Four eyes see more than two: Dataset Distillation with Mixture-of-Experts

Jia-Jiun Yao[1]
tony.yao.10@nycu.edu.tw
Sheng-Feng Yu[12]
robertyu1@mxic.com.tw
Wei-Chen Chiu[1]
walon@cs.nctu.edu.tw

[1] National Yang Ming Chiao Tung University
[2] Macronix International Co., Ltd.

Abstract

The ever-growing size of datasets in deep learning presents a significant challenge in terms of training efficiency and computational cost. Dataset distillation (DD) has emerged as a promising approach to address this challenge by generating compact synthetic datasets that retain the essential information of the original data. However, existing DD methods often suffer from performance degradation when transferring distilled datasets across different network architectures (i.e. the model utilizing distilled dataset for further training is different from the one used in dataset distillation). To overcome this limitation, we propose a novel mixture-of-experts framework for dataset distillation. Our goal focuses on promoting diversity within the distilled dataset by distributing the distillation tasks to multiple expert models. Each expert specializes in distilling a distinct subset of the dataset, encouraging them to capture different aspects of the original data distribution. To further enhance diversity, we introduce a distance correlation minimization strategy to encourage the experts to learn distinct representations. Moreover, during the testing stage (where the distilled dataset is used for training a new model), the mixup-based fusion strategy is applied to better leverage the complementary information captured by each expert. Through extensive experiments, we demonstrate that our framework effectively mitigates the issue of cross-architecture performance degradation in dataset distillation, particularly in low-data regimes, leading to more efficient and versatile deep learning models while being trained upon the distilled dataset.

1 Introduction

While large datasets like ImageNet [4] have demonstrably fueled the success of deep learning models, their sheer size presents a significant challenge in terms of training efficiency. The computational resources required to train models on these massive datasets can be prohibitively expensive, both in terms of time and financial cost. This burden is further amplified in scenarios like continual learning, hyperparameter optimization, and architecture search, where iterative training processes are essential. As the demand for data in deep learning continues to escalate, finding innovative solutions to mitigate the associated training costs becomes increasingly critical.

Two primary approaches have emerged to tackle the challenge associated with large datasets in deep learning: dataset distillation (DD) [1, 9, 15, 21, 22] and coreset selection [3]. Coreset selection focuses on identifying a highly representative subset of the original data, known as a coreset. However, due to storage constraints, this subset may only capture a limited amount of information, potentially failing to well represent the complexities of the entire dataset. DD offers an alternative solution by generating a synthetic dataset optimized to preserve model performance (i.e. the model trained upon original dataset and the one trained upon distilled dataset ideally should perform similarly). While both DD and coreset selection operate under the same storage limitations, DD has the advantage of encoding a richer set of information and details compared to a coreset, thus benefiting the downstream tasks such as image classification, continual learning [16, 18], and federated learning [17].

Since its inception by [15], dataset distillation has witnessed a surge in research efforts, with various approaches exploring different facets of the technique. These approaches can be broadly categorized into three main streams: **1) Gradient Matching** which basically aims to align the gradients generated from synthetic and real data for guiding the distillation process [9, 21]; **2) Trajectory Matching** which attempts to align the training trajectories of synthetic and real datasets, ensuring that the optimization path of the model learnt from the distilled data closely resembles that of the original data [1]; **3) Distribution Matching** which focuses on matching the latent distribution of features extracted from real and synthetic data, encouraging the distilled data to capture the underlying statistical properties of the original data. These diverse approaches highlight the ongoing efforts to refine and optimize DD techniques, facilitating the generation of more compact and informative synthetic training datasets to fuel the deep model training.

While the conventional wisdom in deep learning suggests that larger models generally outperform smaller models when being trained on a given dataset, dataset distillation (DD) presents a contrasting scenario. Synthetic datasets generated through DD often exhibit a tendency to overfit to the specific neural network architecture used during the distillation process. This phenomenon explains the prevalent use of small architectures like ConvNet-3 in DD research for performance comparison. However, when these distilled datasets are transferred to larger and more complex architectures, such as ResNet-18 [7], AlexNet [11], or VGG-11 [12], a significant degradation in classification performance is often observed. This limitation poses a significant challenge to the broader applicability of DD, as it restricts the transferability of distilled knowledge across diverse model architectures. Consequently, there is a pressing need for novel DD methods that explicitly address this issue and promote generalizability across a wider range of network architectures.

To address this challenge, we propose a novel mixture-of-experts (MoE) framework for dataset distillation. The MoE framework, drawing inspiration from [24] and [25], enhances representation diversity by distilling disjoint subsets using different distillation algorithms, thus reducing overfitting and improving cross-architecture transferability. In particularly, unlike alternative strategies such as meta learning or data augmentation, MoE explicitly encourages diverse representations by distributing the distillation task among specialized experts, where each expert captures distinct aspects of data distribution, and our distance correlation minimization further ensures that representations remain diverse rather than redundant.

Our approach leverages the concept of dividing the distillation task among multiple expert models, each focusing on a distinct subset of the original data. This division of labor encourages specialization within each expert, potentially capturing different aspects of the data distribution and leading to a more diverse and informative distilled dataset, in which the enhanced diversity results to benefit the cross-architecture generalizability. For instance, under the constraint upon Image Per Classes (IPC), we split the distillation process into two parallel paths (two experts), with having each expert model responsible for distilling the original dataset into half of the allocated storage budget (noting that overall storage budget, i.e. the number of distilled images, hence is kept the same as the original dataset distillation method which does not have multiple experts). To further promote diversity within the synthetic dataset, we introduce a distance correlation minimization strategy: during each iteration of the distillation process, the distance correlation between the expert models is calculated and subsequently minimized. This step encourages the experts to generate synthetic data that is dissimilar across experts, thereby enriching the overall information content captured by the entire distilled dataset.

Finally, while one attempts to leverage the distilled dataset to train a new model, the fusion strategy based on Mixup [19] can be seamlessly introduced to combine the synthetic images from each expert, thanks to the complementary information captured across individual experts, leading to a more comprehensive and generalizable representation of the original data. To evaluate the effectiveness of our proposed framework, we conducted experiments with various IPC settings as well as adopting different dataset distillation methods to be experts, comparing the performance of single-expert and multiple-expert distillations. Notably, we observed a significant improvement in average cross-architecture performance with multiple experts, particularly in lower IPC conditions (e.g. ≤ 25). This finding suggests that our approach effectively mitigates the transferability issue by promoting diversity and generalizability within the distilled dataset.

In summary, our work makes the following key contributions:

- We introduce a novel mixture-of-experts framework for dataset distillation that addresses the challenge of cross-architecture performance degradation.
- Our approach leverages distance correlation minimization and a fusion strategy to promote diversity and enhance the generalizability of the distilled dataset.
- We demonstrate the effectiveness of our framework by evaluating its performance with popular surrogate distillation methods, including MTT [1], IDC [9], and IDM [23].

2 Related Work

Dataset Distillation. Pioneered by [15], dataset distillation (DD) is a method that aims to condense large training datasets into a compact and informative synthetic set, enabling models trained on this distilled data to achieve comparable performance to those trained on the full dataset. The core of DD lies in a bilevel optimization framework. The inner loop minimizes the training loss of a model using distilled data, while the outer loop optimizes the distilled data based on the matching objective. Several studies have explored various objectives to enhance DD performance: **1) Gradient Matching** – DC [21] optimizes synthetic images by matching gradients of neural networks' weights that are trained on real and distilled data, while DSA [20] incorporating differentiable data augmentation to synthesize more informative images. IDC [9] follows DC but particularly starts the iterative procedure of dataset distillation via training the networks with real data (instead of the synthetic ones) to ensure the quality of gradients. Moreover, under limited storage constraints, IDC stores synthetic images in a low-resolution format, which will be up-sampled before being used to train a new network; **2) Trajectory Matching** – MTT [1] starts from training a set of classification models on real data, recording their optimization trajectories. Subsequently, the distillation process focuses on aligning the training trajectories of models trained on the distilled data with those obtained from real data. Due to the reduced size of the distilled dataset, the training steps for synthetic data are typically larger (approximately 10x to 30x) than those used for real data. FTD [5] further refines MTT by addressing its issue of having interrupted trajectories between iterations, promoting a smoother and flatter trajectory for enhanced performance; **3) Distribution Matching** – DM [22] and IDM [23] aim to align the feature representations learned from real and synthetic datasets, while IDM enhances the classification ability of distilled images by incorporating an additional cross-entropy loss into the objective function. Furthermore, IDM introduces a mechanism of constructing a model pool to maintain diversity among the models used for distillation, alleviating the potential issue of overfitting to specific features or representations. **Mixture-of-Experts.** Building upon multiple expert sub-networks in which each focuses on its specific sub-task or sub-domain within the overall problem space, the Mixture-of-Experts (MoE) architecture [8, 25] presents a compelling approach to leverage the specialized knowledge from these sub-networks within a single and unified model, where a gating module dynamically routes input data to the most relevant expert based on its characteristics, allowing the model to adapt to diverse inputs and learn distinct representations. In this paper, we adapt the core idea of MoE analogously to the problem of dataset distillation, where multiple experts of dataset distillation are encouraged to capture the characteristics of the original dataset from diverse perspectives, making the resultant distilled dataset more generalizable for further use of training models of various network architectures.

3 Methodology

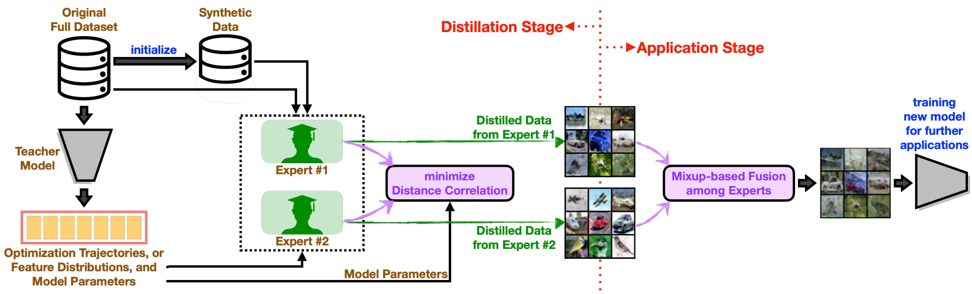

Figure 1: The overall workflow of our proposed Mixture-of-Experts framework for Dataset Distillation. Initially, we pretrain and store multiple model parameters. Then each expert initializes its synthetic data from the original full dataset. During the distillation stage, each expert updates synthetic data by minimizing distance correlation between experts to promote diversity within whole synthetic data. Finally, after the distillation is finished, we apply mixup-based fusion among experts to enhance the transferability and performance of the distilled dataset when training a new model for further applications.

3.1 Preliminary

Dataset Distillation. Given a dataset $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|\mathcal{T}|}$ composed of $|\mathcal{T}|$ real images $\boldsymbol{x}$ and their corresponding labels $\boldsymbol{y}$, the main goal of data distillation is to compress $\mathcal{T}$ into a learnable synthetic dataset $\mathcal{S} = \{(\boldsymbol{x}'_i, \boldsymbol{y}'_i)\}_{i=1}^{|\mathcal{S}|}$ with $|\mathcal{S}| \ll |\mathcal{T}|$, while attempting to encapsulate the knowledge of $\mathcal{T}$ in $\mathcal{S}$. With taking a model parameterized by θ as the bridge, the learning of such synthetic dataset is typically realized by minimizing the discrepancy between the model's behavior when being trained on $\mathcal{T}$ and $\mathcal{S}$ (i.e. the model trained on $\mathcal{S}$ ideally should have comparable performance with the one trained on $\mathcal{T}$) and can be generally written as:

$$\mathcal{S} = \arg\min_{\mathcal{S}} \mathcal{L}(\mathcal{S}, \mathcal{T}), \tag{1}$$

where $\mathcal{L}$ can follow various forms according to the particular designs in different distillation approaches. As described in previous sections, there are three major categories of dataset distillation methods where in this paper we adopt one representative approach per category to conduct our experiments and studies, i.e. IDC [9] for gradient matching, MTT [1] for trajectory matching, and IDM [23] for distribution matching, respectively. Here we briefly review their core designs for building $\mathcal{L}$.

IDC [9] aims to match the gradients produced by real data or distilled synthetic data with respect to the model parameters. Basically, given a model ϕ parameterized by θ which has been trained on the original dataset for a limited number of epochs, with denoting the real data samples and synthetic data samples of the same class c as $Q_c^{\mathcal{T}}$ and $Q_c^{\mathcal{S}}$ respectively, IDC realizes the gradient matching via the following objective:

$$\mathcal{L}_{\text{IDC}} = \sum_{c=1}^{C} \Big(\Big(\frac{1}{|Q_c^{\mathcal{T}}|} \sum_{(x_i, y_i) \in Q_c^{\mathcal{T}}} \nabla_\theta \ell(\phi_\theta(x_i), y_i) - \frac{1}{|Q_c^{\mathcal{S}}|} \sum_{(x'_i, y'_i) \in Q_c^{\mathcal{S}}} \nabla_\theta \ell(\phi_\theta(x'_i), y'_i) \Big)^2 \Big)^{\frac{1}{2}}$$

where C denotes the number of classes and ℓ denotes the cross-entropy loss function.

MTT [1] aims to match the long-range trajectories of updating model parameters between using real data or distilled synthetic data to train the model. Basically, with the same initial model parameters θ, the model variant which is trained upon synthetic data for N steps (resulting to have model parameter $\hat{\theta}_N$) is encouraged to match another variant which is trained upon real data for M steps (resulting to have model parameter $\tilde{\theta}_M$) in terms of parameters, where $N \ll M$. The objective is formulated as:

$$\mathcal{L}_{\text{MTT}} = \frac{\|\hat{\theta}_N - \tilde{\theta}_M\|}{\|\theta - \tilde{\theta}_M\|}$$

where the denominator is independent to the distilled data and serves as a normalized term.

IDM [23] aims to match the feature distribution of real dataset and the one of the distilled synthetic data (where the feature is extracted by a model ϕ parameterized by θ which has been trained on the original dataset) in which the metric for measuring the distance between two distributions is maximum mean discrepancy (MMD [6]). Moreover, IDM has an additional regularization term for encouraging the distilled synthetic data to have minimal classification error. The overall objective for IDM is hence formulated as:

$$\mathcal{L}_{\text{IDM}} = \sum_{c=1}^{C} \| \frac{1}{|Q_c^{\mathcal{T}}|} \sum_{x_i \in Q_c^{\mathcal{T}}} \phi_\theta(x_i) - \frac{1}{|Q_c^{\mathcal{S}}|} \sum_{x_i' \in Q_c^{\mathcal{S}}} \phi_\theta(x_i') \| + \lambda \sum_{(x_i', y_i') \in \mathcal{S}} \ell(\phi_\theta(x_i'), y_i')$$

where λ is to balance the distribution matching term and the regularization term.

3.2 Dataset Distillation with Multiple Experts

As motivated previously, the existing dataset distillation methods often suffers from poor cross-architecture generalization (i.e. when the distilled dataset is adopted to train a model that has different network architecture from the one used in the distillation process, the performance of the resultant model degrades significantly) and we attribute such problem to their relying on a "single" distillation process for encapsulating the entire original dataset, where the distilled dataset might not have sufficient diversity to support various architectures. To this end, we propose to leverage the mixture-of-experts idea for having multiple distillation processes (as experts) in a unified framework and particularly encourage them to capture distinct characteristics of the original data thus enriching the diversity of the resultant distilled dataset as well as improving the cross-architecture generalizability.

In detail, we introduce K experts $\mathbb{E} = \{\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_K\}$, where each expert $\mathcal{E}_i$ is tasked to produce a specific subset $\mathcal{S}_i$ of synthetic dataset $\mathcal{S}$, derived from its corresponding subset $\mathcal{T}_i$ of real data $\mathcal{T}$. Noting that $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_K$ and $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2 \cup \cdots \cup \mathcal{T}_K$ with setting $\mathcal{T}_i \bigcap \mathcal{T}_j = \varnothing, \ \forall i \neq j$ (for simplicity, we assume $|\mathcal{T}|$ is divisible by K). Particularly, in addition to having each expert independently perform dataset distillation on its assigned subset, we advance to introduce a loss function based on distance correlation [14, 24] which minimizes the dependency between the feature representations of synthetic subsets distilled by different experts, thus preventing the experts from converging towards the similar synthetic subsets (in terms of feature representations) as well as prompting the diversity in the distilled dataset. Noting that the distance correlation was proposed by [14] almost two decades ago but is recently revisited by [24] to highlight its advantages and general uses in deep learning as a measure for comparing two feature representations/spaces or the functional behavior between two networks. In implementation, once every few iterations during the distillation process, we randomly select pairs of synthetic subsets (e.g. $\mathcal{S}_i$ and $\mathcal{S}_j$ distilled by two distinct experts $\mathcal{E}_i$ and $\mathcal{E}_j$ respectively) and for each pair we compute its distance correlation. For a pair of synthetic subsets ($\mathcal{S}_i$ and $\mathcal{S}_j$), our loss based on distance correlation is written as:

$$\mathcal{L}_{Corr} = \mathtt{dCor}^2(\phi(\mathcal{S}_i), \phi(\mathcal{S}_j)) \tag{2}$$

where ϕ is a model which has been pretrained on real data and used for feature extraction, and $\mathtt{dCor}^2(\cdot,\cdot)$ is the squared sample distance covariance [24].

The overall objective combines the individual distillation losses of each expert (e.g. $\mathcal{L}_{\text{IDC}}$, $\mathcal{L}_{\text{MTT}}$, or $\mathcal{L}_{\text{IDM}}$, depending on which distillation methods are used to construct our experts) and the loss based on distance correlation:

$$\mathcal{L} = \sum_{i=1}^{K} \mathcal{L}_{DD}(\mathcal{S}_i; \mathcal{T}_i) + \sum_{i=1}^{K} \sum_{j \neq i}^{K} \mathcal{L}_{Corr}(\mathcal{S}_i, \mathcal{S}_j; \phi), \tag{3}$$

where $\mathcal{L}_{DD}(\mathcal{S}_i;\mathcal{T}_i)$ represents the distillation loss (which could be $\mathcal{L}_{\mathrm{MTT}}$, $\mathcal{L}_{\mathrm{IDC}}$ or $\mathcal{L}_{\mathrm{IDM}}$) according to the distillation method used for the expert $\mathcal{E}_i$, operating on its assigned real data subset $\mathcal{T}_i$ and generating the synthetic subset $\mathcal{S}_i$.

3.3 Mixup-based Fusion

While the distilled images can be directly utilized for downstream tasks, this approach may compromise overall accuracy due to the potential for specialized representations learned by individual expert models. Although minimizing distance correlation promotes diversity within the distilled dataset, it may inadvertently lead to experts focusing on limited aspects of the original data distribution. To address this challenge and encourage a more holistic representation, we incorporate a fusion strategy inspired by Mixup [19]. During the evaluation, we combine synthetic images from different expert subsets using a weighted sum:

$$\hat{x}' = \lambda x'_p + (1-\lambda)x'_q \tag{4}$$

where $x'_p \in S_i, x'_q \in S_j$, $p = 1,\ldots,|S_i|, q = 1,\ldots,|S_j|$ and $\forall i \neq j$. Here, λ is a scalar value sampled from a Beta distribution, following the principles outlined in the original Mixup [19] approach. This Mixup-based fusion encourages the model to learn from the combined knowledge of different experts, effectively leveraging their specialized representations to create a more generalizable and comprehensive representation of the original data. This approach aims to improve both the accuracy of downstream tasks and the model's ability to adapt to diverse network architectures.

4 Experimental Results

4.1 Experimental Setup

Datasets. To comprehensively evaluate our multi-expert DD framework, we employed a diverse set of datasets with varying image resolutions and complexities. Our benchmark for low-resolution datasets was CIFAR-10/100 [10], each containing 60,000 color images (32×32) across 10 and 100 classes respectively. For higher resolutions, we utilized ImageNette, a 10-class subset of ImageNet [4], ImageWoof, focusing on 10 dog breeds, and STL-10 [2], a 10-class dataset with labeled and unlabeled images. By evaluating our framework across these datasets (all resized to 64×64), we demonstrate its efficacy and robustness for addressing the challenge of cross-architecture generalizability in dataset distillation task.

Baselines. To showcase the versatility of our proposed framework, we integrate it with three prominent surrogate objectives for dataset distillation: IDC [9], IDM [23], and MTT [1], representing gradient matching, distribution matching, and trajectory matching, respectively. We conduct a comparative analysis of these distillation methods under equivalent storage budget, which means the same images-per-class times the number-of-experts (IPC $\times$ NoE), constraining and reporting their respective cross-architecture performance.

4.2 Comparison between Single Expert and Multiple Experts

We conducted an experiment to evaluate the impact of multiple experts using three different distillation methods: IDC, IDM, and MTT, under two storage budgets: 10 and 20. To compare the effects of multiple experts, we established two scenarios for each storage budget:

Table 1: Distillation and cross-architecture performance on CIFAR-10/100 datasets.

Target	CIFAR-10				CIFAR-100			
Budget	10		20		10		20	
IPC×NoE	5×2	10×1	10×2	20×1	5×2	10×1	10×2	20×1
IDC [9]								
ConvNet-3	**53.32**	52.73	**55.52**	54.05	**31.44**	31.03	**36.03**	35.33
VGG11	**51.41**	49.70	**53.37**	51.74	**29.62**	27.38	**34.53**	31.25
ResNet18	**51.56**	47.51	**54.06**	50.06	**30.74**	28.59	**36.08**	34.09
AlexNet	**46.34**	40.69	**52.01**	44.04	**26.45**	20.05	**33.84**	25.76
IDM [23]								
ConvNet-3	**49.04**	47.61	**52.31**	50.48	26.87	**27.45**	**32.38**	31.67
VGG11	44.53	**44.65**	**48.31**	47.70	**23.06**	22.90	28.57	**28.84**
ResNet18	**45.05**	42.84	**49.52**	46.02	**24.29**	25.35	**32.08**	32.04
AlexNet	**37.19**	34.96	**44.32**	41.71	**19.49**	17.07	**27.16**	22.59
MTT [1]								
ConvNet-3	53.82	**56.17**	60.14	**60.82**	**30.00**	29.53	28.50	**29.80**
VGG11	**49.86**	46.59	**52.04**	47.27	**29.23**	27.99	**27.80**	27.72
ResNet18	**52.14**	46.17	**58.13**	53.94	**30.19**	27.81	**27.55**	26.05
AlexNet	**44.16**	32.97	**52.11**	50.01	**25.27**	23.88	**28.10**	27.67
Full Dataset	84.80				56.20			

allocating the entire budget to a single expert and equally dividing the budget between two experts. Specifically, for a storage budget of 10, we set the number of experts (NoE) to 1 with images per class (IPC) being 10, denoted by IPC $\times$ NoE = 10 $\times$ 1, and to 2 with IPC being 5 for each expert, denoted by IPC $\times$ NoE = 5 $\times$ 2. Similarly, for a storage budget of 20, we set IPC × NoE to 20 $\times$ 1 and 10 $\times$ 2. The distillation process was performed on ConvNet-3. After completing the distillation, we trained new networks, including ConvNet-3, VGG11, ResNet18, and AlexNet, on the distilled dataset to measure the performance of the distillation methods. Table 1 presents the results on CIFAR-10 and CIFAR-100. We observed consistent improvements in multi-experts (i.e. NoE equals to 2) built by IDC over single-expert (i.e. NoE equals to 1) baselines, and most of the multi-expert results built by IDM and MTT outperformed the single-expert baselines. This suggests that the distilled datasets generated by our multi-expert framework are more generalizable for new architectures.

4.3 The Comparison to Single Expert SOTA

We conducted an experiment comparing our framework with a generalization-focused method, namely RDED [13]. To ensure a fair, same-budget (IPC=20) comparison: We created a hybrid MoE ("Ours + RDED"). One expert used IDM (IPC=10); the other directly incorporated 10 RDED-distilled images (IPC=10). We compared this hybrid to RDED distilling 20 images (RDED (IPC=20)). In Table 2, the hybrid MoE outperforms RDED (IPC=20) on ConvNet-3, VGG-11, and ResNet-18. RDED is slightly better on AlexNet. Overall, such preliminary hybrid MoE shows competitive and superior generalization to the RDED, offering an encouraging evidence of our framework's effectiveness. Moreover, our work is actually orthogonal to RDED as it can be simply integrated as an expert in our framework.

Table 2: RDED vs. Hybrid MoE Cross-Architecture Performance

	Ours	RDED
ConvNet-3	**50.28**	45.62
VGG-11	**46.03**	44.36
ResNet-18	**46.79**	45.54
AlexNet	41.94	**43.02**

Table 3: Ablation study on using different mixup strategies.

	Ours	Mixup	w/o Mixup
ConvNet-3	**53.32**	52.46	53.21
VGG11	**51.41**	51.34	50.84
ResNet18	**51.56**	50.20	50.11
AlexNet	**46.34**	44.60	42.98

4.4 Evaluating the Impact of Mixup-Based Fusion.

Table 3 presents an ablation study investigating the efficacy of our proposed mixup-based fusion in comparison to the vanilla mixup technique [19] within the context of dataset distillation using IDC [9] as the surrogate method. The experiment is built using IDC on CIFAR-10 with IPC$\times$NoE = 5×2. The results demonstrate that our strategy ("Ours"), which mixes synthetic images from different expert models (c.f. Section 3.3), consistently outperforms both the baseline without mixup ("w/o Mixup") and the Vanilla Mixup ("Mixup") approach (mixing images from both the same expert and different experts) across all target architectures. This suggests that leveraging the diverse representations learned by multiple experts is crucial for enhancing the generalizability and performance of the distilled dataset. While vanilla mixup also exhibits some performance improvement over the baseline, it falls short of our proposed strategy, indicating that simply mixing images may not sufficiently capture the full range of information required for effective distillation.

4.5 Ablation Studies

Table 4 presents the findings of an ablation study designed to evaluate the individual contributions of two key components within our proposed framework: distance correlation minimization and mixup-based fusion. We analyze their impact on performance using CIFAR-10 and employ IDM [23], IDC [9] and MTT [1] as the surrogate distillation methods and varying IPC settings. Incorporating distance correlation generally leads to an improvement in performance across all three methods and IPC settings. This suggests that encouraging diversity among expert models through distance correlation helps capture a broader range of information from the original data, leading to better generalizability. However, the magnitude of the improvement varies depending on the specific distillation method and IPC level. The inclusion of the mixup-based fusion strategy improves performance across all distillation methods and IPC settings. This highlights the effectiveness of combining complementary information learned by different experts. Finally, the combination of distance correlation and mixup-based fusion results in the best overall performance using the MoE framework.

Table 4: Ablation study on our proposed distance correlation and fusion techniques on CIFAR-10 with using different dataset distillation methods to build experts

Distillation Method	IDM [23]		IDC [9]		MTT [1]	
IPC$\times$NoE	5$\times$2	10$\times$2	5$\times$2	10$\times$2	5$\times$2	10$\times$2
Baseline	46.35	49.50	52.28	53.47	51.02	58.08
+ Distance Correlation	47.55	51.81	52.90	55.01	51.59	58.38
+ Distance Correlation & Mixup-based Fusion	49.04	52.31	53.32	55.52	53.82	60.14

5 Conclusion

This work has presented a novel approach to DD that addresses the challenge of cross-architecture performance degradation by promoting diversity and generalizability within the distilled dataset. Our framework leverages multiple expert models, each specializing in distilling a distinct subset of the data, and incorporates distance correlation minimization as well as image fusion strategies to enhance the richness and informativeness of the distilled data. Through extensive experiments, we have demonstrated the effectiveness of our approach in mitigating the transferability issue and achieving improved performance across various target architectures, particularly in low-data regimes. Our findings highlight the importance of diversity in DD and provide valuable insights for future research in this area.

Acknowledgement

This work is sponsored by Macronix International Co., Ltd. (MXIC) and funded by the National Science and Technology Council (NSTC) under Project Numbers 112-2221-E-A49-087-MY3 and 111-2628-E-A49-018-MY4.

References

[1] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

[3] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Jiawei Du, Yidi Jiang, Vincent T. F. Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3749–3758, 2023.

[6] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, March 2012. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2503308.2188410`.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pages 1339–1344 vol.2, 1993. doi: 10.1109/IJCNN.1993.716791.

[9] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning (ICML)*, pages 11102–11118, 2022.

[10] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[13] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. *arXiv preprint arXiv:2312.03526*, 2023.

[14] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

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

[16] Felix Wiewel and Bin Yang. Condensed composite memory continual learning, 2021.

[17] Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. Feddm: Iterative distribution matching for communication-efficient federated learning, 2022.

[18] Enneng Yang, Li Shen, Zhenyi Wang, Tongliang Liu, and Guibing Guo. An efficient dataset condensation plugin and its application to continual learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Murj6wcjRw.

[19] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

[20] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning (ICML)*, pages 12674–12685, 2021.

[21] Bo Zhao and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations (ICLR)*, 2021.

[22] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 6514–6523, 2023.

[23] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7856–7865, 2023.

[24] Xingjian Zhen, Zihang Meng, Rudrasis Chakraborty, and Vikas Singh. On the versatile uses of partial distance correlation in deep learning. In *Proceedings of the European conference on computer vision (ECCV)*, 2022.

[25] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.