INTERNATIONAL NEURAL NETWORK SOCIETY

Dataset Pruning &Distillation

---Streamlining Machine Learning Performance

Joey Tianyi Zhou

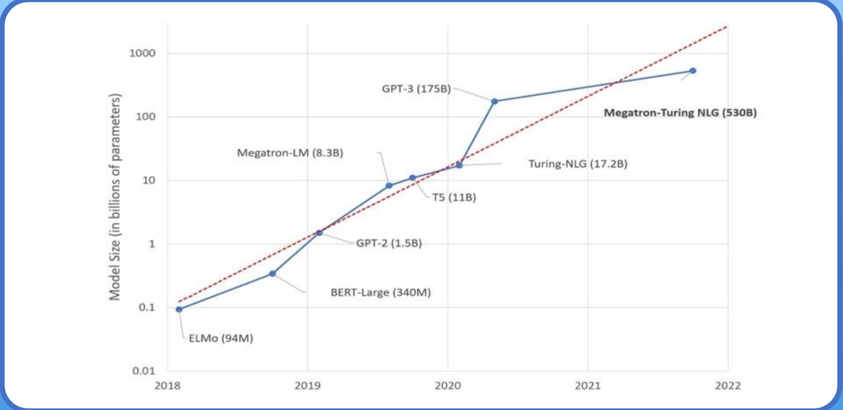
Deputy Director/Principal Scientist

A*STAR Centre for Frontier AI Research, Singapore

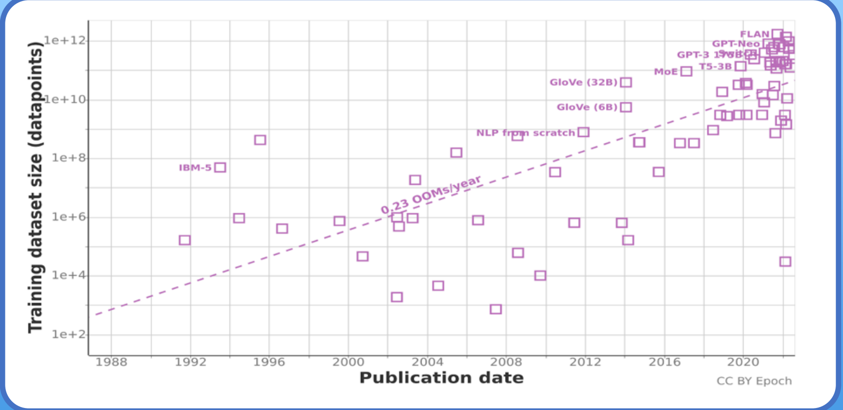
Why Dataset Distillation

Sustainable AI

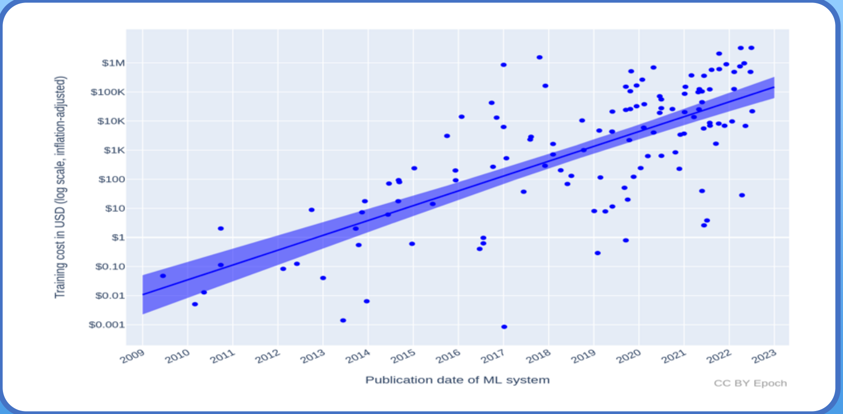
Increasing Model Size



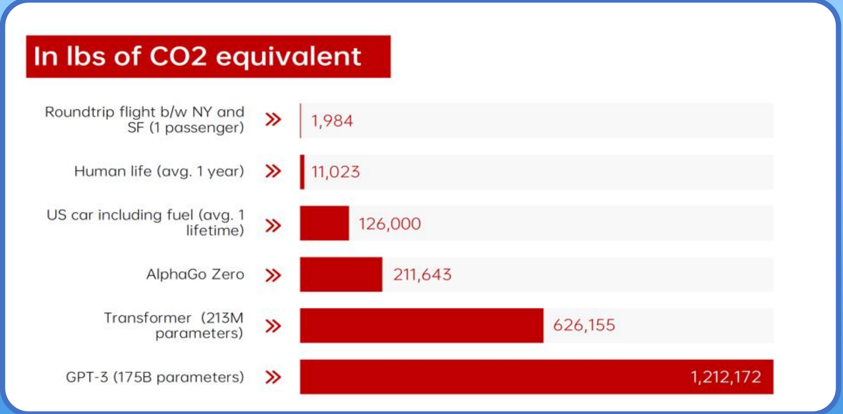
Massive Training Data



Exponential increasing cost

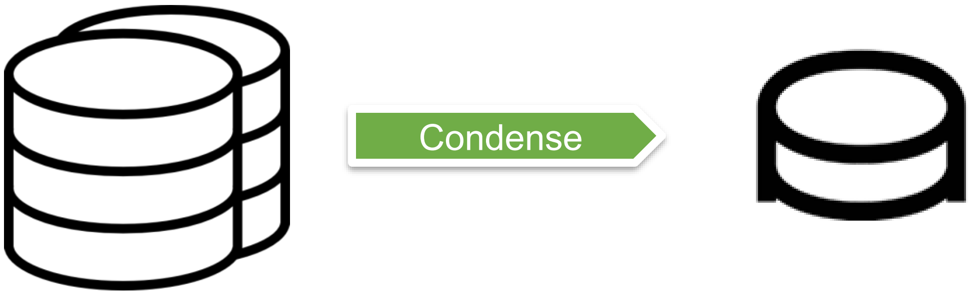


Exponential increasing CO2



CREATING GROWTH, ENHANCING LIVES

Dataset distillation:



A real dataset
large but redundant

50,000 Images

A synthetic dataset
small but compact

500 Images

By doing so

100%	Data Storage	1%
1X	Training Speed	≫10X
$100M	Evaluation Cost (Cost per year for HPE)	≪ $1M

LLMs evaluation Federate Learning Continual Learning

Outline

Dataset Pruning

Select a subset images in a full dataset without performance drop.

Dataset Distillation

Learn a few synthetic images (alter image pixels) to replace full dataset.

Outline

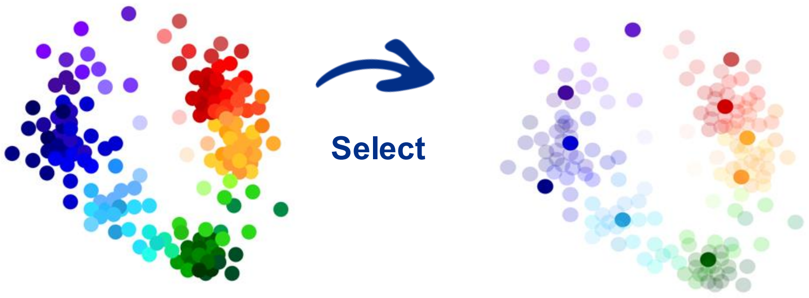
Dataset Pruning

Select a subset images in a full dataset without performance drop.

Dataset Distillation

Learn a few synthetic images (alter image pixels) to replace full dataset.

Dataset Pruning



Full dataset

Coreset

Also known as **Subset Selection**.

The Objective:

- **Select a subset images** in a full dataset witho ut performance drop.

Dataset Pruning

A Milestone Paper: Forgetting (ICLR, 2019)

- Some samples are consistently forgotten across subsequent training;
- Some examples are never forgotten.

Prune unforgettable ones

Forgetting statistics:

Sample i undergoes a forgetting event when it is correctly classified in the current update, but misclassified in the next update.

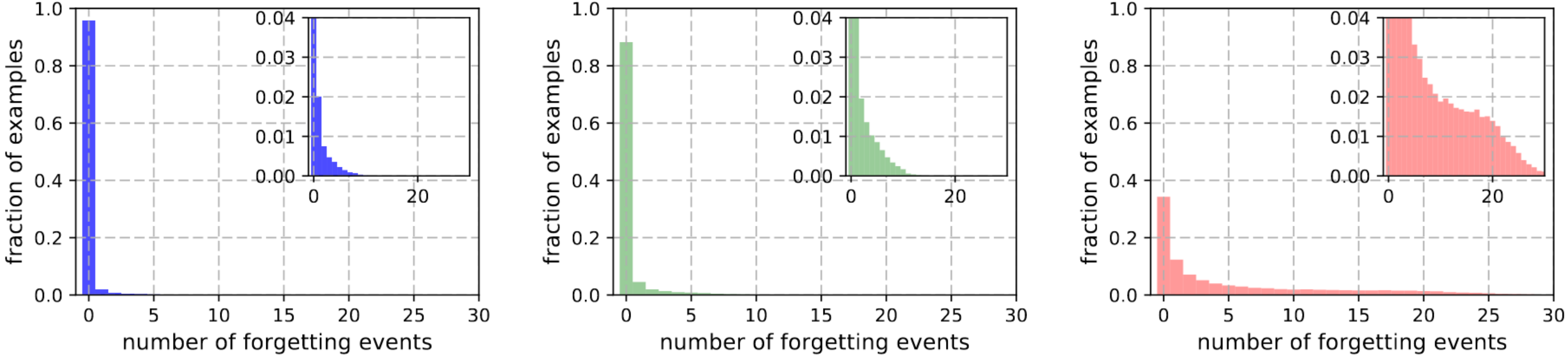
Algorithm 1 Computing forgetting statistics.

initialize prev_acc$_i = 0, i \in \mathcal{D}$
initialize forgetting $T[i] = 0, i \in \mathcal{D}$
while not training done **do**
  $B \sim \mathcal{D}$ # sample a minibatch
  for example $i \in B$ **do**
    compute acc$_i$
    if prev_acc$_i >$ acc$_i$ **then**
      $T[i] = T[i] + 1$
    prev_acc$_i =$ acc$_i$
  gradient update classifier on B
return T

T[i]: the number of forgetting events for samp
i.

Dataset Pruning

A Milestone Paper: Forgetting (ICLR, 2019)



Histograms of forgetting events on (from left to right) MNIST, permutedMNIST and CIFAR-10

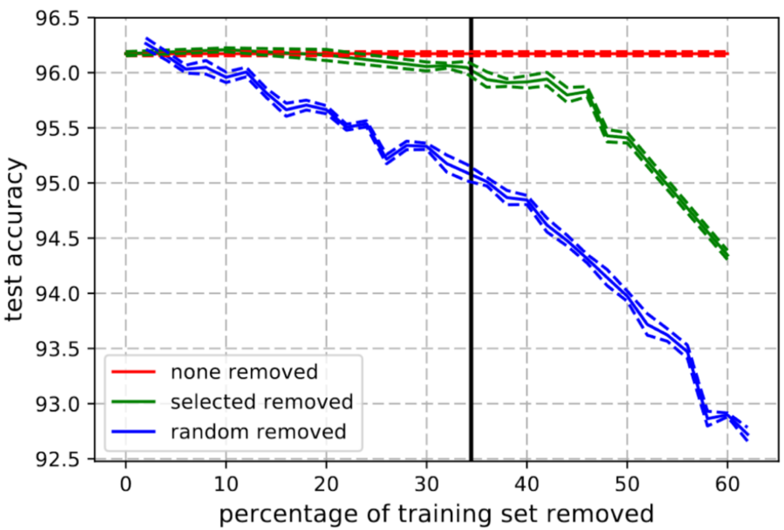
- **Most samples are unforgettable.**
- **More complex datasets contain significantly fewer unforgettable samples.**

CREATING GROWTH, ENHANCING LIVES

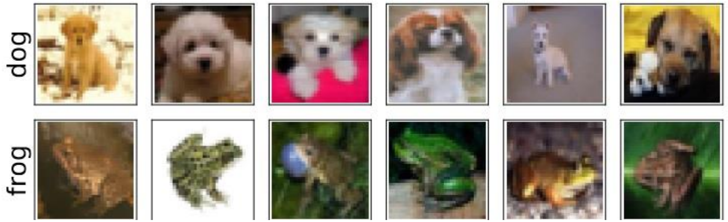
Dataset Pruning

A Milestone Paper: Forgetting (ICLR, 2019)

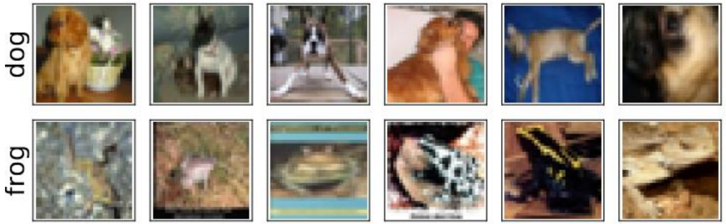
Performance on CIFAR-10 of ResNet18



The most unforgettable samples (CIFAR-10)



The most forgettable samples (CIFAR-10)



Forgettable samples seem to exhibit peculiar or uncommon features.

CREATING GROWTH, ENHANCING LIVES

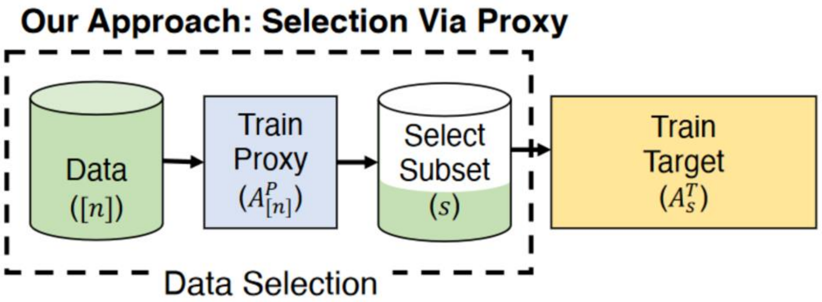
Dataset Pruning

Problem/Challenge of “Forgetting”:

The collection of “Forgetting” statistics for large models is time-consuming.

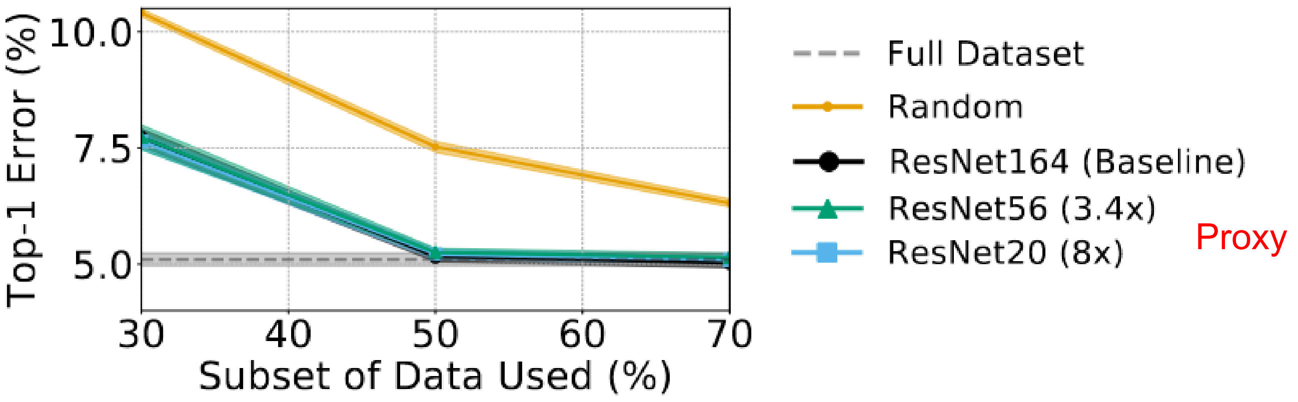
A Solution: Selection via Proxy (ICLR, 2020)

- Use a **smaller** proxy network $A^P_{[n]}$ with **fewer** epochs to speed up the training process.



Dataset Pruning

“Selection via Proxy” Performance



CIFAR10 forgetting events

Performance is on-par a large network.

CREATING GROWTH, ENHANCING LIVES

Dataset Pruning

Problem/Challenge of "Forgetting":

Collecting "Forgetting" statistics necessitates a complete training.

A Solution: EL2N (NeurIPS, 2021)

EL2N (New Metric) : The **early (less than 20 epochs) error vector score**, averaged over several weight initializations.

$$\mathbb{E}\|p(\mathbf{w}_t, x) - y\|_2.$$

logits vector

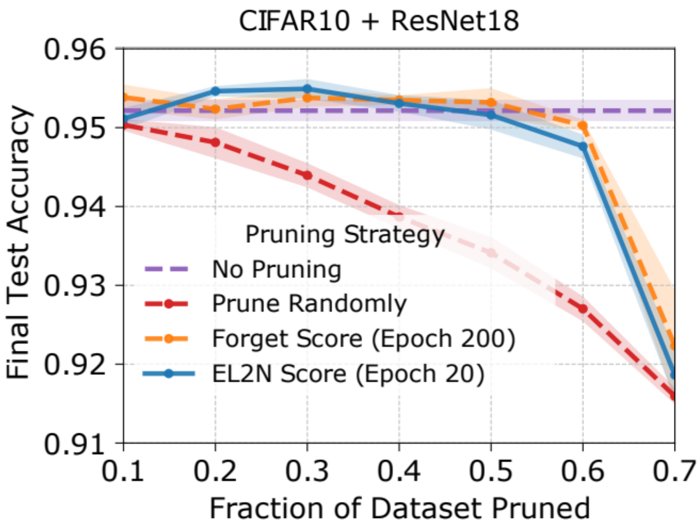
one-hot label

Prune samples with smaller error vector scores.

CREATING GROWTH, ENHANCING LIVES

Dataset Pruning: performance

EL2N Performance



Comparison of **forgetting scores at the end of training** and **EL2N scores early in training** (at epoch 20).

Dataset Pruning

Motivation: **Snapshot-based dataset pruning**

❖ Fluctuating Coreset Distributions.

Sample importance scores fluctuate with epochs during training, leading to significantly different coreset distributions at various training snapshots.

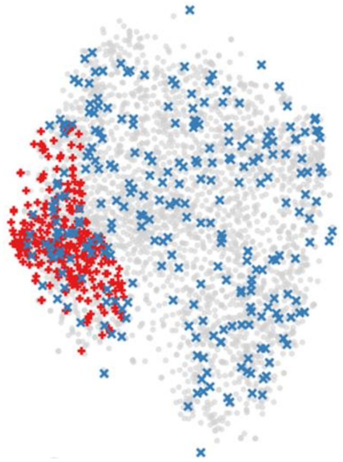


Figure 1. Distributions of coresets selected from Epoch 10 and 100.

Dataset Pruning

A Solution: Temporal Dual-Depth Scoring (TDDS), CVPR 2024.

- Problem Formulation

$$\mathbb{S}^* = \underset{\mathbb{S}\subset\mathbb{U}}{\arg\min} \|\boldsymbol{G}_{t,\mathbb{U}} - \tilde{\boldsymbol{G}}_{t,\mathbb{S}}\|, \quad \text{where} \quad \boldsymbol{G}_{t,\mathbb{U}} = \sum_{\substack{n=1,\\ \boldsymbol{x}_n\in\mathbb{U}}}^{|\mathbb{U}|} \boldsymbol{g}_t(\boldsymbol{x}_n), \quad \tilde{\boldsymbol{G}}_{t,\mathbb{S}} = \sum_{\substack{m=1,\\ \boldsymbol{x}_m\in\mathbb{S}}}^{|\mathbb{S}|} \boldsymbol{g}_t(\boldsymbol{x}_m),$$

+training dynamics

$$\mathbb{S}^* = \underset{\mathbb{S}\subset\mathbb{U}}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{\mathcal{G}}_{t,\mathbb{U}} - \tilde{\boldsymbol{\mathcal{G}}}_{t,\mathbb{S}}\|^2, \quad \text{where} \quad \boldsymbol{\mathcal{G}}_{t,\mathbb{U}} = [\|\boldsymbol{g}_t(\boldsymbol{x}_n)\|]_{n=1}^{N}$$

Dataset Pruning: performance

Spanning Training Progress: Temporal Dual-Depth Scoring (TDDS)

Experiments:

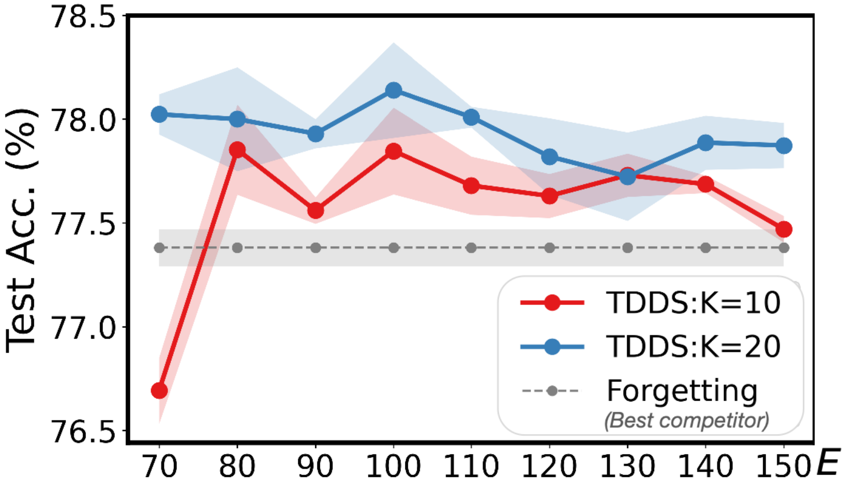
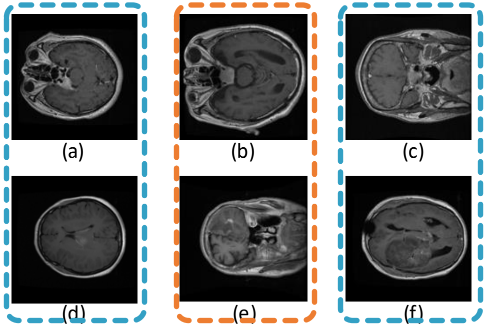


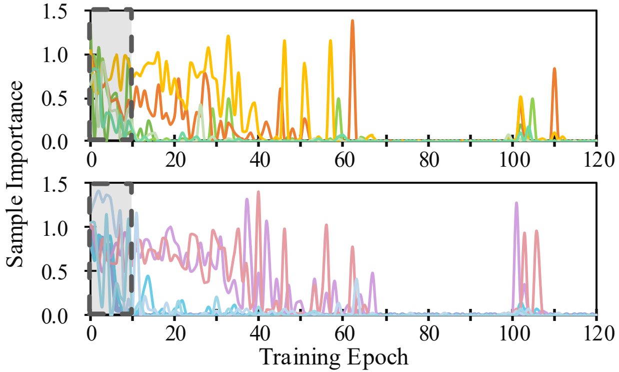
Figure 5. Parameter sensitivity study.

Dataset Pruning

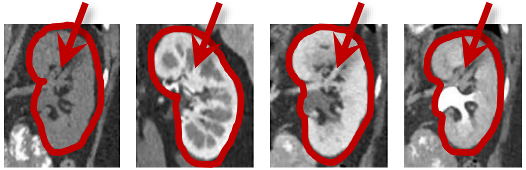
- Medical image datasets present fine-grained **intra-class variation** and **inter-class similarity**.

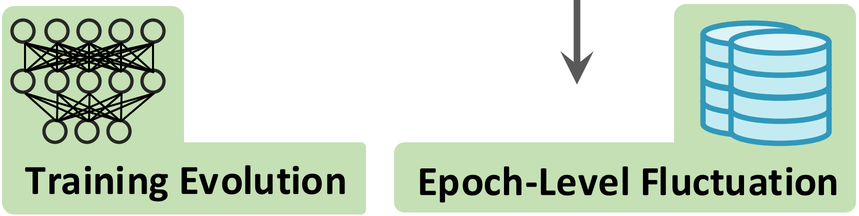


- The **importance of samples** in enhancing the model performance **varies** across different training epochs.



- Deep learning process exhibits **evolutionary nature** from simple to more complex stages.

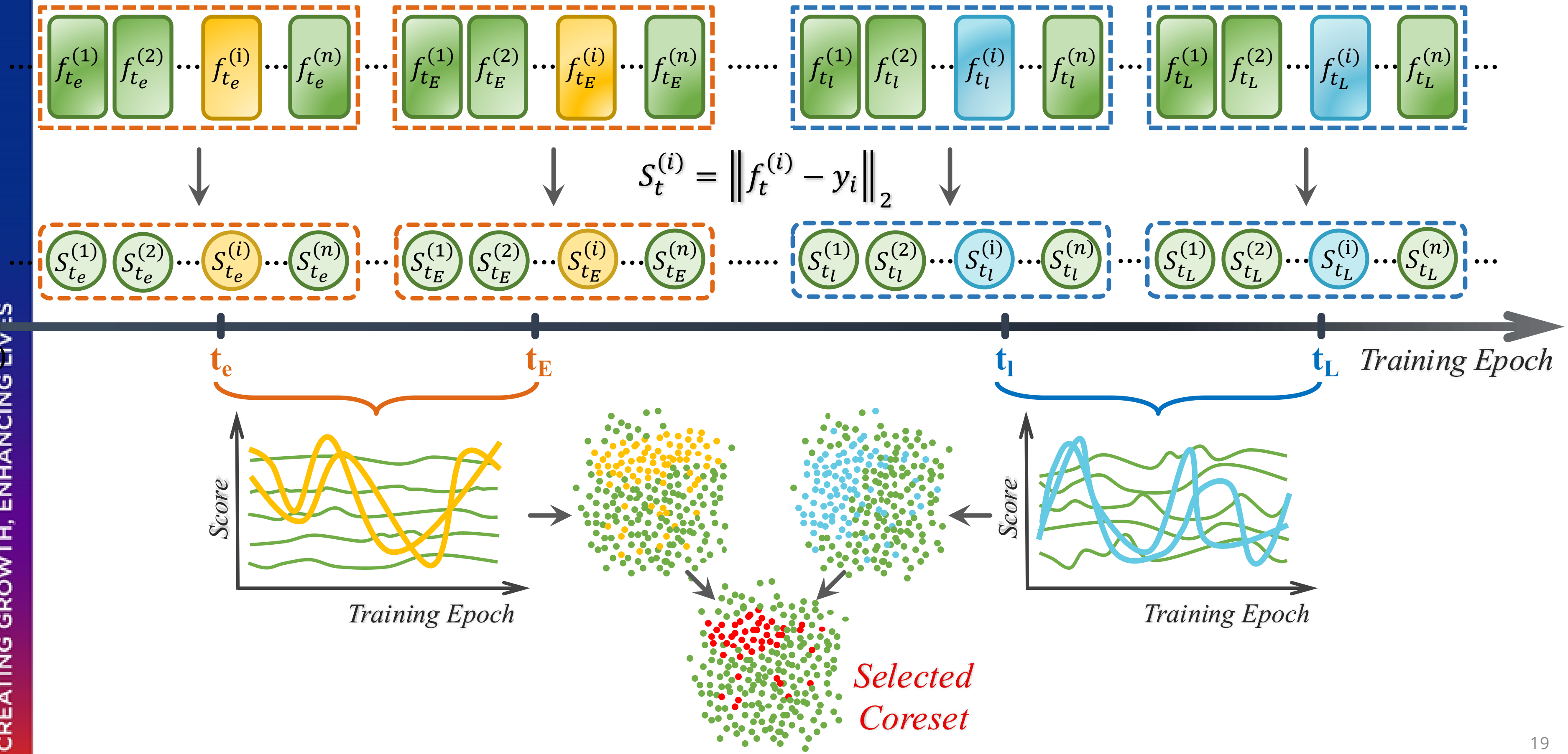




Evolution-aware VAriance (EVA)

"EVA: Evolution-aware VAriance Coreset Selection for Medical Image Classification", MM 2024 (Best Paper Nomination)

Capturing Training Evolution with Dual-Window



Outline

Dataset Pruning

Select a subset images in a full dataset without performance drop.

Dataset Distillation

Learn a few synthetic images (alter image pixels) to replace full dataset.

- Performance Matching
- Gradient Matching
- Distribution Matching
- Trajectory Matching
- Sequential Matching

Dataset Distillation

Compared with Dataset Pruning:

Dataset Distillation Objective:

- **Learn a few synthetic image s (alter image pixels**) to repla ce full dataset.

1 image per class (IPC)

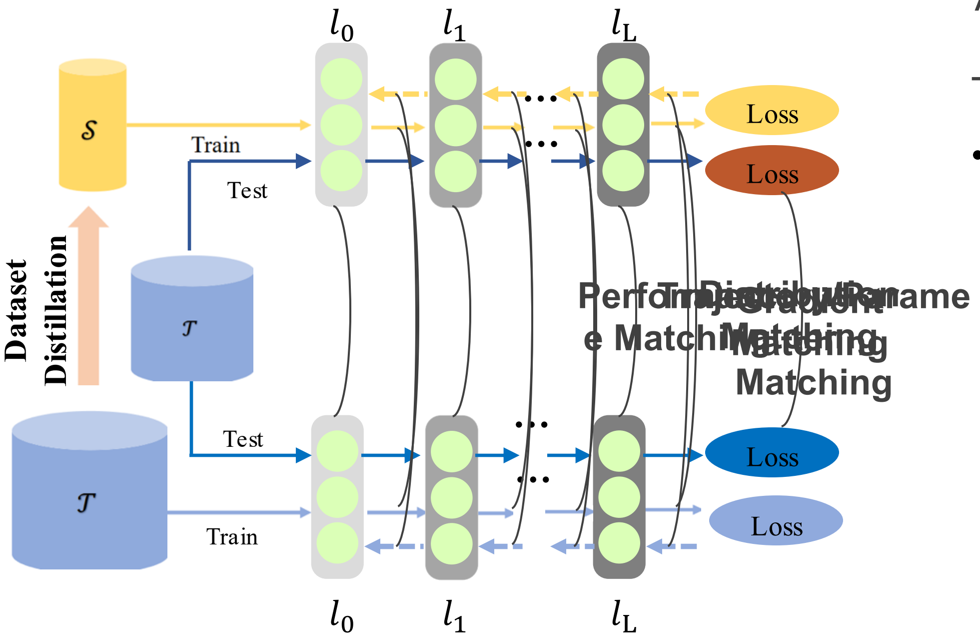
Dataset Pruning Objective:

- **Select a subset images** in a full dataset without performance drop.

Whether the image pixels are learnable

CREATING GROWTH, ENHANCING LIVES

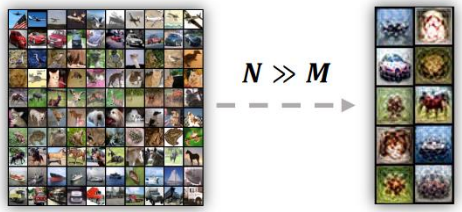
Dataset Distillation



Also known as **Dataset Condensation**.

The Objective:

- **Learn a few synthetic images (alter image pixels)** to replace full dataset.



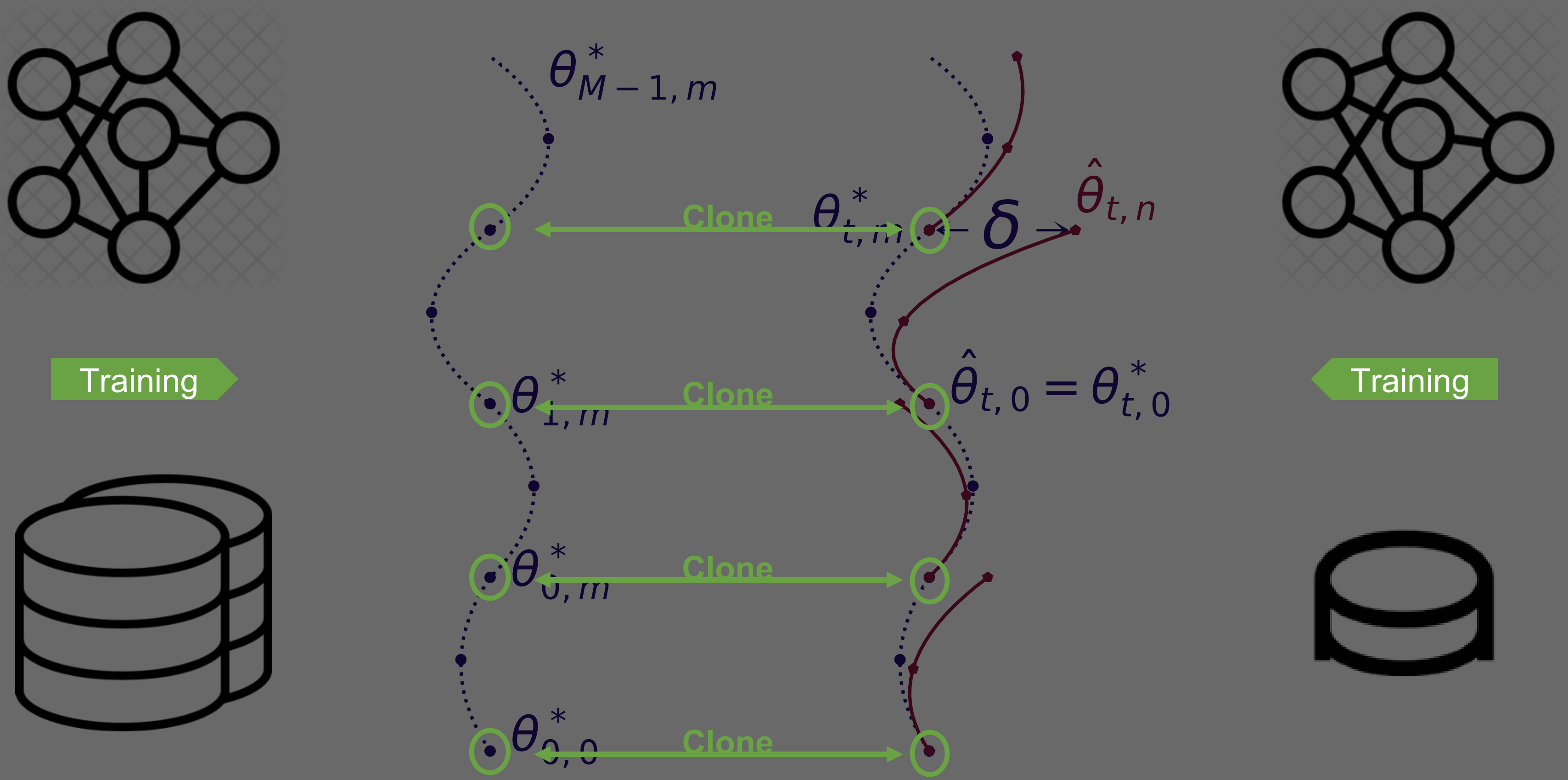
N images

M images

1 image per class (IPC)

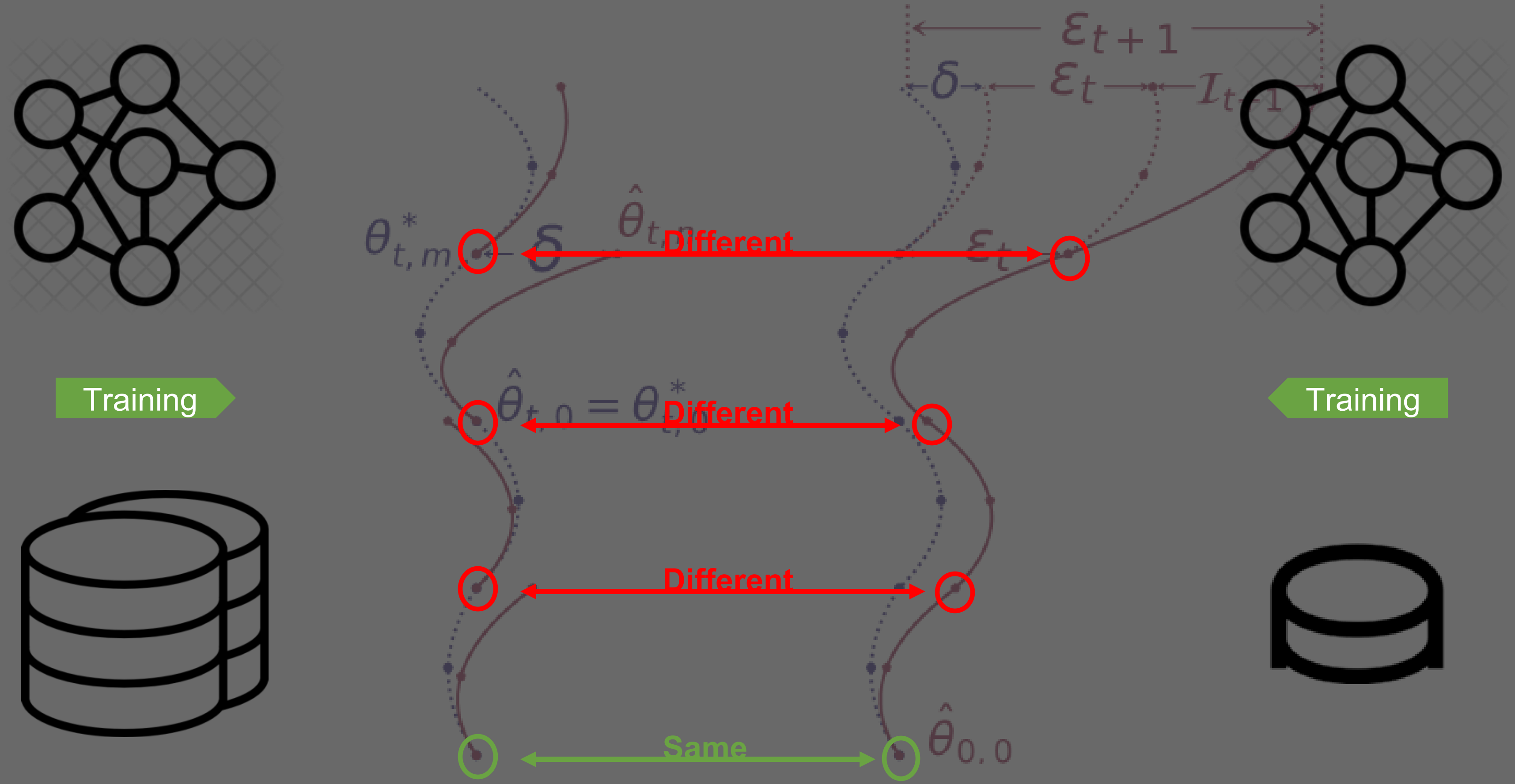
However, a trajectory error exists

In training, two trajectories start at the same points.



However, a trajectory error exists

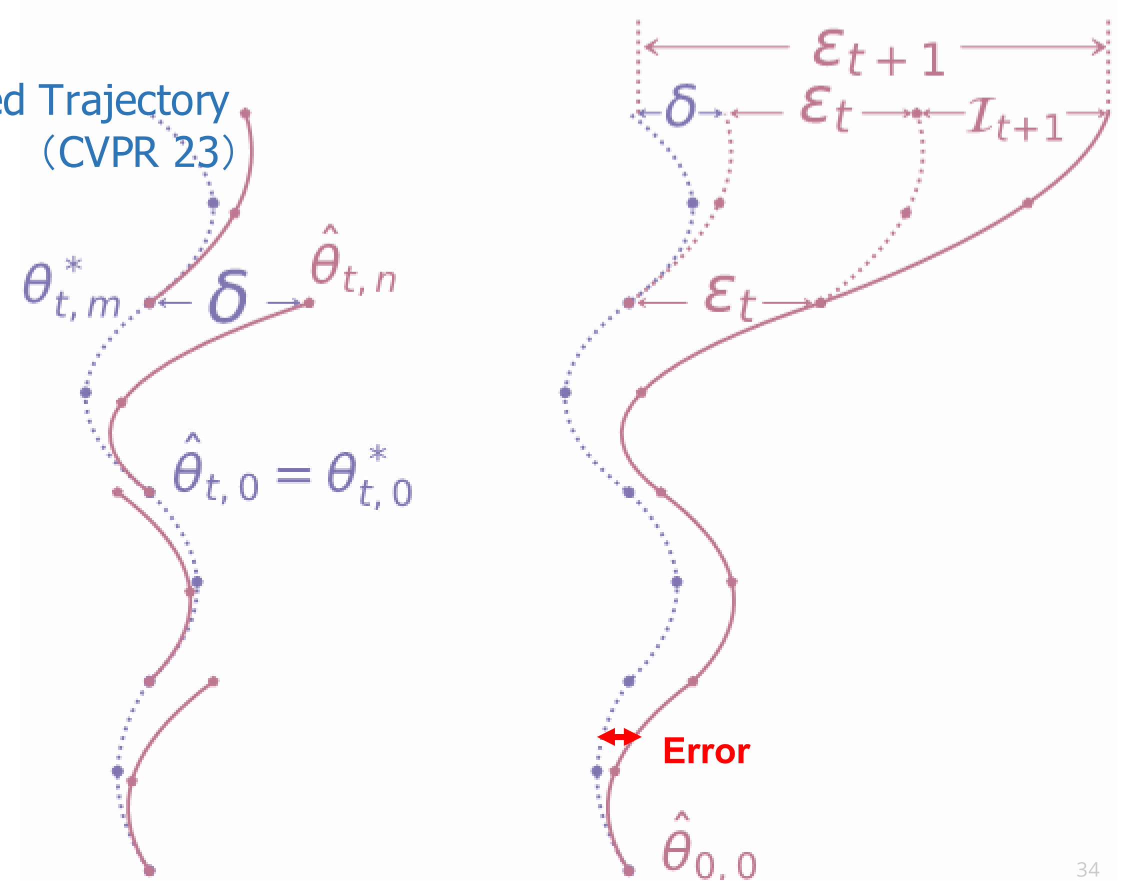
In testing, two trajectories start at the different points.



Trajectory error is accumulating

In testing, trajectories will not be vanishing, but accumulating.

Solution: Minimizing the Accumulated Trajectory Error to Improve Dataset Distillation （CVPR 23）



CREATING GROWTH, ENHANCING LIVES

Trajectory error is accumulating

In testing, trajectories will not be vanishing, but accumulating.

The gap between training and testing results in the accumulated trajectory error

We formulate it as below,

(1) Initial error: matching error

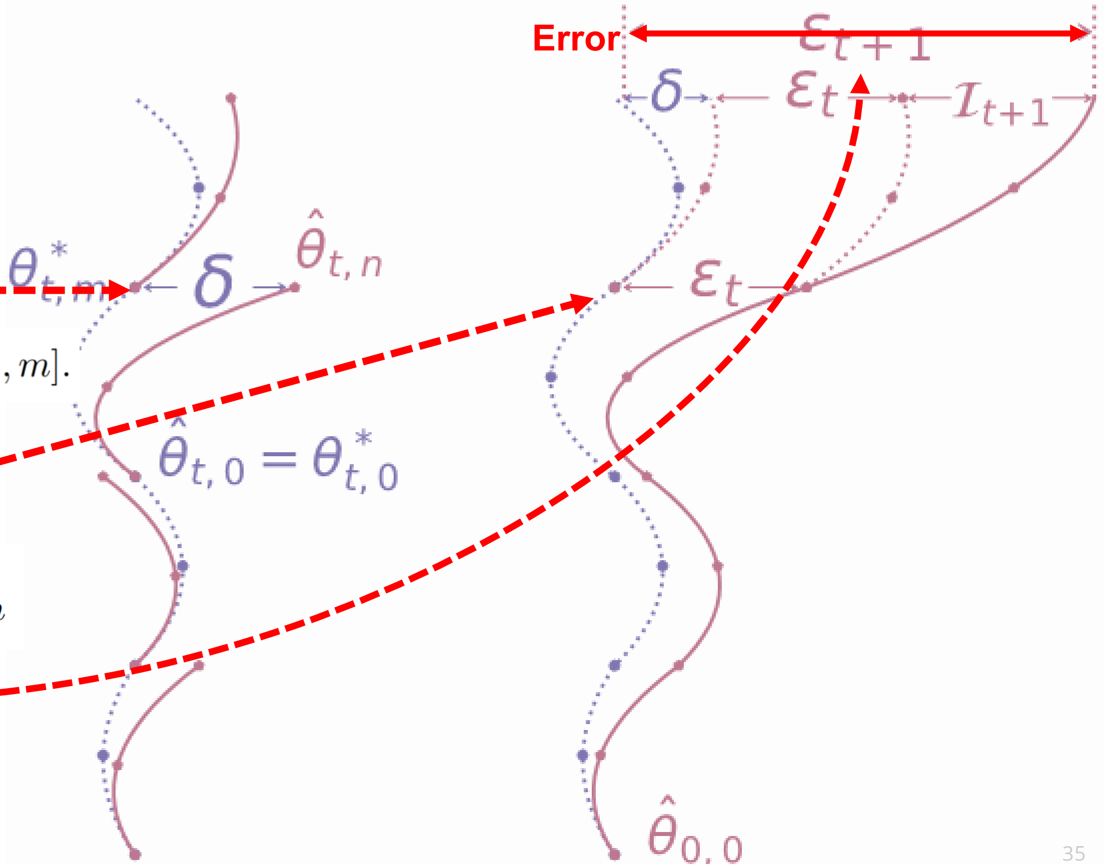
$$\delta_t = \mathcal{A}[\nabla_\theta L_{\mathcal{S}}(f_{\theta^*_{t-1,m}}), n] - \mathcal{A}[\nabla_\theta L_{\mathcal{T}}(f_{\theta^*_{t-1,m}}), m].$$

(2) Match error makes trajectory error

$$\epsilon_t = \hat{\theta}_{t+1,0} - \theta^*_{t+1,0} = \hat{\theta}_{t,n} - \theta^*_{t,m}$$

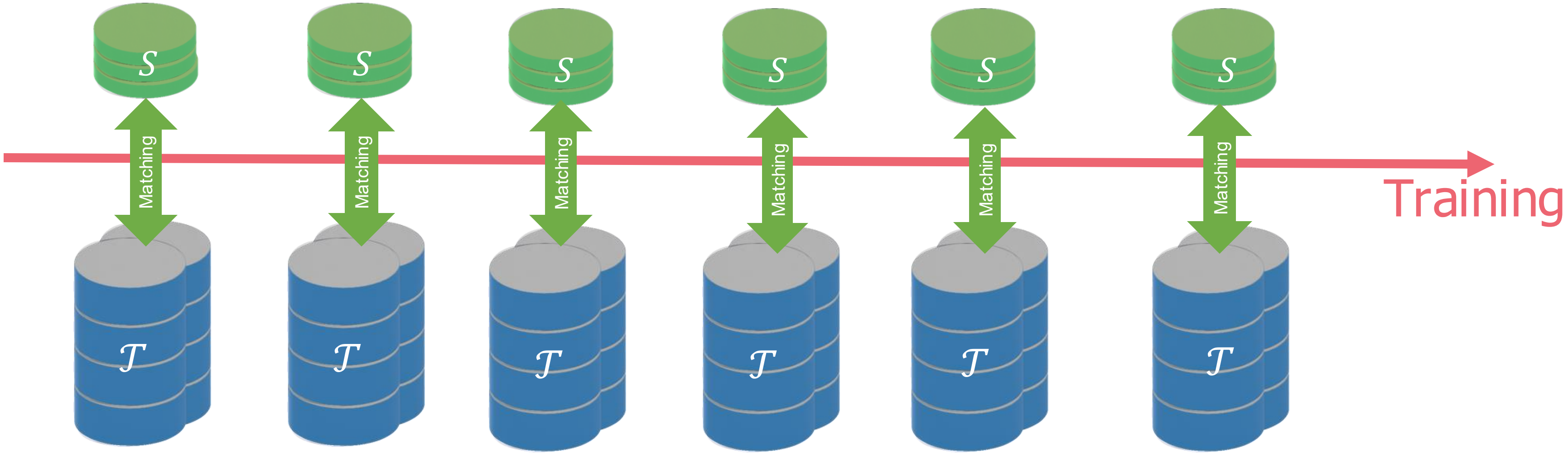
(3) trajectory error will be accumulated

$$\epsilon_{t+1} = \epsilon_t + \delta_{t+1} + \mathcal{I}(\theta^*_{t,m}, \epsilon_t)$$



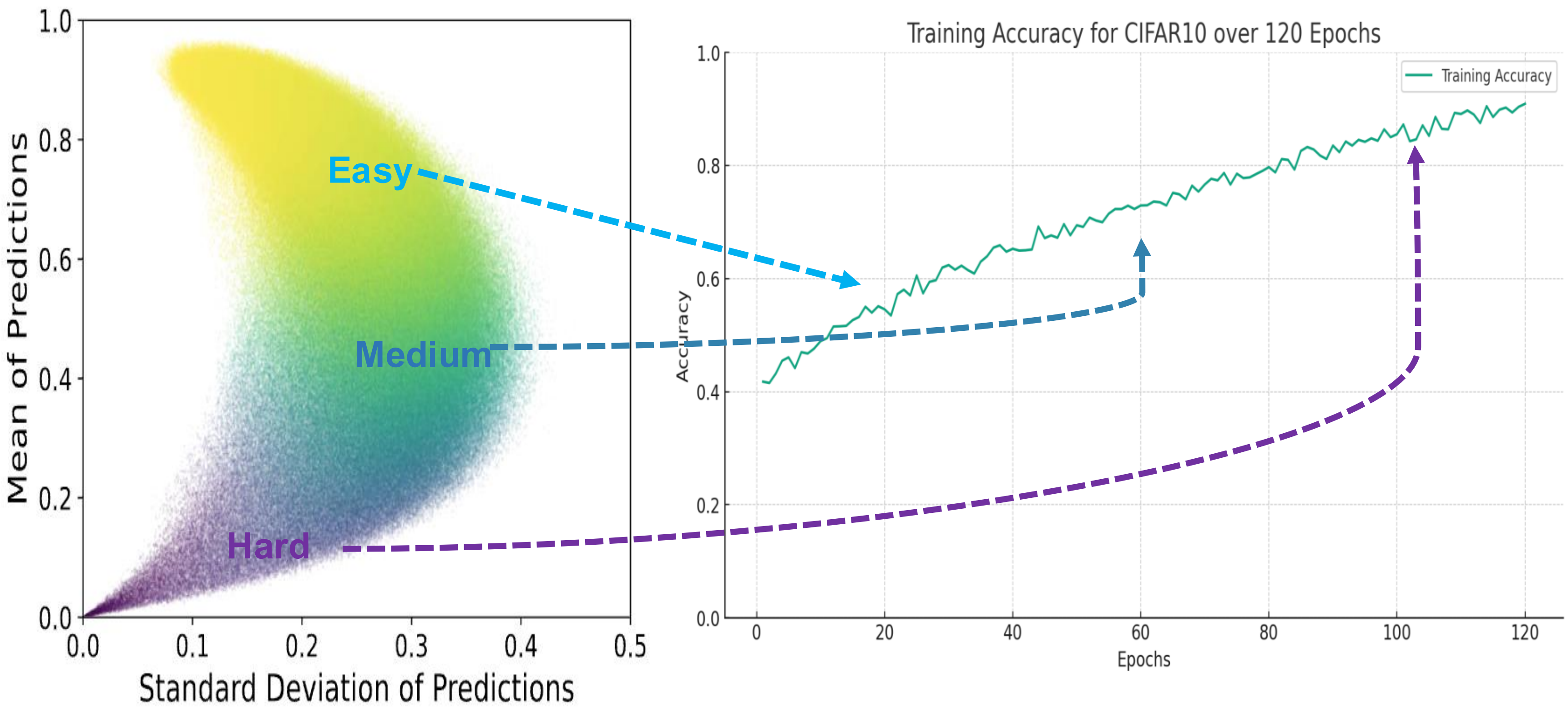
NeurIPS 2023: Sequential Subset Matching for Dataset Distillation

All the distillation methods follow:



However, data is learned **in sequence** in original dataset.

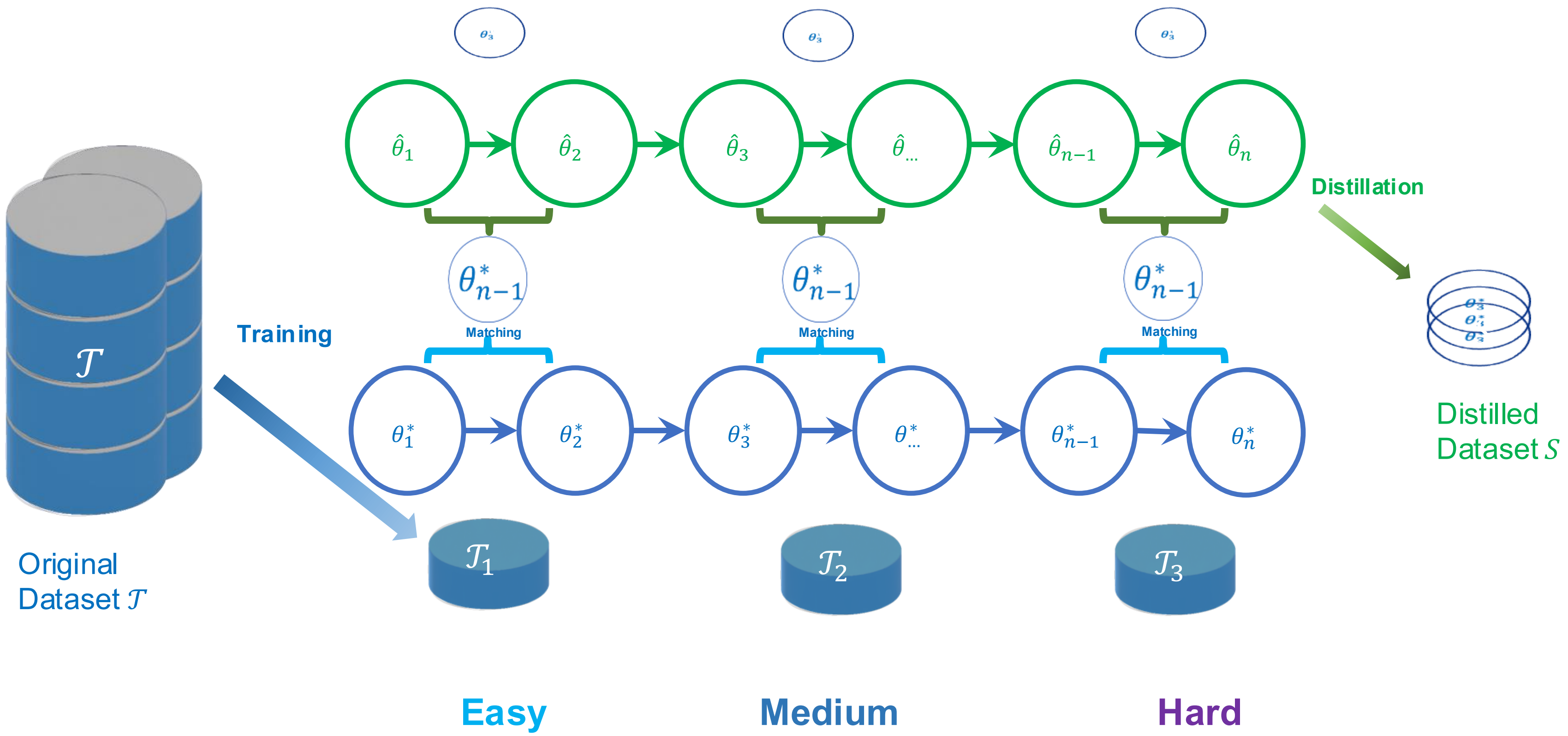
Data can be grouped as "Easy", "Medium", and "Hard" subsets.



Simply group data by mean and derivation of predictions

Data is learned in sequence

Thus, we propose Sequential Matching Method



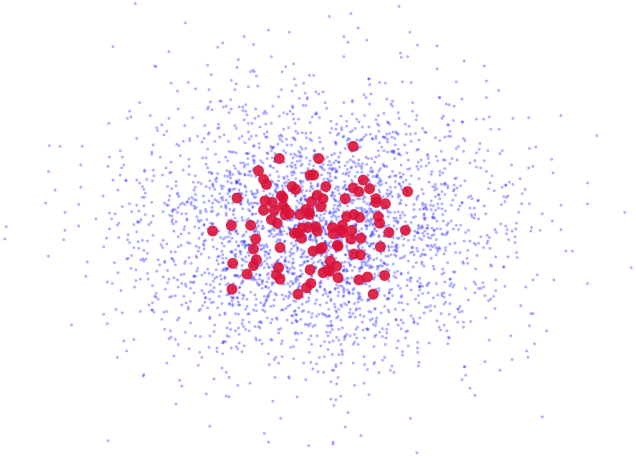
CREATING GROWTH, ENHANCING LIVES

Diversity-Driven Synthesis （NeurIPS 24）

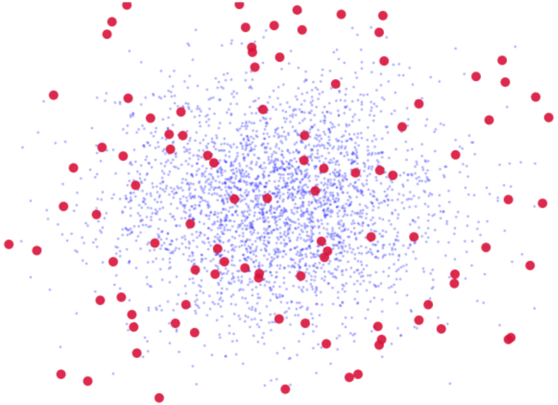
Distillation progress is to solve

$$\underset{\boldsymbol{s}_i \in \mathbb{R}^d}{\arg\min} \left[\ell\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right) + \lambda \mathcal{L}_{\mathrm{BN}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)\right]$$

However, distilled data is clustered at the central



Sre2L distilled data



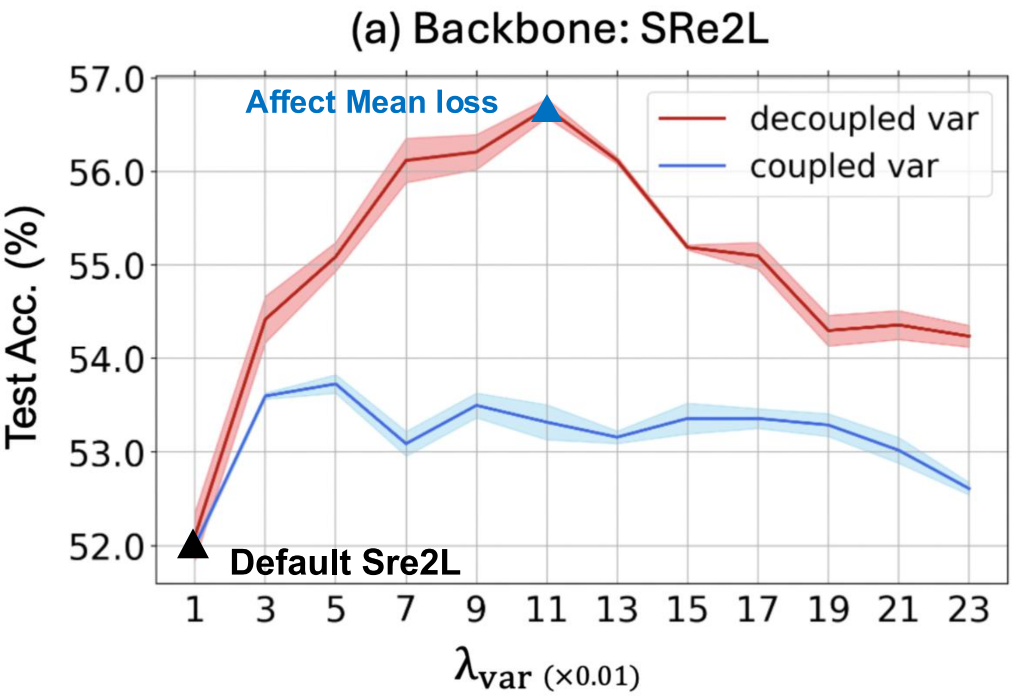
Ideal distilled data

Diversity limitations

Our feasibility experiments

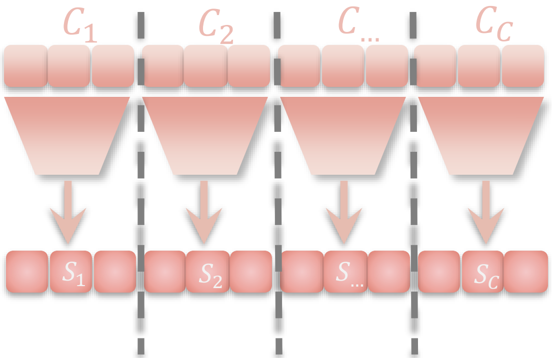
We decouple the BN loss and **emphasize** the variation loss only

$$\mathcal{L}_{\text{mean}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)+\boxed{\lambda_{\text{var}}}\mathcal{L}_{\text{var}}\left(f_{\theta_{\mathcal{T}}}, \boldsymbol{s}_i\right)$$

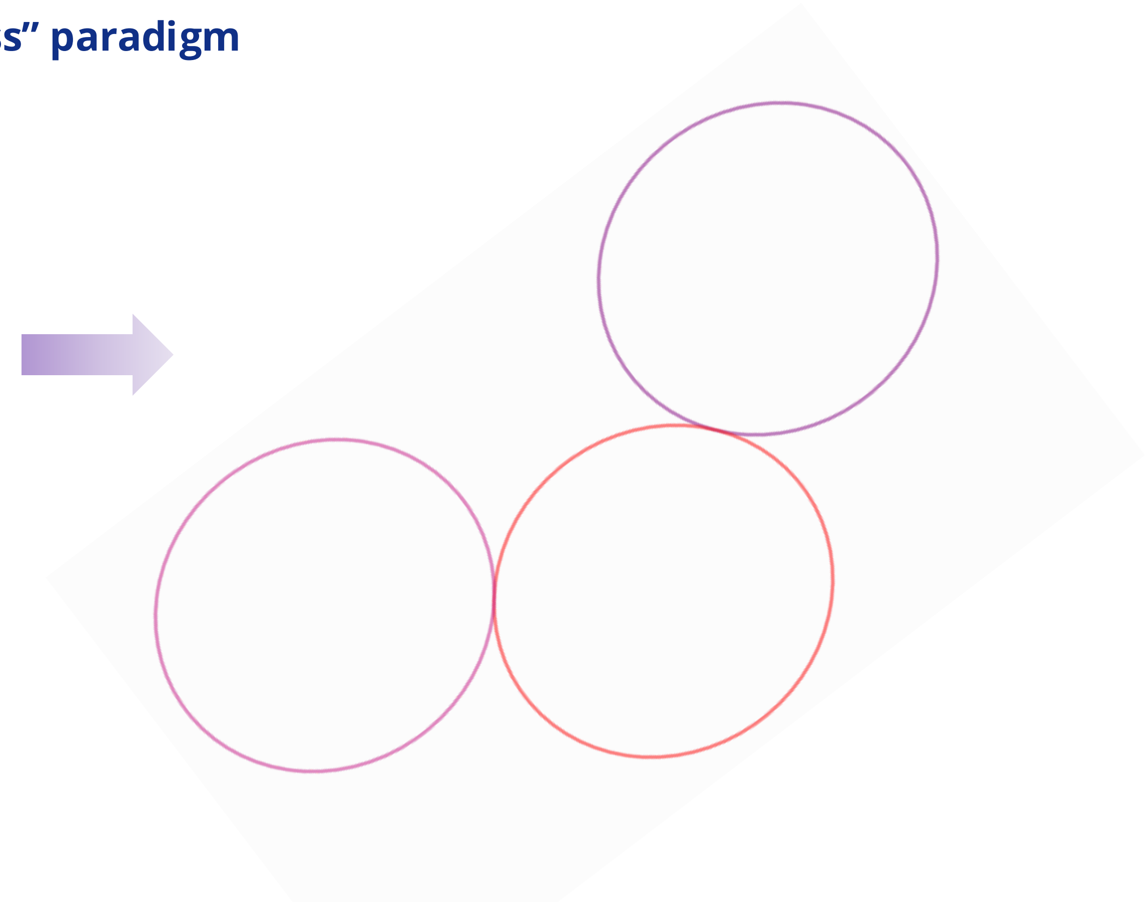


How do existing methods achieve DD?

- “One instance for one class” paradigm

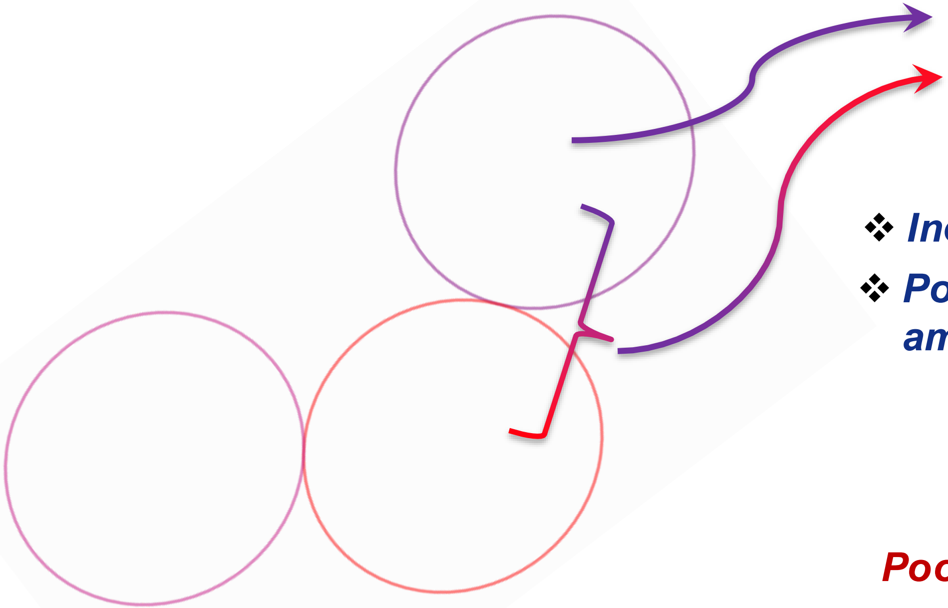


Class Barriers



How do existing methods achieve DD?

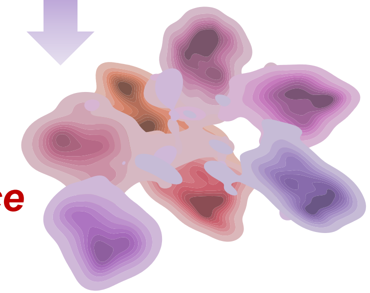
- ❖ "One instance for one class" paradigm



- ❖ *Duplicated intra-class features*
- ❖ *Oversight of inter-class features*

- ❖ ***Inefficient utilization of the distillation budget***
- ❖ ***Poor generalization in complex or ambiguous scenarios***

Poor performance

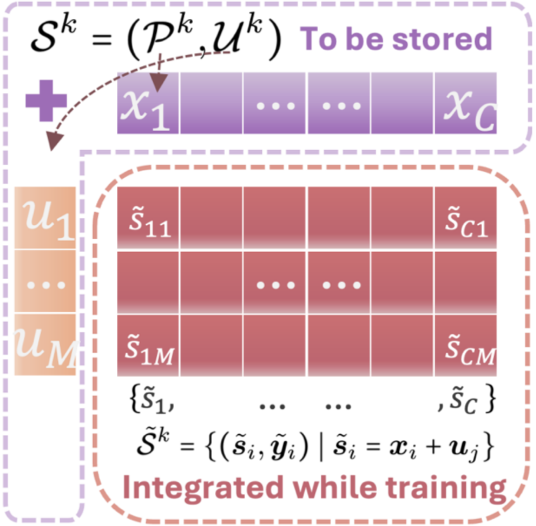


evaluated on validation set

Breaking Class Barriers: Efficient Dataset Distillation via Inter-class Feature Compensator

Breaking Class Barriers:
Efficient DD via Inter-class Feature Compensator

❖ **“One instance for all classes” paradigm**



- *Design :*

$$\tilde{\mathcal{S}}^k = \{(\tilde{\boldsymbol{s}}_i, \tilde{\boldsymbol{y}}_i) \mid \tilde{\boldsymbol{s}}_i = \boldsymbol{x}_i + \boldsymbol{u}_j, \\ \text{for each } \boldsymbol{x}_i \in \mathcal{P}^k \text{ and each } \boldsymbol{u}_j \in \mathcal{U}^k\}$$

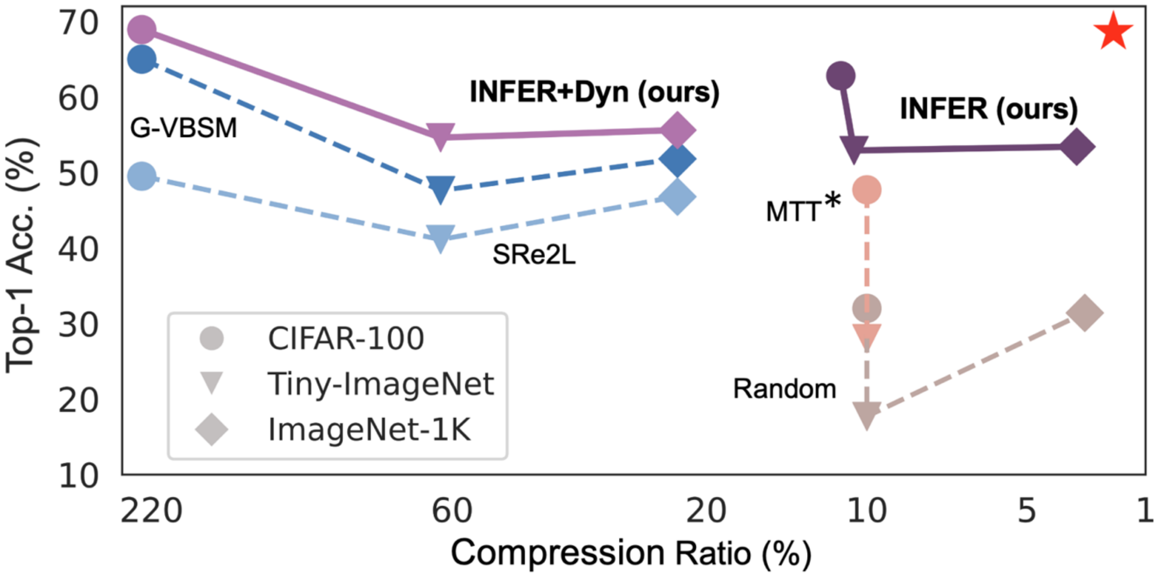
- *Optimization :*

$$\underset{\boldsymbol{u}_j \in \mathbb{R}^d}{\arg\min} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{P}^k} \left[\ell\left(f_{\theta_\mathcal{T}}, \boldsymbol{x}_i + \boldsymbol{u}_j, \boldsymbol{y}_i\right) + \alpha \mathcal{L}_{\mathrm{BN}}\left(f_{\theta_\mathcal{T}}, \boldsymbol{x}_i + \boldsymbol{u}_j\right)\right],$$

$$\text{where} \quad \mathcal{L}_{\mathrm{BN}}\left(f_{\theta_\mathcal{T}}, \boldsymbol{x}_i + \boldsymbol{u}_j\right) = \sum_l \|\mu_l(\tilde{\mathcal{S}}_j^k) - \mu_l(\mathcal{T})\|_2 + \sum_l \|\sigma_l^2(\tilde{\mathcal{S}}_j^k) - \sigma_l^2(\mathcal{T})\|_2$$

Breaking Class Barriers:
Efficient DD via Inter-class Feature Compensator

❖ Experiment Results



Better performance & higher compression ratio!!!

Recently Published Papers on this domain

[Sequential Subset Matching for Dataset Distillation]
Jiawei Du, Qin Shi, **Joey Tianyi Zhou***
In NeurIPS 2023

[Minimizing the accumulated trajectory error to improve dataset distillation]
Jiawei Du, Yidi Jiang, Vincent TF Tan, **Joey Tianyi Zhou***, Haizhou Li
In CVPR 2023

[You Only Condense Once: Two Rules for Pruning Condensed Datasets]
Yang He, Lingao Xiao, **Joey Tianyi Zhou***
In NeurIPS 2023

[Meta Knowledge Condensation for Federated Learning]
Ping Liu, Xin Yu, and **Joey Tianyi Zhou***
In ICLR 2023

[Multisize Dataset Condensation]
Yang He, Lingao Xiao, **Joey Tianyi Zhou***, Ivor Tsang
*In ICLR 2024 (**Oral, 1.2%**)*

[Spanning Training Progress: Temporal Dual-Depth Scoring (TDDS) for Enhanced Dataset Pruning]
Xin Zhang, Jiawei Du, Weiying Xie, Yunsong Li, **Joey Tianyi Zhou***
In CVPR 2024

[Evolution-aware VAriance (EVA) Coreset Selection for Medical Image Classification]
Yuxin Hong, Xiao Zhang, Xin Zhang, **Joey Tianyi Zhou**
*In ACM MM 2024 (**Best Paper Nomination**)*

[Diversity-Driven Synthesis: Enhancing Dataset Distillation through Directed Weight Adjustment]
Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, **Joey Tianyi Zhou***
*In NeurIPS 2024 (**Spotlight Paper**)*

[Breaking Class Barriers: Efficient Dataset Distillation via Inter-Class Feature Compensator]
Xin Zhang, Jiawei Du, Ping Liu, **Joey Tianyi Zhou***
In ICLR 2025

CREATING GROWTH, ENHANCING LIVES

Acknowledgement

The presenter wishes to acknowledge the International Neural Network Society for their sponsorship of the Webinar Series.

Join the International Neural Network Society

Computational, perceptual, and brain-inspired since 1987

scan
for
more

- Your gateway to Neural Networks excellence
- Exploring NNs in hardware, software and wetware
- Discounts for IJCNN & OA to Neural Networks
- Nurturing and investing in young talent