FADRM: Fast and Accurate Data Residual Matching for Dataset Distillation

Jiacheng Cui *, Xinyue Bi *, Yaxin Luo, Xiaohan Zhao, Jiacheng Liu, Zhiqiang Shen †

*: *Equal Contribution*; †: *Corresponding Author*

Problem Statement

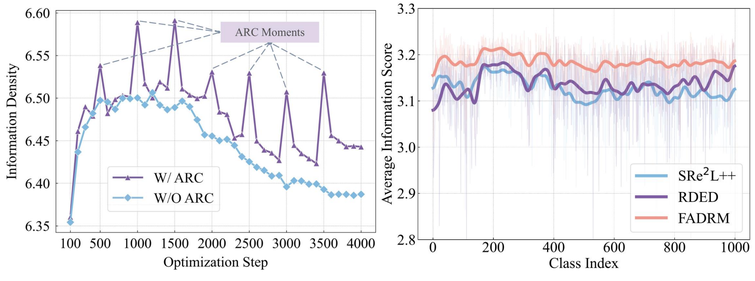
Dataset Distillation (DD) aims to synthesize a compact dataset $\mathcal{C}$ that allows a model trained on it to perform comparably to one trained on the full dataset $\mathcal{O}$. The objective is to *minimize the performance gap* between the two models:

$$\underset{\mathcal{C},|\mathcal{C}|}{\operatorname{argmin}} \sup_{(x,y)\sim\mathcal{O}} |L(f_{\theta_\mathcal{O}}(x), y) - L(f_{\theta_\mathcal{C}}(x), y)|$$

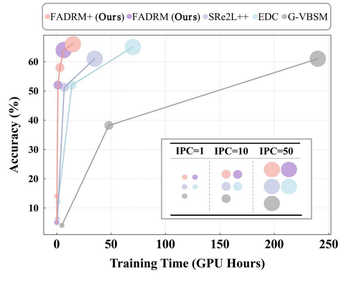
The key challenge is to *preserve essential information* and *generalization ability* while drastically reducing data size and computational cost.

Motivation

1) ***Information Vanishing:*** As optimization progresses, information density first increases but later declines due to local feature loss, causing information vanishing that reduces image fidelity and downstream performance.



2) ***High Computational Overhead:*** Large-scale data synthesis demands high computation, with EDC taking ~70 hours for a 50-IPC dataset, limiting large-scale or repeated experiments.



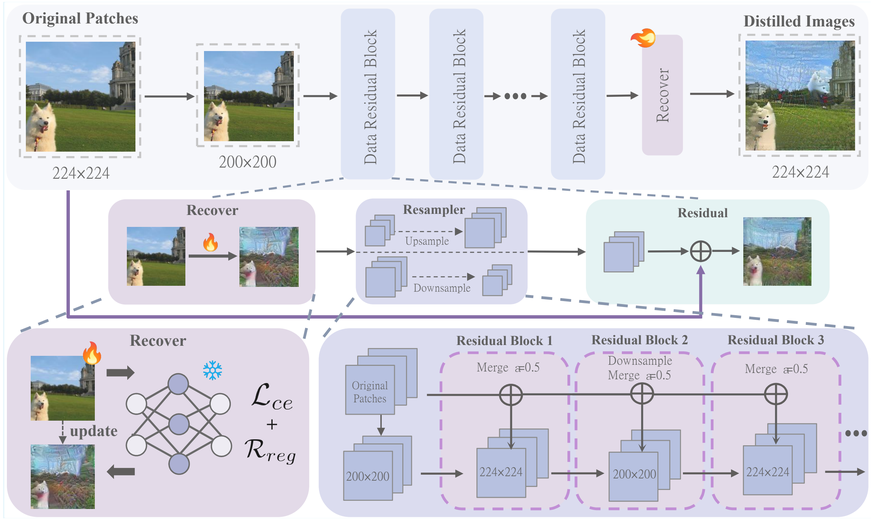
Method	Time Cost (s)	Peak Memory (GB)
SRe^{2}L++ [6]	2.52	5.3
FADRM	**0.47**	**2.9**
G-VBSM [30]	17.28	21.4
CV-DD [6]	8.20	23.4
EDC [31]	4.99	17.9
FADRM+	**1.09**	**11.0**

Methodology

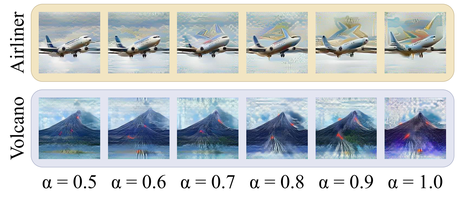
Our method is based on three components:

1) Adjustable Residual Connection – ***mitigating Information Vanishing.***
2) Muti-resolution Optimization
3) Mixed-Precision Training

} ***improve generation efficiency.***



In ARC, a large merge ratio favors optimized patches; a small one favors original patches.



Experimental Results

Dataset	IPC (Ratio)	ResNet18					ResNet50					ResNet101				
		RDED	EDC	CV-DD	FADRM	**FADRM+**	RDED	EDC	CV-DD	FADRM	**FADRM+**	RDED	EDC	CV-DD	FADRM	**FADRM+**
CIFAR-100	1 (0.2%)	17.1	39.7	28.3	31.8	**40.6**	10.9	36.1	28.7	27.3	**37.4**	11.2	32.3	29.0	29.2	**40.1**
	10 (2.0%)	56.9	63.7	62.7	67.4	**67.9**	41.6	62.1	61.5	66.5	**67.4**	54.1	61.7	63.8	68.3	**68.9**
	50 (10.0%)	66.8	68.6	67.1	71.0	**71.3**	64.0	69.4	68.2	71.5	**72.1**	67.9	68.5	67.6	71.9	**72.1**
	Whole Dataset	78.9					79.9					79.5				
Tiny-ImageNet	1 (0.2%)	11.8	39.2	30.6	28.6	**40.4**	8.2	35.9	25.1	28.4	**39.4**	9.6	40.6	28.0	27.9	**41.9**
	10 (2.0%)	41.9	51.2	47.8	48.9	**52.8**	38.4	50.2	43.8	47.3	**53.7**	22.9	51.6	47.4	47.8	**53.6**
	50 (10.0%)	58.2	57.2	54.1	56.4	**58.7**	45.6	58.8	54.7	57.0	**60.3**	41.2	58.6	54.1	57.2	**60.8**
	Whole Dataset	68.9					71.5					70.6				
ImageNette	1 (0.1%)	35.8	-	36.2	36.2	**39.2**	27.0	-	27.6	31.1	**31.9**	25.1	-	25.3	26.3	**29.3**
	10 (1.0%)	61.4	-	64.1	64.8	**69.0**	55.0	-	61.4	64.1	**68.1**	54.0	-	61.0	61.9	**63.7**
	50 (5.2%)	80.4	-	81.6	83.6	**84.6**	81.8	-	82.0	84.1	**85.4**	75.0	-	80.0	80.3	**82.3**
	Whole Dataset	93.8					89.8					89.3				
ImageWoof	1 (0.1%)	20.8	-	21.4	21.0	**22.8**	17.8	-	19.1	19.5	**19.9**	19.6	-	19.9	20.0	**21.8**
	10 (1.1%)	38.5	-	49.3	44.5	**57.3**	35.2	-	47.8	44.9	**54.1**	31.3	-	42.6	40.4	**51.4**
	50 (5.3%)	68.5	-	71.9	72.3	**72.6**	67.0	-	71.2	71.0	**71.7**	59.1	-	69.9	70.3	**70.6**
	Whole Dataset	88.2					77.8					82.7				
ImageNet-1k	1 (0.1%)	6.6	12.8	9.2	9.0	**14.7**	8.0	13.3	10.0	12.2	**16.2**	5.9	12.2	7.0	6.8	**14.1**
	10 (0.8%)	42.0	48.6	46.0	48.4	**50.9**	49.7	54.1	51.3	54.5	**57.5**	48.3	51.7	51.7	54.8	**58.1**
	50 (3.9%)	56.5	58.0	59.5	60.1	**61.2**	62.0	64.3	63.9	65.4	**66.9**	61.2	64.9	62.7	66.0	**67.0**
	Whole Dataset	72.3					78.6					79.8				

Theoretical Guarantee

Theorem 2 (Proof in Appendix A.3.1). *Let $\mathcal{H}$ be a class of functions $h : \mathbb{R}^d \to \mathbb{R}$, and let h be Lipschitz-continuous with constant $L_h > 0$, and the loss function ℓ be Lipschitz-continuous with constant $L_l > 0$ and bounded within a finite range $[0, B]$. Consider: 1. Optimized perturbation added to the original data: $\tilde{\mathcal{C}}^{res} = \{\tilde{x}_i^{res}, \tilde{y}_i^{res}\}_{i=1}^n$. 2. residual injected dataset (FADRM): $\tilde{\mathcal{C}}_{\text{FADRM}} = \{\tilde{x}_i, \tilde{y}_i\}_{i=1}^n$. 3. patches selected from the original dataset: $\mathcal{O} = \{x_i, y_i\}_{i=1}^n$. 4. discrepancy $\Delta := \frac{1}{n}\sum_{i=1}^n \|\tilde{x}_i^{res} - x_i\|$. Let $h_{\text{res}} \in \mathcal{H}$ denote the hypothesis trained on $\tilde{\mathcal{C}}^{res}$, and $h_{\text{FADRM}} \in \mathcal{H}$ be trained on $\tilde{\mathcal{C}}_{\text{FADRM}}$. Define the corresponding empirical risks: $\widehat{\mathcal{L}}_{\text{res}} := \frac{1}{n}\sum_{i=1}^n \ell(h_{\text{res}}(\tilde{x}_i^{res}), \tilde{y}_i^{res})$, $\widehat{\mathcal{L}}_{\text{FADRM}} := \frac{1}{n}\sum_{i=1}^n \ell(h_{\text{FADRM}}(\tilde{x}_i), \tilde{y}_i)$. Assume:*

$$\mathfrak{R}_n(\mathcal{H} \circ \mathcal{O}) < \mathfrak{R}_n(\mathcal{H} \circ \tilde{\mathcal{C}}^{res})$$

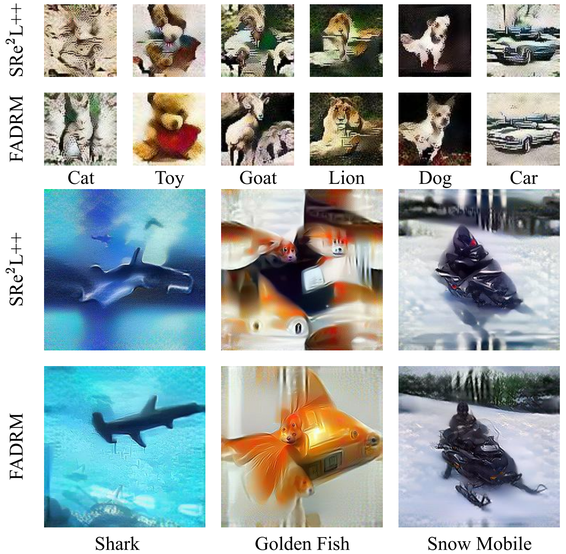
Then the generalization bound of h_{FADRM} is rigorously shown to be tighter than that of h_{res}, i.e.,

$$\widehat{\mathcal{L}}_{\text{FADRM}} + 2B \cdot \mathfrak{R}_n(\mathcal{H} \circ \tilde{\mathcal{C}}_{\text{FADRM}}) < \widehat{\mathcal{L}}_{\text{res}} + 2B \cdot \mathfrak{R}_n(\mathcal{H} \circ \tilde{\mathcal{C}}^{res}).$$

Cross-Architecture Generalization Analysis

Model	#Params	RDED	EDC	CV-DD	**FADRM+**
ResNet18 [11]	11.7M	42.0	48.6	46.0	**50.9**
ResNet50 [11]	25.6M	49.7	54.1	51.3	**57.5**
ResNet101 [11]	44.5M	48.3	51.7	51.7	**58.1**
EfficientNet-B0 [35]	39.6M	42.8	51.1	43.2	**51.9**
MobileNetV2 [28]	3.4M	34.4	45.0	39.0	**45.5**
ShuffleNetV2-0.5x [44]	1.4M	19.6	29.8	27.4	**30.2**
Swin-Tiny [22]	28.0M	29.2	38.3	–	**39.1**
Wide ResNet50-2 [11]	68.9M	50.0	–	53.9	**59.1**
DenseNet121 [13]	8.0M	49.4	–	50.9	**55.4**
DenseNet169 [13]	14.2M	50.9	–	53.6	**58.5**
DenseNet201 [13]	20.0M	49.0	–	54.8	**59.7**

Distilled Image Visualization



Paper is here

Code is here