathbf{P}_{ij} F(\boldsymbol{x}_t^i) F(\boldsymbol{x}_s^j), \tag{5}$$

where $\mathbf{P}_{ij}$ represents the contribution matrix, indicating the level of contribution from each real instance $F(\boldsymbol{x}_t^i)$ to shaping each synthetic instance $F(\boldsymbol{x}_s^j)$. Such a reformulation introduces a weighted contribution mechanism, enabling dynamic adjustment of emphasis in matching the most promising real-synthetic pairs for approximation. In accordance with the sizes of the original and the distilled dataset, the matrix $\mathbf{P}$ is subject to the marginal constraints on quantity:

$$\sum_j \mathbf{P}_{ij} = |\mathcal{S}|, \quad \sum_i \mathbf{P}_{ij} = |\mathcal{T}|. \tag{6}$$

It ensures that the total contributions from real to synthetic data are optimized via the strategy of assignment, where the influence of each real instance gets distributed across the synthetic set in a pairwise manner. We demonstrate that such a reformulation still guarantees the alignment of mean values upon convergence, where the average representation

of all synthetic instances approximates that of real ones:

$$\sum_i \frac{\partial \mathbf{D}}{\partial F(\boldsymbol{x}_s^i)} = 0 \iff \sum_{j\in\mathcal{S}} \frac{F(\boldsymbol{x}_s^j)}{|\mathcal{S}|} = \sum_{j\in\mathcal{T}} \frac{F(\boldsymbol{x}_t^j)}{|\mathcal{T}|}. \quad (7)$$

Therefore, our reformulation not only preserves the advantages of previous methods but also introduces instance-wise flexibility and explicit supervision on the overall structure.

Revisiting the Optimal Transport for Estimation of the Contribution Assignment. The problem of finding the optimal $\mathbf{P}$ can be implemented via an OT process. Specifically, we aim to determine $\mathbf{P}$ that minimizes the transport cost between real and synthetic distributions:

$$U = \{\mathbf{P} \in \mathbb{R}_+^{|\mathcal{T}|\times|\mathcal{S}|} | \mathbf{P}\mathbf{1}_{|\mathcal{S}|} = |\mathcal{S}|\mathbf{1}_{|\mathcal{T}|}, \mathbf{P}^{\mathrm{T}}\mathbf{1}_{|\mathcal{T}|} = |\mathcal{T}|\mathbf{1}_{|\mathcal{S}|}\}, \quad (8)$$

The optimal $\mathbf{P}$ can be efficiently approximated as:

$$\mathbf{P}^\lambda = \underset{\mathbf{P}\in U}{\operatorname{argmin}} \sum_{i,j} \mathbf{P}_{ij}\mathbf{C}(F(\boldsymbol{x}_t^i), F(\boldsymbol{x}_s^j)) - \lambda H(\mathbf{P}), \quad (9)$$

where $\mathbf{C}(F(\mathcal{T}), F(\mathcal{S}))$ denotes the cost matrix that measures the "distance" between the real $F(\boldsymbol{x}_t^i)$ and synthetic $F(\boldsymbol{x}_s^j)$ instances. The entropic regularization $H(\mathbf{P})$ is balanced by the hyper-parameter λ for smoothness.

Improving Perception of Distribution Nuances via Hilbert Space Conversion. Most existing dataset distillation methods simply use the low-order term of distance measurement for instance representations either in the gradient or the embedding space, neglecting the high-order terms that indicate the subtle nuances for a high level of alignment. To effectively capture the distribution structures of $\mathcal{T}$ and $\mathcal{S}$, we propose to map their representations into the "infinite-dimensional" Hilbert space via Gaussian kernels.

$$\mathcal{K}_{\sigma_k}(F(\boldsymbol{x}_t^i), F(\boldsymbol{x}_s^j)) = e^{-\frac{\|F(\boldsymbol{x}_t^i) - F(\boldsymbol{x}_s^j)\|^2}{2{\sigma_k}^2}}, \quad (10)$$

where σ_k controls the scale of the kernel's sensitivity to differences. By leveraging multiple σ_k, the complex relationship between $(\boldsymbol{x}_t^i, \boldsymbol{x}_s^j)$ pairs can be perceived properly:

$$\mathbf{R}_{ij} = \frac{1}{n}\sum_{k=1}^{n} \mathcal{K}_{\sigma_k}(F(\boldsymbol{x}_t^i), F(\boldsymbol{x}_s^j)), \quad (11)$$

where n is the number of kernels and $\mathbf{R}_{ij}$ denotes the entry of i-th row and j-th column of the relevance matrix $\mathbf{R}$. Compared with the vanilla measurement like cosine or Euclidean distance, the intricate pairwise discrepancy between $\mathcal{T}$ and $\mathcal{S}$ is comprehensively identified in our Hilbert space.

Cost Matrix Computation. The cost matrix $\mathbf{C}$ is implemented to be negatively associated with $\mathbf{R}$, implying that any pair of instances that share a high relevance $\mathbf{R}_{ij}$ should be assigned low cost for distributional "movement":

$$\mathbf{C} = \mathbf{J} - \mathbf{R}, \quad (12)$$

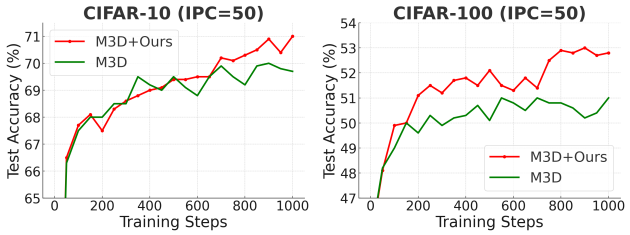

Figure 3. Accuracy progression over iteration.

where $\mathbf{J} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{S}|}$ is the all-ones matrix, $\mathbf{J}_{ij} = 1, \forall i, j$.

Sinkhorn Normalization. To efficiently solve the OT problem, we use the Sinkhorn algorithm to iteratively normalize a candidate transport matrix $\mathbf{K}^t$ to satisfy the marginal constraints [7]. It applies entropy regularization to efficiently approximate the OT matrix without sacrificing precision. We initialize $\mathbf{K}^0 \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{S}|}$ at $t = 0$ as:

$$\mathbf{K}^0 = \exp(-\frac{\mathbf{C}}{\lambda}). \quad (13)$$

Then, Sinkhorn normalization is performed by:

$$\begin{aligned} \widehat{\mathbf{K}}^t &\leftarrow \operatorname{diag}\left(\mathbf{K}^{t-1}\mathbf{1}_{|\mathcal{S}|} \oslash (|\mathcal{S}|\mathbf{1}_{|\mathcal{T}|})\right)^{-1}\mathbf{K}^{t-1}, \\ \mathbf{K}^t &\leftarrow \widehat{\mathbf{K}}^t \operatorname{diag}\left(\left(\widehat{\mathbf{K}}^t\right)^{\mathrm{T}}\mathbf{1}_{|\mathcal{T}|} \oslash (|\mathcal{T}|\mathbf{1}_{|\mathcal{S}|})\right)^{-1}, \end{aligned} \quad (14)$$

where $\oslash$ denotes element-wise division. These iterative updates can be simplified into the following component-wise forms that are well-suited for efficient processing:

$$\widehat{\mathbf{K}}_{ij}^t = \frac{|\mathcal{S}|\mathbf{K}_{ij}^{t-1}}{\sum_j \mathbf{K}_{ij}^{t-1}}, \quad \mathbf{K}_{ij}^t = \frac{|\mathcal{T}|\widehat{\mathbf{K}}_{ij}^t}{\sum_i \widehat{\mathbf{K}}_{ij}^t}. \quad (15)$$

After T iterations, we obtain the contribution matrix:

$$\mathbf{P}^\lambda = \mathbf{K}^T. \quad (16)$$

Boosting Synthesis-Based Distillation in a Plug-and-Play Manner. During the first-level optimization, we complete "matching" to explicitly quantify the contributions $\mathbf{P}^\lambda$. In the subsequent second-level "approximating", we perform the actual minimization of divergence between real and synthetic sets for authentic, informative data synthesis.

Specifically, Eq. (5) is minimized strictly following the batch-wise allocation scheme. By altering $F(\cdot)$ (see Table 1), we can flexibly incorporate both optimization-oriented and DM-based methods. It is noted that our reformulated distance is consistently updated batch-by-batch to ensure dynamic adjustment to the latest generated samples.

4. Experiments

4.1. Experimental Settings

Datasets. Our evaluation encompasses datasets of low (MNIST [25], FashionMNIST [47], SVHN [30], CIFAR-

Table 2. Top-1 accuracy of test models trained on distilled synthetic images. We use † to denote the results reproduced by us. The distillation is conducted with ConvNet-3. We apply our OPTICAL to existing synthesis-based distillation methods and respectively show its slight (blue) and non-trivial (red) performance gains (Δ). The best results are emphasized in **bold**.

Dataset	MNIST			FashionMNIST			SVHN			CIFAR-10			CIFAR-100		
IPC	1	10	50	1	10	50	1	10	50	1	10	50	1	10	50
Ratio (%)	0.017	0.17	0.83	0.017	0.17	0.83	0.014	0.14	0.7	0.02	0.2	1	0.2	2	10
DD [42]	-	79.5	-	-	-	-	-	-	-	-	36.8	-	-	-	-
LD [2]	60.9	87.3	93.3	-	-	-	-	-	-	25.7	38.3	42.5	11.5	-	-
DC [57]	91.7	97.4	98.8	70.5	82.3	83.6	31.2	76.1	82.3	28.3	44.9	53.9	12.8	25.2	-
KIP [31, 32]	90.1	97.5	98.3	73.5	86.8	88.0	57.3	75.0	80.5	49.9	62.7	68.6	15.7	28.3	-
DSA [55]	88.7	97.8	99.2	70.6	84.6	88.7	27.5	79.2	84.4	28.8	52.1	60.6	13.9	32.3	42.8
IDC [19]	94.2	98.4	99.1	81.0	86.0	86.2	68.5	87.5	90.1	50.6	67.5	74.5	-	45.1	-
MTT [3]	-	-	-	-	-	-	-	-	-	46.3	65.3	71.6	24.3	40.1	47.7
FTD [15]	-	-	-	-	-	-	-	-	-	46.8	66.6	73.8	25.2	43.4	50.7
ATT [26]	-	-	-	-	-	-	-	-	-	48.3	67.7	74.5	26.1	44.2	51.2
DREAM [27]	95.7	98.6	99.2	81.3	86.4	86.8	69.8	87.9	90.5	51.1	69.4	74.8	29.5	46.8	52.6
DREAM+Ours	**96.0**	**98.8**	**99.3**	81.7	86.7	87.4	**70.7**	**88.2**	90.8	**52.0**	69.7	75.0	**29.7**	47.1	52.9
Δ	+0.3	+0.2	+0.1	+0.4	+0.3	+0.6	+0.9	+0.3	+0.3	+0.9	+.0.3	+0.2	+0.2	+0.3	+0.3
CAFE [41]	93.1	97.2	98.6	77.1	83.0	84.8	42.6	75.9	81.3	30.3	46.3	55.5	12.9	27.8	37.9
CAFE+DSA	-	-	-	73.7	83.0	88.2	-	-	-	31.6	50.9	62.3	14.0	31.5	42.9
DataDAM [37]	-	-	-	-	-	-	-	-	-	32.0	54.2	67.0	14.5	34.8	49.4
DM [56]	89.7	97.5	98.6	70.7	83.5	88.1	30.3	73.5	82.0	26.0	48.9	63.0	11.4	29.7	43.6
DM+Ours	-	97.9	98.8	-	85.0	**88.9**	-	75.2	84.8	-	50.2	63.4	-	30.8	44.5
Δ	-	+0.4	+0.2	-	+1.5	+0.8	-	+1.7	+2.8	-	+1.3	+0.4	-	+1.1	+0.9
IDM [58]	-	-	-	-	-	-	-	-	-	45.6	58.6	67.5	20.1	45.1	50.0
IDM+Ours	-	-	-	-	-	-	-	-	-	46.3	59.1	67.8	24.4	45.8	50.4
Δ	-	-	-	-	-	-	-	-	-	+0.7	+0.5	+0.3	+4.3	+0.7	+0.4
M3D [51]	94.4	97.6	98.2	80.7	85.0	86.2	58.1†	82.5†	89.0	45.3	63.5	69.9	26.2	42.4	50.9
M3D+Ours	95.9	97.9	98.7	81.8	86.0	86.3	59.9	83.5	89.6	46.1	64.0	71.0	26.8	42.9	53.0
Δ	+1.5	+0.3	+0.5	+1.1	+1.0	+0.1	+1.8	+1.0	+0.6	+0.8	+0.5	+1.1	+0.6	+0.5	+2.1
DANCE [50]	94.5†	96.2†	96.5†	81.5	86.3	86.9	65.8†	86.6†	90.2†	47.1	70.8	76.1	27.9	49.8	52.8
DANCE+Ours	95.4	97.5	97.8	**82.0**	**86.9**	87.7	67.3	88.1	**91.1**	48.0	**71.1**	**76.5**	28.3	**50.7**	**53.5**
Δ	+0.9	+1.3	+1.3	+0.5	+0.6	+0.8	+1.5	+1.5	+0.9	+0.9	+0.3	+0.4	+0.4	+0.9	+0.7
Full Dataset		99.6			93.5			95.4			84.8			56.2	

10 [21], CIFAR-100 [21]), medium (TinyImageNet [24]), and high (ImageNet subsets [11]) resolutions.

Network Architectures. Following previous works, we use a depth-3 ConvNet [36] for the low-resolution datasets, a depth-4 ConvNet for TinyImageNet [24], and a ResNetAP-10 [19] for the high-resolution ImageNet subsets.

Baselines. We evaluate the impact of incorporating our OPTICAL into multiple approaches, including DREAM [27], DM [56], IDM [58], M3D [51] and DANCE [50]. We also compare our results against various methods including DD [42], LD [2], DC [57], KIP [32], DSA [55], MTT [3], FTD [15], ATT [26], CAFE, and CAFE+DSA [41].

Metric. Following previous works [50, 51], we employ the test accuracy of networks trained on the distilled examples as the evaluation metric. Each network is trained from scratch multiple times: 10 times for low-resolution datasets and TinyImageNet, and 3 times for the ImageNet subsets.

Implementation Details. We follow the same experimental settings (e.g., α) as previous works for comparability. As for the OT, we set ${\sigma_k}^2 = [0.1, 1, 5]$ and $T = 20$. Our $1/\lambda$ is optimized via grid search in $\{1.0, 1.1, 1.2, 1.3\}$.

4.2. Results and Discussions

Results on Low-Resolution Datasets. As shown in Table 2, we apply the OPTICAL framework to one optimization-oriented method and four DM-based methods. Results indicate that OPTICAL consistently improves performance across various datasets and image-per-class (IPC) settings. This enhancement stems from our OT matrix that captures the geometric structures of the real data distribution. It leverages intra-class relationship for more effective alignment between real and synthetic data. Notably, due to the factor and up-sampling techniques in prior methods [27, 50, 51, 58], *the synthetic set typically contains more images than the IPC*, further amplifying the benefits of the OT matrix. The greatest performance gains are observed with DM [56], which employs a basic form of $\mathbf{D}_s$ and suffers from instance-wise homogeneity (Sec. 3.2). For other methods that adopt more complex forms of $\mathbf{D}_s$ to convey category-specific knowledge, the limitation of uniform contribution still remains. In this case, our method introduces adaptive non-uniform contributions in return for more representative synthesis. Even state-of-the-art ap-

Table 3. Performance on TinyImageNet with ConvNet-4.

Type	IPC / Ratio (%)	1 / 0.2	10 / 2	50 / 10
Optimization-Oriented	DC [57]	$5.3_{\pm 0.1}$	$12.9_{\pm 0.1}$	$12.7_{\pm 0.4}$
	DSA [55]	$5.7_{\pm 0.1}$	$16.3_{\pm 0.2}$	$5.1_{\pm 0.2}$
	DREAM [27]	$10.0_{\pm 0.4}$	-	$29.5_{\pm 0.3}$
	MTT [3]	$6.2_{\pm 0.4}$	$17.3_{\pm 0.2}$	$26.5_{\pm 0.3}$
Distribution-Matching-Based	DataDAM [37]	$8.3_{\pm 0.4}$	$18.7_{\pm 0.3}$	$28.7_{\pm 0.3}$
	DM [56]	$3.9_{\pm 0.2}$	$12.9_{\pm 0.4}$	$24.1_{\pm 0.3}$
	DM+Ours	$4.7_{\pm 0.3}$	$13.8_{\pm 0.3}$	$25.1_{\pm 0.2}$
	Δ	+0.8	+0.9	+1.0
	IDM [58]	$10.1_{\pm 0.2}$	$21.9_{\pm 0.2}$	$27.7_{\pm 0.3}$
	IDM+Ours	$10.7_{\pm 0.2}$	$22.5_{\pm 0.4}$	$28.1_{\pm 0.4}$
	Δ	+0.6	+0.6	+0.4
	DANCE [50]	$11.6_{\pm 0.2}$	$26.4_{\pm 0.3}$	$28.9_{\pm 0.4}$
	DANCE+Ours	**12.3**$_{\pm 0.2}$	**26.9**$_{\pm 0.4}$	**29.6**$_{\pm 0.4}$
	Δ	+0.7	+0.5	+0.7
Full Dataset		$37.6_{\pm 0.4}$		

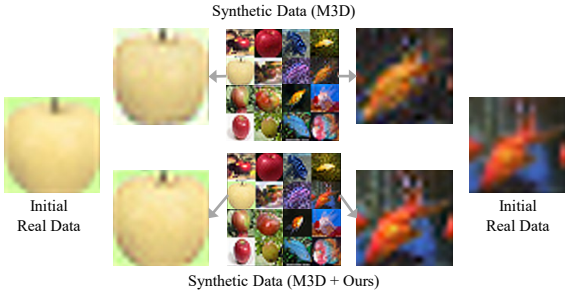

Figure 4. Visualization of the distilled set of CIFAR-100.

proaches, such as the optimization-oriented DREAM [27] and the DM-based DANCE [50], benefit significantly from our OPTICAL. DREAM selects representative original images, providing a strong initialization based on a fixed number of subclasses. However, as sample-to-subclass relationships should ideally be dynamic, our adaptive contribution allocation further enhances overall performance.

To further demonstrate our advantages, we present test accuracy over training steps in Fig. 3. The baseline's performance plateaus around 700 steps, whereas our approach continues to improve. This suggests that the baseline stucks in a suboptimal convergence by solely matching the mean statistics of real data. In contrast, OPTICAL exploits geometric information of the real data throughout training to closely approximate their distributions.

We visualize synthesized images in Fig. 4. M3D produces over-smoothed synthetic data that lack fine-grained details. It alters irrelevant attributes (e.g., background) due to unintended distribution shifts. In contrast, our OPTICAL maintains category-irrelevant elements untouched while highlighting category-specific main bodies, indicating improved retention of discriminative information.

Results on Higher-Resolution Datasets. Following M3D [51] and DANCE [50], we further extend our experiments to higher-resolution datasets to evaluate the scalability and robustness of our approach. Results are shown in Table 3 and 4. In these scenarios, where the image complexity is considerably higher, our method demonstrates its superior generalization capabilities.

Table 4. Performance on ImageNet-subsets with ResNetAP-10.

IPC	ImageNet-10 10		ImageNet-10 20		ImageNet-100 10		ImageNet-100 20	
	Base	Ours	Base	Ours	Base	Ours	Base	Ours
DM [56]	52.3	**53.6**	59.3	**60.1**	22.3	**24.0**	30.4	**31.1**
M3D [51]	73.4	**74.7**	76.8	**78.5**	46.9	**48.1**	55.5	**56.4**

Table 5. Performance on CIFAR-10 with the **same architectures** for both dataset distillation and evaluation.

IPC	Method	ConvNet-3	ResNet-10	DenseNet-121
10	DM [56]	$48.9_{\pm 0.6}$	$39.4_{\pm 0.6}$	$33.3_{\pm 0.8}$
	DM+Ours	$50.2_{\pm 0.7}$	$40.8_{\pm 0.7}$	$36.5_{\pm 0.6}$
	Δ	+1.3	+1.4	+3.2
	M3D [51]	$63.5_{\pm 0.2}$	$53.4_{\pm 0.3}$	$45.3_{\pm 0.3}$
	M3D+Ours	$64.0_{\pm 0.3}$	$54.1_{\pm 0.3}$	$46.0_{\pm 0.3}$
	Δ	+0.5	+0.7	+0.7
	DANCE [50]	$70.8_{\pm 0.2}$	$61.2_{\pm 0.3}$	$49.3_{\pm 1.1}$
	DANCE+Ours	$71.1_{\pm 0.1}$	$61.7_{\pm 0.2}$	$51.1_{\pm 0.7}$
	Δ	+0.3	+0.5	+0.8
50	DM [56]	$63.0_{\pm 0.4}$	$63.1_{\pm 0.2}$	$50.2_{\pm 0.1}$
	DM+Ours	$63.4_{\pm 0.3}$	$64.0_{\pm 0.1}$	$50.9_{\pm 0.1}$
	Δ	+0.4	+0.9	+0.7
	M3D [51]	$69.9_{\pm 0.5}$	$64.9_{\pm 0.5}$	$57.6_{\pm 0.2}$
	M3D+Ours	$71.0_{\pm 0.6}$	$65.9_{\pm 0.4}$	$59.0_{\pm 0.3}$
	Δ	+1.1	+1.0	+1.4
	DANCE [50]	$76.1_{\pm 0.1}$	$66.8_{\pm 0.5}$	$64.9_{\pm 0.2}$
	DANCE+Ours	$76.5_{\pm 0.1}$	$67.5_{\pm 0.5}$	$65.4_{\pm 0.3}$
	Δ	+0.4	+0.7	+0.5

Results on Different Networks. Unlike most studies limited to certain architectures, we demonstrate our versatility across multiple architectures. Table 5 shows that our approach consistently enhances performance across different backbones. Notably, it yields even greater improvements in deeper architectures like ResNet and DenseNet, which are more sensitive to synthetic data quality, underscoring our method's adaptability to diverse network structures.

Cross-Architecture Evaluation. One of the key strengths of our method lies in its robustness when the synthetic dataset is trained on one architecture and evaluated on another, a scenario we termed as cross-architecture evaluation. As shown in Tables 6 and 7, our method exhibits even greater gains when the network used for distillation differs from the one used for evaluation. This demonstrates strong generalization ability of our approach, which is essential for real-world applications where the deployment environment may differ from the training setup.

Table 6. Performance on CIFAR-10 with **different architectures** across dataset distillation and evaluation. Synthetic data are condensed using ConvNet-3 and evaluated with three architectures.

IPC	Method	ConvNet-3	ResNet-10	DenseNet-121
10	DM [56]	$48.9_{\pm 0.6}$	$42.3_{\pm 0.5}$	$39.0_{\pm 0.1}$
	DM+Ours	$50.2_{\pm 0.7}$	$44.1_{\pm 0.4}$	$41.9_{\pm 0.1}$
	Δ	+1.3	+2.2	+2.9
	M3D [51]	$63.5_{\pm 0.2}$	$56.7_{\pm 0.3}$	$54.6_{\pm 0.2}$
	M3D+Ours	$64.0_{\pm 0.3}$	$57.1_{\pm 0.2}$	$55.0_{\pm 0.2}$
	Δ	+0.5	+0.4	+0.4
	DANCE [50]	$70.8_{\pm 0.2}$	$67.0_{\pm 0.2}$	$64.5_{\pm 0.3}$
	DANCE+Ours	$71.1_{\pm 0.1}$	$67.6_{\pm 0.2}$	$66.8_{\pm 0.2}$
	Δ	+0.3	+0.6	+2.3
50	DM [56]	$63.0_{\pm 0.4}$	$58.6_{\pm 0.3}$	$57.4_{\pm 0.3}$
	DM+Ours	$63.4_{\pm 0.3}$	$59.0_{\pm 0.2}$	$58.2_{\pm 0.3}$
	Δ	+0.4	+0.4	+0.8
	M3D [51]	$69.9_{\pm 0.5}$	$66.6_{\pm 0.3}$	$66.1_{\pm 0.4}$
	M3D+Ours	$71.0_{\pm 0.6}$	$67.3_{\pm 0.2}$	$66.8_{\pm 0.3}$
	Δ	+1.1	+0.7	+0.7
	DANCE [50]	$76.1_{\pm 0.1}$	$68.0_{\pm 0.1}$	$64.8_{\pm 0.3}$
	DANCE+Ours	$76.5_{\pm 0.1}$	$72.7_{\pm 0.1}$	$70.3_{\pm 0.2}$
	Δ	+0.4	+4.7	+5.5

Table 7. Performance on CIFAR-10 with **different architectures** across dataset distillation and evaluation. Synthetic data are condensed using three architectures and evaluated with ConvNet-3.

IPC	Method	ConvNet-3	ResNet-10	DenseNet-121
10	DM [56]	$48.9_{\pm 0.6}$	$42.0_{\pm 0.4}$	$33.4_{\pm 0.4}$
	DM+Ours	$50.2_{\pm 0.7}$	$44.5_{\pm 0.5}$	$36.2_{\pm 0.6}$
	Δ	+1.3	+2.5	+2.8
	M3D [51]	$63.5_{\pm 0.2}$	$57.3_{\pm 0.2}$	$47.6_{\pm 0.4}$
	M3D+Ours	$64.0_{\pm 0.3}$	$57.9_{\pm 0.1}$	$48.6_{\pm 0.2}$
	Δ	+0.5	+0.6	+1.0
	DANCE [50]	$70.8_{\pm 0.2}$	$59.5_{\pm 0.3}$	$49.3_{\pm 0.5}$
	DANCE+Ours	$71.1_{\pm 0.1}$	$60.5_{\pm 0.2}$	$51.1_{\pm 0.3}$
	Δ	+0.3	+1.0	+1.8
50	DM [56]	$63.0_{\pm 0.4}$	$60.2_{\pm 0.1}$	$50.2_{\pm 0.1}$
	DM+Ours	$63.4_{\pm 0.3}$	$60.6_{\pm 0.3}$	$50.9_{\pm 0.1}$
	Δ	+0.4	+0.4	+0.7
	M3D [51]	$69.9_{\pm 0.5}$	$67.5_{\pm 0.3}$	$60.9_{\pm 0.3}$
	M3D+Ours	$71.0_{\pm 0.6}$	$68.4_{\pm 0.2}$	$61.8_{\pm 0.2}$
	Δ	+1.1	+0.9	+0.9
	DANCE [50]	$76.1_{\pm 0.1}$	$67.9_{\pm 0.3}$	$66.8_{\pm 0.1}$
	DANCE+Ours	$76.5_{\pm 0.1}$	$68.3_{\pm 0.2}$	$67.3_{\pm 0.1}$
	Δ	+0.4	+0.4	+0.5

Table 8. Effect of kernel function choices on CIFAR-10 performance using DM+Ours (Images Per Class = 10).

	N/A	Linear	Gauss (single)	Gauss (ours)
Acc	49.5	47.8	49.8	**50.2**

Impact of the Kernal Function. As shown in Table 9, our Gaussian kernel achieves superior alignment between real and synthetic data distributions by capturing higher-order feature interactions, leading to the best performance.

Table 9. Impact of the entropic regularization parameter λ and the number of Sinkhorn iterations T on CIFAR-100 performance with M3D+Ours (Images Per Class = 50).

$1/\lambda$	0.8	1	1.1	1.2	1.3	1.5
Acc	52.2	52.8	52.6	53.0	52.8	53.0
T	2	5	10	20	40	100
Acc	52.3	52.5	52.8	53.0	52.8	53.0

Table 10. Run-time per iteration (b=128) on CIFAR-10 and ImageNet with one RTX 4090 GPU (measured in seconds).

Dataset	IPC	DM	w/ Ours	M3D	w/ Ours	DANCE	w/ Ours
CIFAR-10	50	0.066	0.097	0.075	0.105	0.135	0.168
CIFAR-10	10	0.065	0.095	0.068	0.097	0.111	0.143
ImgNet-10	10	0.912	0.942	1.139	1.168	1.603	1.634

Sensitivity Analysis. Table 9 shows the effect of varying the entropic regularization parameter λ and the number of Sinkhorn iterations T. Results indicate that accuracy remains stable across a range of λ values, showing minimal variance around 53.0%. For T, performance stabilizes once T reaches 20 iterations, suggesting that further increasing T offers limited additional benefit. These observations underscore the robustness of the method.

Complexity Analysis. When integrated with existing distillation methods, our OPTICAL introduces an additional complexity of $O(Tb|\mathcal{S}|)$ per batch where b is the batch size ($b \ll |\mathcal{T}|$) and T denotes iterations. The Sinkhorn normalization allows parallel GPU computation, with runtime per iteration consistently around 0.03 seconds (see Table 10) across architectures, IPC, and datasets. This confirms the scalability and training efficiency for practical applications.

5. Conclusion

In this paper, we address the limitations of homogeneous distance minimization in dataset distillation. We propose the OPTICAL to offer an adaptive contribution allocation scheme that works with any synthesis-based methods in a versatile plug-and-play manner. OPTICAL employs an OT matrix to dynamically allocate contributions in the distance measure between real-synthetic pairs, enhancing the representativeness and generalization of synthetic data. Extensive experiments across diverse datasets, frameworks, and model architectures demonstrate the effectiveness, generalizability, and adaptability of our OPTICAL.

Acknowledgement. This work was supported by National Natural Science Foundation of China under Contract 62021001 , and the Youth Innovation Promotion Association CAS. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. 3

[2] Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Flexible dataset distillation: Learn labels instead of images. *arXiv preprint arXiv:2006.08572*, 2020. 1, 6

[3] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, pages 4750–4759, 2022. 2, 3, 6, 7

[4] Pengfei Chen, Rongzhen Zhao, Tianjing He, Kongyuan Wei, and Qidong Yang. Unsupervised domain adaptation of bearing fault diagnosis based on join sliced wasserstein distance. *ISA transactions*, 129:504–519, 2022. 3

[5] Xiao Cui, Wengang Zhou, Yang Hu, Weilun Wang, and Houqiang Li. Heredity-aware child face image generation with latent space disentanglement. *arXiv preprint arXiv:2108.11080*, 2021. 1

[6] Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkd: Sinkhorn distance minimization for knowledge distillation. *TNNLS*, 2024. 3

[7] Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkhorn distance minimization for knowledge distillation. In *LREC-COLING*, pages 14846–14858, 2024. 5

[8] Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. *arXiv preprint arXiv:2412.14528*, 2024. 3

[9] Xiao Cui, Qi Sun, Min Wang, Li Li, Wengang Zhou, and Houqiang Li. Layoutenc: Leveraging enhanced layout representations for transformer-based complex scene synthesis. *TOMM*, 2025. 1

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26, 2013. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[12] 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 *CVPR*, pages 17057–17066, 2024. 1

[13] 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 *CVPR*, pages 17057–17066, 2024. 3

[14] Bin Du, Wei Xie, Yang Li, Qisong Yang, Weidong Zhang, Rudy R Negenborn, Yusong Pang, and Hongtian Chen. Safe adaptive policy transfer reinforcement learning for distributed multiagent control. *TNNLS*, 2023. 3

[15] Jiawei Du, Yidi Jiang, Vincent Y. F. Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *CVPR*, pages 3749–3758, 2023. 6

[16] Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: concepts, models, algorithms and case studies*. Springer Science & Business Media, 2009. 1, 2

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *NeurIPS*, 30, 2017. 3

[18] Shuncheng He, Yuhang Jiang, Hongchang Zhang, Jianzhun Shao, and Xiangyang Ji. Wasserstein unsupervised reinforcement learning. In *AAAI*, pages 6884–6892, 2022. 3

[19] 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. *arXiv preprint arXiv:2205.14959*, 2022. 2, 3, 6

[20] Pascal Klink, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. On the benefit of optimal transport for curriculum reinforcement learning. *TPAMI*, 2024. 3

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[22] Yixing Lan, Xin Xu, Qiang Fang, and Jianye Hao. Sample efficient deep reinforcement learning with online state abstraction and causal transformer model prediction. *TNNLS*, 2023. 3

[23] Tung Le, Khai Nguyen, Shanlin Sun, Nhat Ho, and Xiaohui Xie. Integrating efficient optimal transport and functional maps for unsupervised shape correspondence learning. In *CVPR*, pages 23188–23198, 2024. 3

[24] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. 5

[26] Dai Liu, Jindong Gu, Hu Cao, Carsten Trinitis, and Martin Schulz. Dataset distillation by automatic training trajectories. *arXiv preprint arXiv:2407.14245*, 2024. 1, 6

[27] Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. Dream: Efficient dataset distillation by representative matching. In *ICCV*, pages 17314–17324, 2023. 2, 4, 6, 7

[28] Yanbin Liu, Linchao Zhu, Xiaohan Wang, Makoto Yamada, and Yi Yang. Bilaterally normalized scale-consistent sinkhorn distance for few-shot image classification. *TNNLS*, 2023. 3

[29] Brian B Moser, Federico Raue, Sebastian Palacio, Stanislav Frolov, and Andreas Dengel. Latent dataset distillation with diffusion models. *arXiv preprint arXiv:2403.03881*, 2024. 1

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[31] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *ICLR*, 2020. 6

[32] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *NeurIPS*, 34:5186–5198, 2021. 6

[33] Thong Thanh Nguyen and Anh Tuan Luu. Improving neural cross-lingual abstractive summarization via employing opti-

mal transport distance for knowledge distillation. In *AAAI*, pages 11103–11111, 2022. 3

[34] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *FTML*, 11(5-6): 355–607, 2019. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[36] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv*, 2017. 6

[37] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *ICCV*, pages 17097–17107, 2023. 1, 4, 6, 7

[38] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 1, 2

[39] 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. In *ICLR*, 2019. 1, 2

[40] Cédric Villani and Cédric Villani. The wasserstein distances. *Optimal Transport: Old and New*, pages 93–111, 2009. 3

[41] 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 *CVPR*, pages 12196–12205, 2022. 2, 3, 6

[42] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1, 3, 6

[43] Yi Wei, Xiaofei Li, Lihui Lin, Dengming Zhu, and Qingyong Li. Causal discovery on discrete data via weighted normalized wasserstein distance. *TNNLS*, 2022. 3

[44] Chen Weilin, Qiao Jie, Cai Ruichu, and Hao Zhifeng. On the role of entropy-based loss for learning causal structure with continuous optimization. *TNNLS*, 2023. 3

[45] Max Welling. Herding dynamical weights to learn. In *ICML*, pages 1121–1128, 2009. 1, 2

[46] Xiaobo Xia, Jiale Liu, Shaokun Zhang, Qingyun Wu, Hongxin Wei, and Tongliang Liu. Refined coreset selection: Towards minimal coreset size under model performance constraints. In *ICML*, 2024. 1

[47] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, 2017. 5

[48] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *TPAMI*, 46(1):150–170, 2024. 1

[49] Zhichen Zeng, Boxin Du, Si Zhang, Yinglong Xia, Zhining Liu, and Hanghang Tong. Hierarchical multi-marginal optimal transport for network alignment. In *AAAI*, pages 16660–16668, 2024. 3

[50] Hansong Zhang, Shikun Li, Fanzhao Lin, Weiping Wang, Zhenxing Qian, and Shiming Ge. Dance: Dual-view distribution alignment for dataset condensation. *arXiv preprint arXiv:2406.01063*, 2024. 2, 3, 4, 6, 7, 8

[51] Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. M3d: Dataset condensation by minimizing maximum mean discrepancy. In *AAAI*, pages 9314–9322, 2024. 2, 3, 4, 6, 7, 8

[52] Hansong Zhang, Shikun Li, Dan Zeng, Chenggang Yan, and Shiming Ge. Coupled confusion correction: Learning from crowds with sparse annotations. In *AAAI*, pages 16732–16740, 2024. 1

[53] Jingwei Zhang, Tongliang Liu, and Dacheng Tao. An optimal transport analysis on generalization in deep learning. *TNNLS*, 34(6):2842–2853, 2021. 3

[54] Qiyuan Zhang, Shu Leng, Xiaoteng Ma, Qihan Liu, Xueqian Wang, Bin Liang, Yu Liu, and Jun Yang. Cvar-constrained policy optimization for safe reinforcement learning. *TNNLS*, 2024. 3

[55] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, pages 12674–12685. PMLR, 2021. 2, 6, 7

[56] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, pages 6514–6523, 2023. 2, 3, 4, 6, 7, 8

[57] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2020. 2, 3, 4, 6, 7

[58] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *CVPR*, pages 7856–7865, 2023. 2, 3, 4, 6, 7