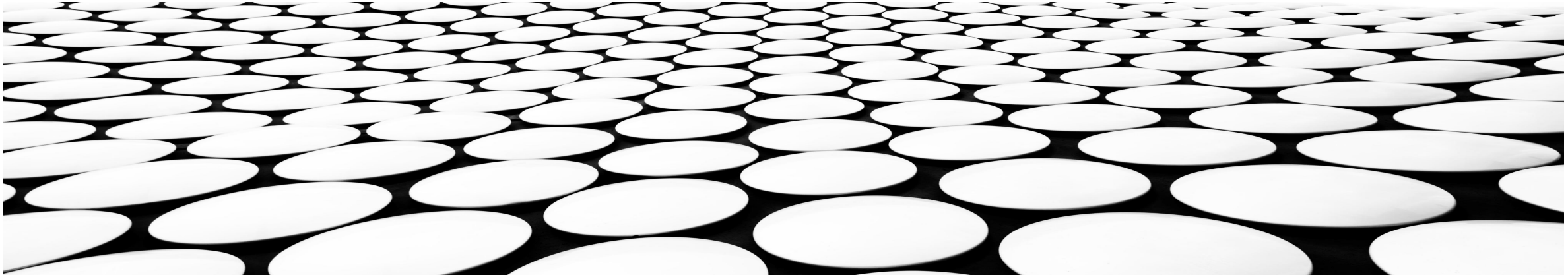
Paper Share: Dataset Condensation

- **数据集压缩：目的**

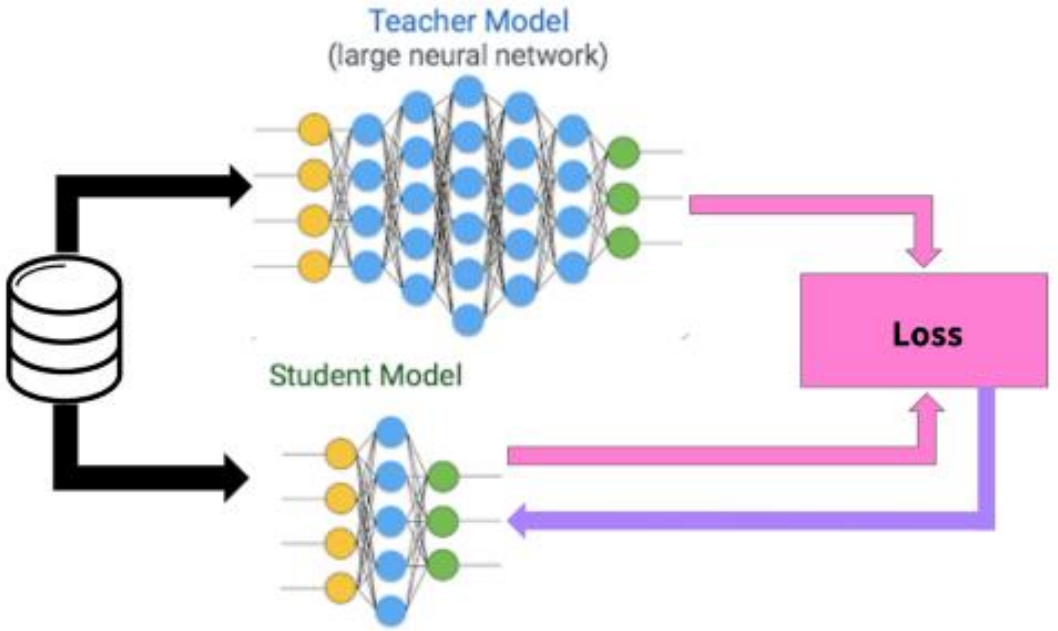
给定数据集大小$R^{M\times d} \rightarrow R^{N\times d}$，其中$N \ll M$



Dataset Distillation

- **涉及distillation——抽象提取**

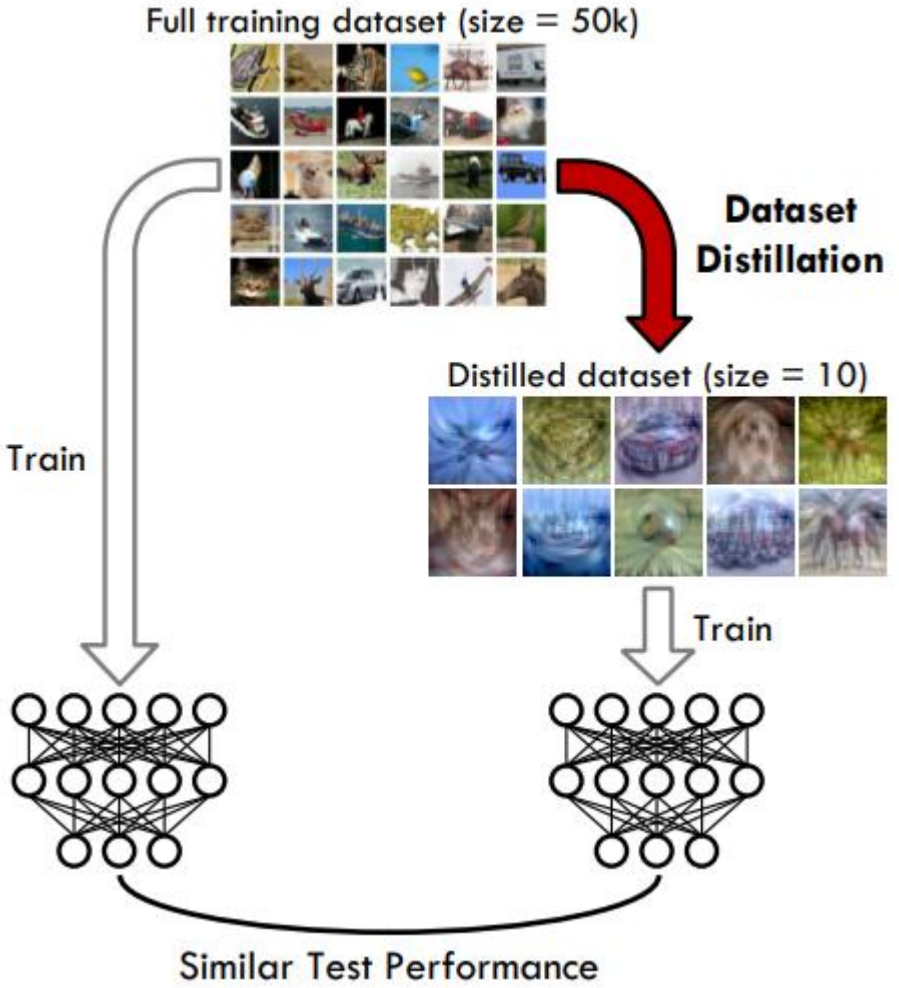
- **模型蒸馏（知识蒸馏）：**
  - 学生网络是比教师网络规模小很多的网络模型
  - 学生网络通过**模仿**教师网络的**行为**达到学习的效果
  - 不同的知识蒸馏方法定义了不同的需要模仿的行为（如：中间层特征蒸馏、输出层蒸馏等）
  - 当学生网络在定义的行为上达到和教师网络相似的表现时，认为达到了蒸馏的效果，获得了好的学生网络



Dataset Distillation

- **涉及distillation——抽象提取**

- **数据集蒸馏(dataset distillation)：**
  - **学生**：合成数据集　**教师**：原数据集
  - 在数据集蒸馏中合成数据集**模仿**原数据集的什么**行为?**
    - 仍然离不开模型，因为数据只有输入到模型中才有意义
    - 模仿原数据集在模型中的**某种表现——对应不同的数据集蒸馏方法**
  - 当合成数据集在**指定的表现**上和原数据集相似时，认为获得了好的合成数据



Dataset Distillation

- **一般流程**

- **定义合成数据集的学习目标**

Dataset Condensation With Gradient Matching

- 同一网络，如果输入真实图像计算的loss的梯度与输入合成图像计算的loss的梯度相近，认为获得了好的合成数据

Dataset Condensation With Distribution Matching

- 在不同嵌入空间中真实图像分布与合成图像分布之间的距离尽可能小时，认为获得了好的合成数据

Dataset Meta-learning From Kernel Ridge-regression

- 在合成数据集S上利用KRR预测来自真实数据集的输入，如果预测的正确率高说明合成了好的数据。

Dataset Distillation By Matching Training Trajectories

- 当合成数据的训练轨迹（训练后的模型参数）和真实数据相近时，认为获得了好的合成数据

Dataset Distillation

- **一般流程**
  - **具体算法**
  - **需要给出对应方法的具体算法，算法包括：**
    - 初始化细节（合成数据集的初始化、模型的初始化、其他参数……）
    - 如何更新网络和合成数据集（交替更新、单层/双层循环、内部/外部循环……）
  - **举例**

Algorithm 1: Dataset condensation with gradient matching

Input: Training set $\mathcal{T}$

Required: Randomly initialized set of synthetic samples $\mathcal{S}$ for C classes, probability distribution over randomly initialized weights P_{θ_0}, deep neural network $\phi_{\boldsymbol{\theta}}$, number of outer-loop steps K, number of inner-loop steps T, number of steps for updating weights $\varsigma_{\boldsymbol{\theta}}$ and synthetic samples $\varsigma_{\mathcal{S}}$ in each inner-loop step respectively, learning rates for updating weights $\eta_{\boldsymbol{\theta}}$ and synthetic samples $\eta_{\mathcal{S}}$.

for $k = 0, \cdots, K-1$ **do**

  Initialize $\boldsymbol{\theta}_0 \sim P_{\theta_0}$

  for $t = 0, \cdots, T-1$ **do**

    for $c = 0, \cdots, C-1$ **do**

      Sample a minibatch pair $B_c^{\mathcal{T}} \sim \mathcal{T}$ and $B_c^{\mathcal{S}} \sim \mathcal{S}$ ▷ $B_c^{\mathcal{T}}$ and $B_c^{\mathcal{S}}$ are of the same class c.

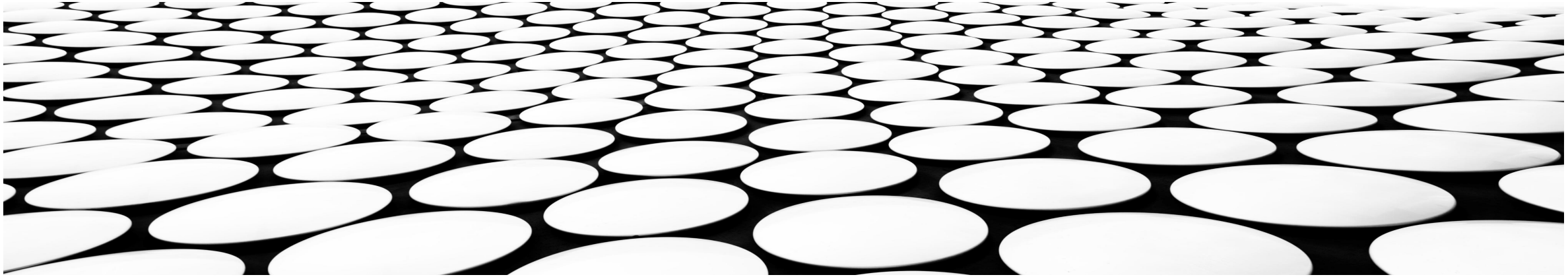
      Compute $\mathcal{L}_c^{\mathcal{T}} = \frac{1}{|B_c^{\mathcal{T}}|} \sum_{(\boldsymbol{x},y)\in B_c^{\mathcal{T}}} \ell(\phi_{\boldsymbol{\theta}_t}(\boldsymbol{x}), y)$ and $\mathcal{L}_c^{\mathcal{S}} = \frac{1}{|B_c^{\mathcal{S}}|} \sum_{(\boldsymbol{s},y)\in B_c^{\mathcal{S}}} \ell(\phi_{\boldsymbol{\theta}_t}(\boldsymbol{s}), y)$

      Update $\mathcal{S}_c \leftarrow \texttt{opt-alg}_{\mathcal{S}}(D(\nabla_{\boldsymbol{\theta}}\mathcal{L}_c^{\mathcal{S}}(\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}}\mathcal{L}_c^{\mathcal{T}}(\boldsymbol{\theta}_t)), \varsigma_{\mathcal{S}}, \eta_{\mathcal{S}})$

    Update $\boldsymbol{\theta}_{t+1} \leftarrow \texttt{opt-alg}_{\boldsymbol{\theta}}(\mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta}_t), \varsigma_{\boldsymbol{\theta}}, \eta_{\boldsymbol{\theta}})$ ▷ Use the whole $\mathcal{S}$

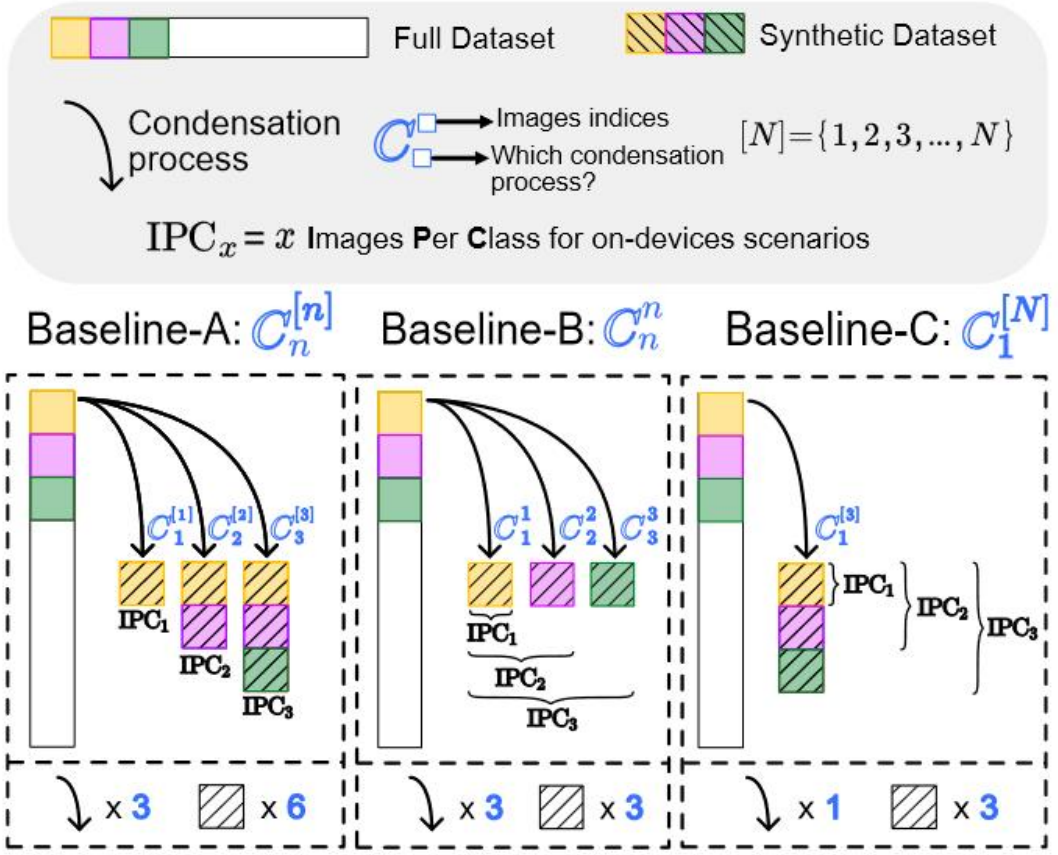
Output: $\mathcal{S}$

MULTISIZE DATASET CONDENSATION ICLR 2024

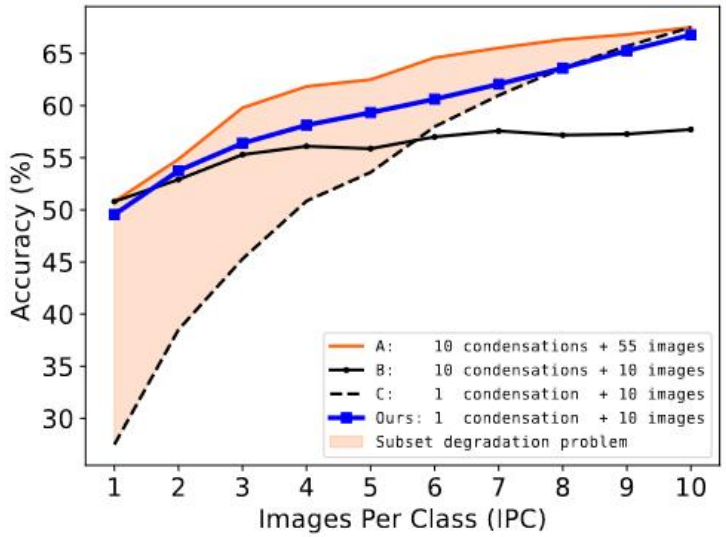


Multisize Dataset Condensation

- **问题阐述：**当需要灵活大小(image per class, IPC)的压缩数据集时，会面临“子集退化问题”(subset degradation problem)



- **subset degradation problem：**
  - 当需要一个较小尺寸的数据集时，从压缩好的数据集中再选择具有更小IPC子集相比直接从原数据集中压缩一个目标尺寸的数据集的方法产生了性能的损失
  - 定量实验：右图



- **随着子集大小的变小，准确性差异也会增大**

Multisize Dataset Condensation

- **MDC：**
  - 解决子集退化问题——**adaptive subset loss**
  - 单个压缩过程

	Symbol	Condense	Storage
A	$\mathbb{C}_n^{[n]}, n \in \{1, 2, \ldots, N\}$	N	$1+2+\ldots+N$
B	$\mathbb{C}_n^n, n \in \{1, 2, \ldots, N\}$	N	N
C	$\mathbb{C}_1^{[N]}$	1	N
Ours	$\mathbb{C}_1^{[N]}$	1	N

- **Gradient Matching方法：**

$$\min_{\mathcal{S}\in\mathbb{R}^{N\times d}} D\left(\nabla_\theta \ell(\mathcal{S};\theta), \nabla_\theta \ell(\mathcal{B};\theta)\right) = D(\mathcal{S},\mathcal{B};\theta)$$

$$\mathcal{S} \leftarrow \mathcal{S} - \lambda \nabla_{\mathcal{S}} D\left(\mathcal{S},\mathcal{B};\theta\right), \quad \theta \leftarrow \theta - \eta \nabla_\theta \ell\left(\theta;\mathcal{S}\right)$$

- **新的压缩目标(基于Gradient Matching)：**

$$\min_{\mathcal{S}\in\mathbb{R}^{N\times d}} D\left(\nabla_\theta \ell\left(\mathcal{S}_{[1]}, \mathcal{S}_{[2]}, \ldots \mathcal{S}_{[N]};\theta\right), \nabla_\theta \ell\left(\mathcal{B};\theta\right)\right)$$

$$\mathcal{S}_{[n]} = \mathcal{S}_{\{1,2,\ldots,n\}} \subset \mathcal{S} = \mathcal{S}_{[N]}$$

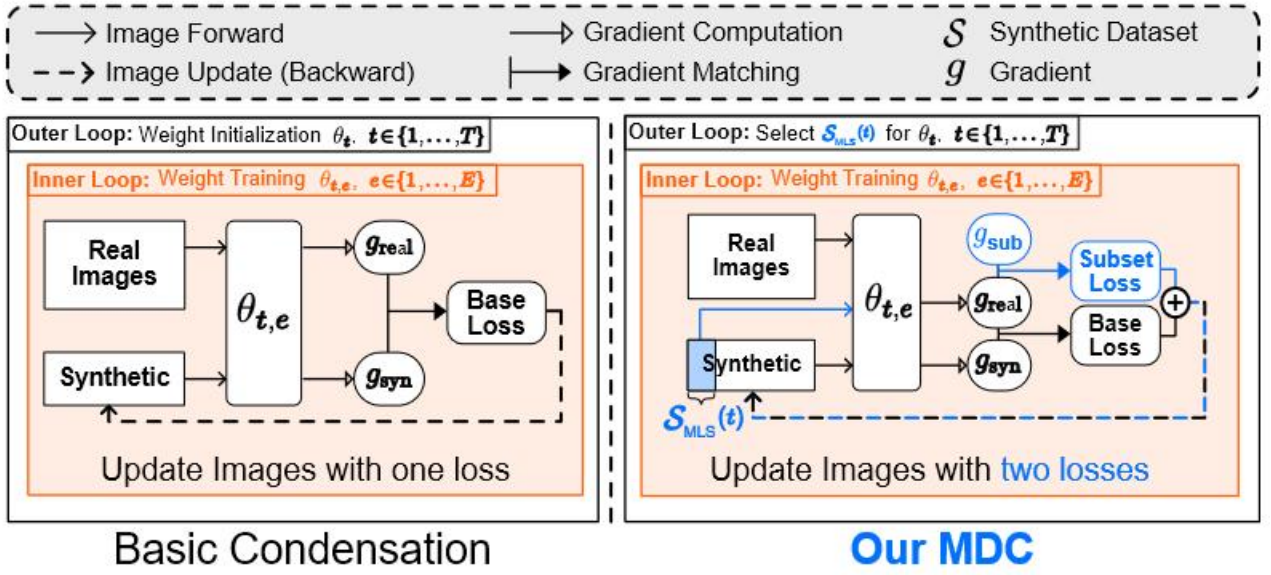
- 让每一个$S_{[n]}$都参与和原数据集B之间的Gradient Matching：
  - $S_{[N]}$：contributes to **“base loss”**
  - $S_{[1],[2],\ldots\ldots,[N-1]}$：contributes to **“subset loss”**
  - 初步的想法是让S的每个大小的子集都参与到Loss的计算中，这样能让[1, N]每个尺寸的子集都有一定的代表性

Multisize Dataset Condensation

- **新的压缩目标(基于Gradient Matching)：**

$$\min_{\mathcal{S}\in\mathbb{R}^{N\times d}} D\left(\nabla_\theta \ell\left(\mathcal{S}_{[1]}, \mathcal{S}_{[2]}, \ldots \mathcal{S}_{[N]}; \theta\right), \nabla_\theta \ell\left(\mathcal{B}; \theta\right)\right)$$

$$\mathcal{S}_{[n]} = \mathcal{S}_{\{1,2,\ldots,n\}} \subset \mathcal{S} = \mathcal{S}_{[N]}$$



Basic Condensation　　**Our MDC**

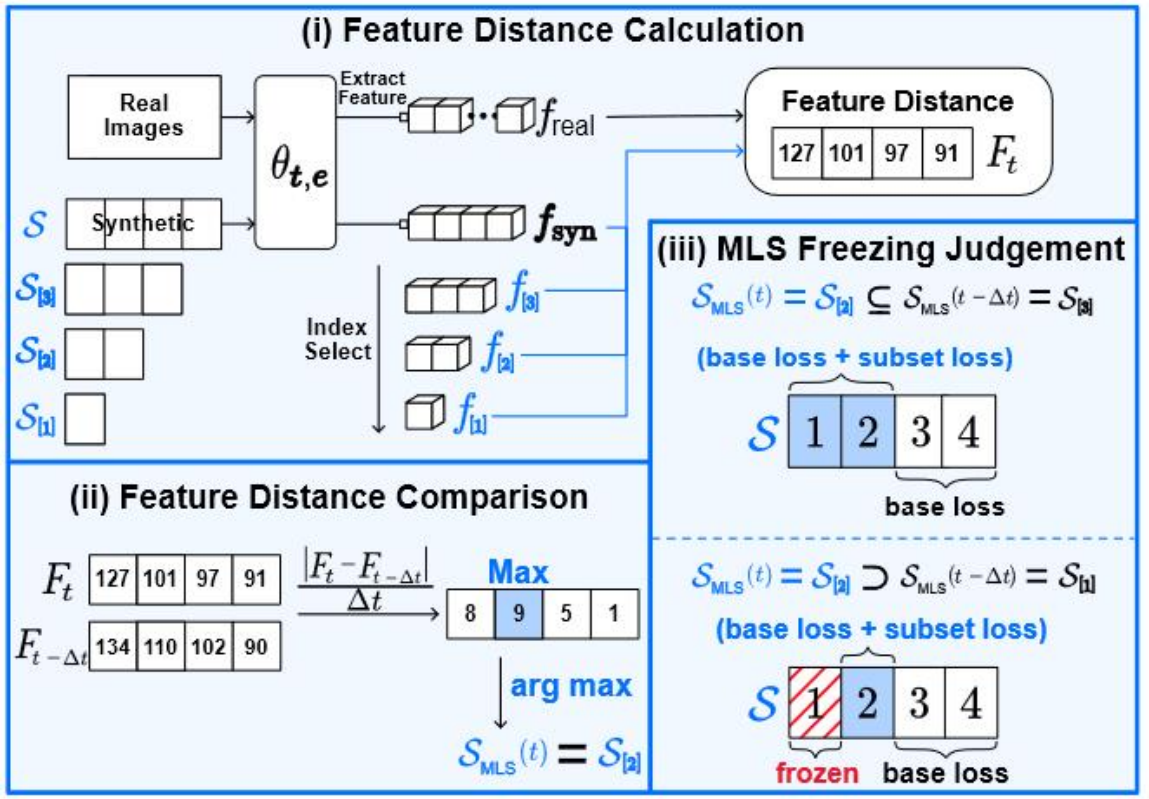
- 在单个压缩过程中将subset loss 融入base loss：
  - **新的更新策略**

$$\mathcal{S} \leftarrow \mathcal{S} - \lambda\left(\nabla_{\mathcal{S}} D\left(\mathcal{S}, \mathcal{B}; \theta\right) + \nabla_{\mathcal{S}_{[n]}} D\left(\mathcal{S}_{[n]}, \mathcal{B}; \theta\right)\right), \quad n \in [1, N-1]$$

  - 每隔一段时间选择一个当前最值得被优化的子集$S_{[n]}$，计算base loss和$S_{[n]}$的subset loss，反向传播更新合成数据集

Multisize Dataset Condensation

- **如何选择当前最具代表性的（最值得被优化的）子集：**
  - **Most Learnable Subset (MLS)**



- 三步选出MLS：
  - **I. Feature Distance Calculation:**
    - 如果按之前预定义的压缩目标，需要在N−1个子集上进行N−1次前向传播和N−1次反向传播
    - 使用**feature distance**代替**gradient distance**
    - feature distance的计算只涉及1次整体的forward计算feature，因为大小为n的子集feature可以直接从大小为N的合成数据集feature中提取
    - 子集 S[n] 在迭代次数 t 处的feature distance可以表示为：

$$F_t\left(\mathcal{S}_{[n]}, \mathcal{B}\right) = D\left(f_t\left(\mathcal{S}_{[n]}\right), f_t(\mathcal{B})\right)$$

- ✓ D(·)可以是MSE等距离度量函数

Multisize Dataset Condensation

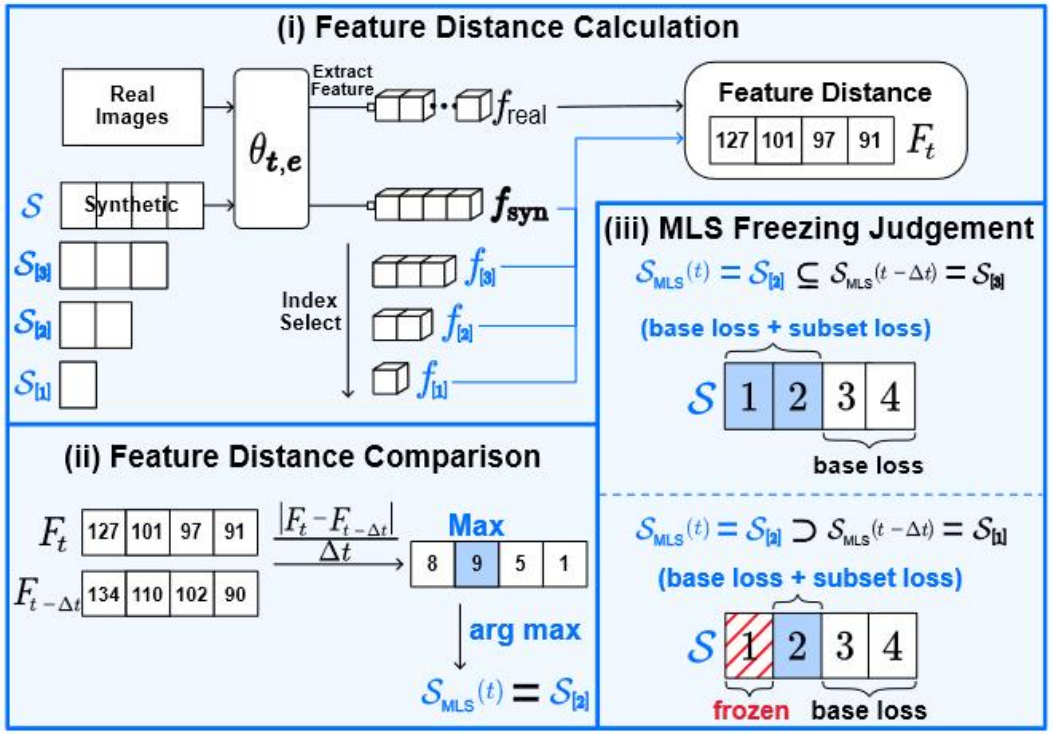
- **选择MLS：**

II. Feature Distance Comparison:

- 一般认为较大子集会比较小的子集更具有代表性(即与原数据集的feature distance更小)，因此初始情况下，认为$\mathbf{S_{[1]}}$与原数据集的feature distance在所有子集中是最大的。因此初始阶段认为$S_{[1]}$最需要被优化，即**选择$\mathbf{S_{[1]}}$为初始的MLS**
- 在后续的迭代过程中，随着压缩的进行，子集变得更具代表性。同一子集与原数据集的feature distance应该是逐渐减小的

$$F_{t-\Delta t}\left(\mathcal{S}_{[p]}, \mathcal{B}\right) > F_t\left(\mathcal{S}_{[p]}, \mathcal{B}\right)$$

- 此时的MLS应是特征距离缩减率(**feature distance reduction rate**)最高的子集



$$\mathcal{S}_{\mathrm{MLS}}(t) = \mathcal{S}_{[n_t^*]} = \underset{\mathcal{S}_{[n]}}{\arg\max}\left(R\left(\mathcal{S}_{[n]}, t\right)\right) \quad \text{where} \quad n \in [1, N-1].$$

✓ **定义feature distance reduction rate**

$$R(\mathcal{S}_{[n]}, t) = \frac{\Delta F_{\mathcal{S}_{[n]}}}{\Delta t} = \frac{\left|F_t\left(\mathcal{S}_{[n]}, \mathcal{B}\right) - F_{t-\Delta t}\left(\mathcal{S}_{[n]}, \mathcal{B}\right)\right|}{\Delta t}$$

✓ 寻找在时间间隔 Δt 内距 B 的特征距离下降最快的子集。这表明该子集正在以最快的速度“学习”，因此认为它是最可学习的子集（MLS）

Multisize Dataset Condensation

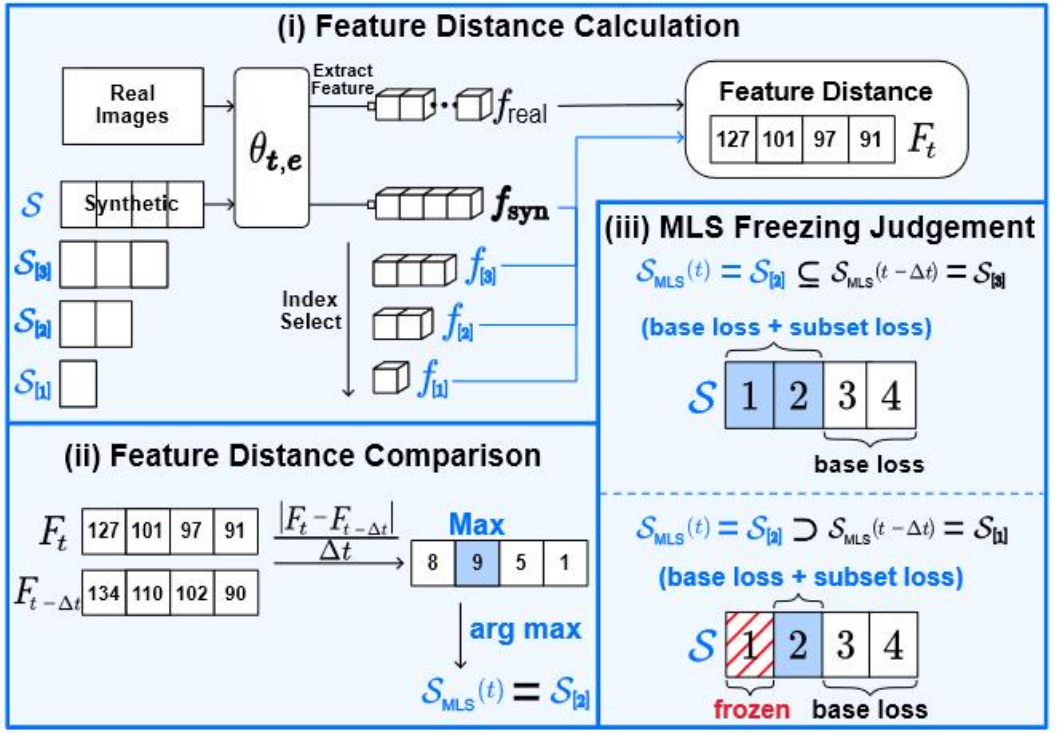
- **选择MLS：** $$\mathcal{S} \leftarrow \mathcal{S} - \lambda \left(\nabla_{\mathcal{S}} D\left(\mathcal{S}, \mathcal{B}; \theta\right) + \nabla_{\mathcal{S}_{[n]}} D\left(\mathcal{S}_{[n]}, \mathcal{B}; \theta\right) \right), \quad n \in [1, N-1]$$

III. MLS Freezing Judgement:

- 根据当前MLS及上一阶段MLS的大小，某些图像被冻结不被更新

$$\begin{aligned} &\text{Update} \quad \mathcal{S} && \text{if } \mathcal{S}_{\text{MLS}}(t) \subseteq \mathcal{S}_{\text{MLS}}(t-\Delta t) \\ &\text{Update} \quad \mathcal{S} \setminus \mathcal{S}_{\text{MLS}}(t-\Delta t) && \text{if } \mathcal{S}_{\text{MLS}}(t) \supset \mathcal{S}_{\text{MLS}}(t-\Delta t) \end{aligned}$$

- \ 是集合减法的符号
- 如果当前 MLS(t) 的大小小于等于前一阶段MLS(t − Δt)的大小，更新整个合成数据 S
- 当前MLS的大小大于MLS(t − Δt)大小时，更新整个S将导致上一阶段的优化受到新梯度的负面影响。因此，冻结前面的 MLS (t − Δt) 以保留已经学习的信息
- S \ S_{MLS}(t − Δt)被更新



Multisize Dataset Condensation

- **整体算法**

Algorithm 1 Multisize Dataset Condensation

Input: Full dataset $\mathcal{B}$, model Θ, MLS selection period Δt, learning rate of the synthetic dataset λ, learning rate of the model η, outer loop iterations T, inner loop epochs E, and class loop iterations C.

Output: Synthetic dataset $\mathcal{S}$

Initialize synthetic dataset $\mathcal{S}$
Initialize the most learnable subset (MLS) $\mathcal{S}_{\text{MLS}}$
for $t = 1$ **to** T **do** ▷ Outer loop
    Randomly initialize model weight θ_t
    for $e = 1$ **to** E **do** ▷ Inner loop
        for $c = 1$ **to** C **do** ▷ Class loop
            Sample class-wise mini-batches $B_c \sim \mathcal{B}$, $S_c \sim \mathcal{S}$
            Update $\mathcal{S}_c$ with **subset loss** according to Eq. 10
        end for
        $\theta_{t,e+1} \leftarrow \theta_{t,e} - \eta \nabla_\theta \ell\left(\theta_{t,e}; B\right)$ ▷ Update model with real image mini-batch $B \sim \mathcal{B}$
    end for
    if $t \,\%\, \Delta t$ is 0 **then** ▷ Every Δt iterations
        Select $\mathcal{S}_{\text{MLS}}$ according to Eq. 9
    end if
end for
return Synthetic dataset $\mathcal{S}$

Multisize Dataset Condensation

- **Experiments**
  - 使用最后一层特征计算feature distance
  - 在outer loop中每隔100次重新选择一次MLS(Δt = 50 for ImageNet-10)

Dataset		1	2	3	4	5	6	7	8	9	10	Avg.	Diff.
SVHN	A	68.50†	75.27	79.55	81.85	83.33	84.53	85.66	86.05	86.25	87.50†	81.85	-
	B	68.50†	71.65	71.27	71.92	73.28	70.74	71.83	71.08	71.97	71.55	71.38	-
	C	35.48	51.55	60.42	67.97	74.38	77.65	81.70	83.86	85.96	87.50†	70.65	0
	Ours	63.26	67.91	72.15	74.09	77.54	78.17	80.92	82.82	84.27	86.38	76.75	+6.10
CIFAR-10	A	50.80	54.85	59.79	61.84	62.49	64.59	65.53	66.33	66.82	67.50†	62.05	-
	B	50.80	53.17	55.09	56.17	55.80	56.98	57.60	57.78	58.22	58.38	56.00	-
	C	27.49	38.50	45.29	50.85	53.60	57.98	60.99	63.60	65.71	67.50†	53.15	0
	Ours	49.55	53.75	56.39	59.33	58.13	60.62	62.06	63.59	65.25	66.79	59.55	+6.40
CIFAR-100	A	28.90†	34.28	37.35	39.13	41.15	42.65	43.62	44.48	45.07	45.40	40.20	-
	B	28.90†	30.63	31.64	31.76	32.61	32.85	33.03	33.04	33.32	33.39	32.12	-
	C	14.38	21.76	28.01	32.21	35.27	39.09	40.92	42.69	44.28	45.40	34.40	0
	Ours	27.58	31.83	33.59	35.42	36.93	38.95	40.70	42.05	43.86	44.34	37.53	+3.13

(a) Results of SVHN, CIFAR-10, CIFAR-100 targeting IPC_{10}.

Multisize Dataset Condensation

- **Experiments**

Dataset		1	2	3	4	5	6	7	8	9	10	20	30	40	50	Avg.	Diff.
SVHN	A	68.50†	75.27	79.55	81.85	83.33	84.53	85.66	86.05	86.25	87.50†	89.54	90.27	91.09	91.38	84.34	-
	C	34.90	46.52	52.23	56.30	62.25	65.34	68.84	69.57	71.95	74.69	83.73	87.83	89.73	91.38	68.23	0
	Ours	58.77	67.72	69.33	72.26	75.02	73.71	74.50	74.63	76.21	76.87	83.67	87.08	89.46	91.39	76.47	+8.24
CIFAR-10	A	50.80	54.85	59.79	61.84	62.49	64.59	65.53	66.33	66.82	67.50†	70.82	72.86	74.30	75.07	65.26	-
	C	27.87	35.69	41.93	45.29	47.54	51.96	53.51	55.59	56.62	58.26	66.77	70.50	72.98	74.50	54.21	0
	Ours	47.83	52.18	56.29	58.52	58.75	60.67	61.90	62.74	62.32	62.64	66.88	70.02	72.91	74.56	62.01	+7.80
CIFAR-100	A	28.90	34.28	37.35	39.13	41.15	42.65	43.62	44.48	45.07	45.40	49.50	52.28	52.54	53.47	43.56	-
	C	12.66	18.35	23.76	26.92	29.12	32.23	34.21	35.71	37.18	38.25	45.67	49.60	52.36	53.47	34.96	0
	Ours	26.34	29.71	31.74	32.95	34.49	36.36	38.49	39.59	40.43	41.35	46.06	49.40	51.72	53.67	39.45	+4.49

(b) Results of SVHN, CIFAR-10, CIFAR-100 targeting IPC_{50}.

Dataset		1	2	3	4	5	6	7	8	9	10	15	20	Avg.	Diff.
ImageNet-10	A	60.40	63.87	67.40	68.80	71.33	70.60	70.47	71.93	72.87	72.80†	75.50	76.60†	70.21	-
	B	60.40	62.07	62.80	63.40	64.67	63.13	62.67	63.60	64.13	63.60	62.73	64.13	63.11	-
	C	44.00	57.27	62.80	66.13	64.33	69.47	69.53	70.53	71.73	73.00	74.47	75.73	66.58	0
	Ours	55.87	61.60	63.40	64.40	63.80	67.73	67.13	70.07	71.07	71.13	76.00	79.20	67.62	+1.04

(c) Results of ImageNet-10 targeting IPC_{20}.

Multisize Dataset Condensation

- 与SOTA的比较

	DC	DSA	MTT	IDC	DREAM	Ours
1	15.35	16.76	18.80	27.49	32.52	49.55
2	19.75	21.22	24.90	38.50	39.57	53.75
3	22.54	26.78	31.90	45.29	48.21	56.39
4	26.28	30.18	38.10	50.85	53.84	59.33
5	30.37	33.43	43.20	53.60	55.25	58.13
6	33.99	38.15	49.20	57.98	60.46	60.62
7	36.36	41.18	51.60	60.99	63.27	62.06
8	39.83	45.37	56.30	63.60	65.04	63.59
9	42.68	49.21	58.50	65.71	67.40	65.25
10	44.90†	52.10†	62.80†	67.50†	69.40†	66.79
Avg.	31.21	35.44	43.53	53.15	55.50	**59.55**
Diff.	-28.34	-24.11	-16.02	-6.40	-4.05	-

(a) CIFAR-10, IPC_{10}.

	DC	DSA	MTT	IDC	DREAM	Ours
1	16.32	12.50	15.13	27.87	27.57	47.83
2	18.77	15.19	23.92	35.69	36.57	52.18
3	21.24	19.69	26.53	41.93	43.50	56.29
4	21.42	22.02	30.30	45.29	47.35	58.52
5	23.32	23.28	32.71	47.54	49.81	58.75
6	23.63	24.79	35.54	51.96	53.38	60.67
7	25.35	25.62	34.12	53.51	54.58	61.90
8	27.40	27.84	40.60	55.59	56.78	62.74
9	27.93	29.57	43.43	56.62	58.91	62.32
10	28.00	32.51	45.99	58.26	60.10	62.64
20	36.53	40.94	60.41	66.77	68.07	66.88
30	42.82	48.05	67.68	70.50	70.48	70.02
40	48.90	54.24	69.71	72.98	72.79	72.91
50	53.90†	60.60†	71.60†	74.50†	74.80†	74.56
Avg.	29.68	31.20	42.69	54.21	55.33	**62.01**
Diff.	-32.33	-30.81	-19.32	-7.80	-6.68	-

(b) CIFAR-10, IPC_{50}