NEURAL INFORMATION PROCESSING SYSTEMS

Frequency Domain-based Dataset Distillation

Donghyeok Shin*[1], Seungjae Shin*[1], and Il-Chul Moon[1,2]

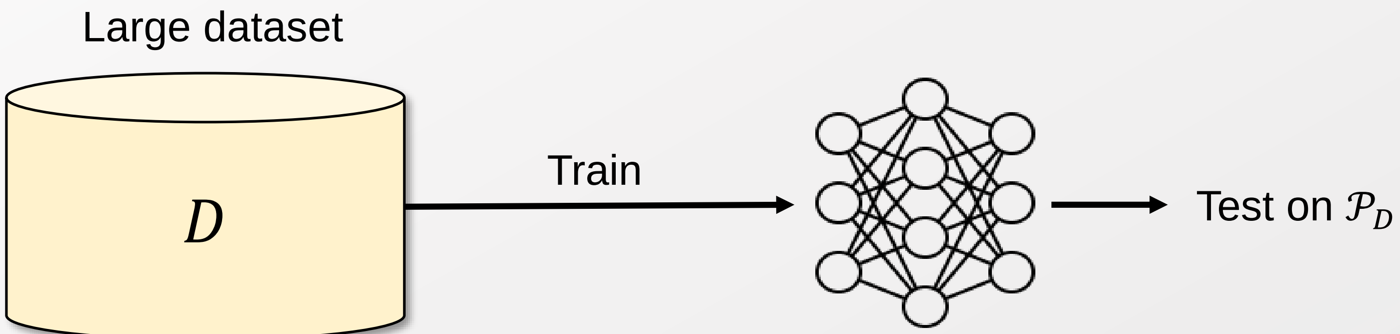
* Equal contribution

[1]KAIST, [2]Summary.AI

NeurIPS 2023

Introduction Large-scale dataset

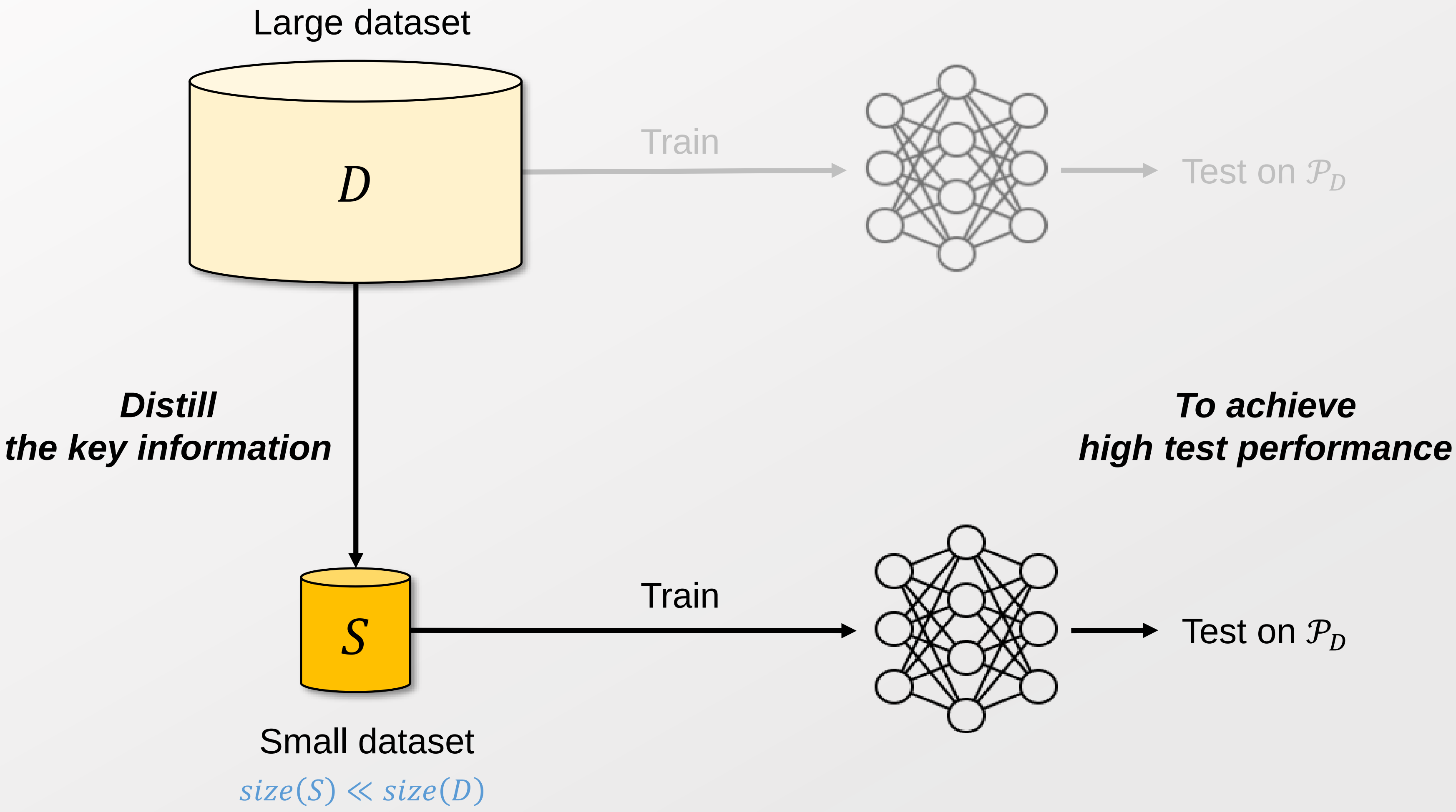
- Handling large-scale datasets is crucial for high performance, but it is accompanied by many burdens.



- Large data storage
- Expensive computation cost
- High energy usage

Introduction Dataset Distillation

- Dataset distillation aims at synthesizing a small-size dataset which can achieve the high test performance.

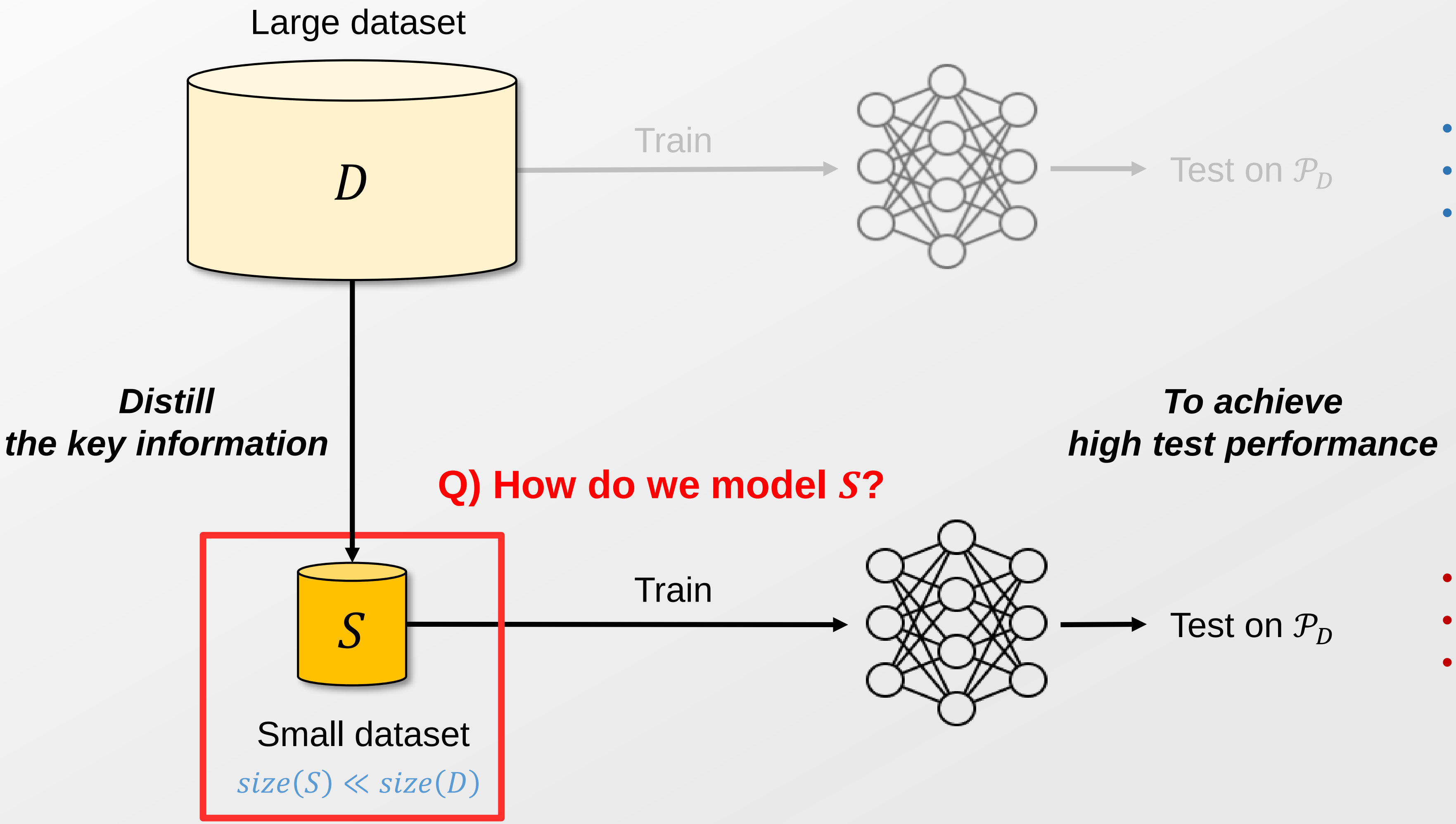


- Large data storage
- Expensive computation cost
- High energy usage

- Small data storage
- Cheap computation cost
- Low energy usage

Introduction Dataset Distillation

- Dataset distillation aims at synthesizing a small-size dataset which can achieve the high test performance.



- Large data storage
- Expensive computation cost
- High energy usage

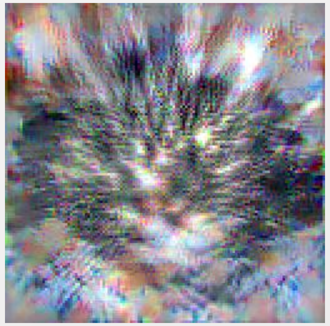
- Small data storage
- Cheap computation cost
- Low energy usage

Q) How do we model S?

Input-sized variable

: Same size as the original instance

Distill →



$C \times H \times W$ $\quad$ $C \times H \times W$

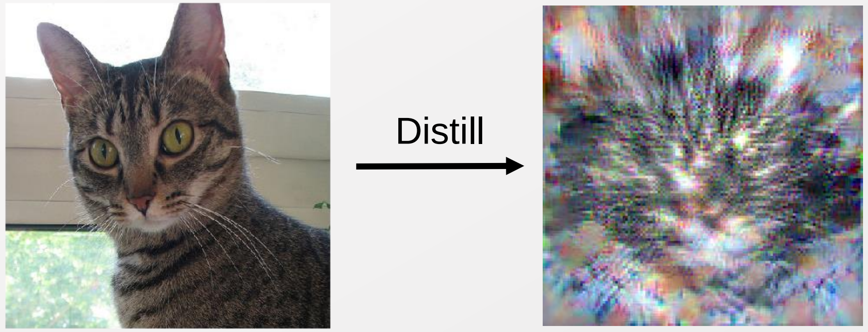
- Difficulty in specifying the importance of spatial dimension
  ➔ Superfluous dimensions are included.

Cazenavette et. al., Dataset distillation by matching training trajectories, CVPR, 2022

Introduction Previous Researches

Q) How do we model S?

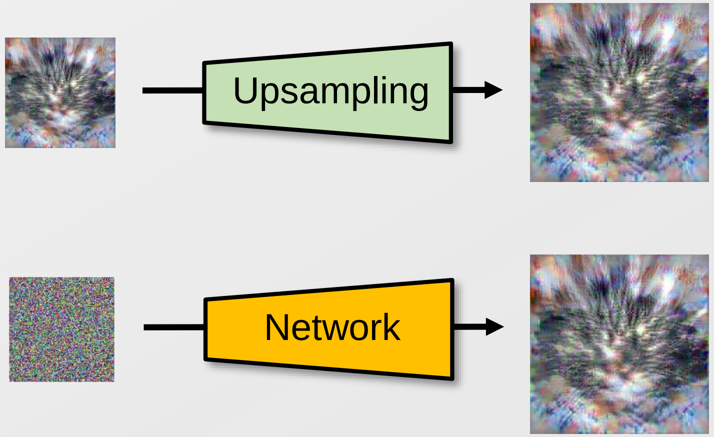
Input-sized variable
: Same size as the original instance



$C \times H \times W$ $C \times H \times W$

- Difficulty in specifying the importance of spatial dimension
  ➔ Superfluous dimensions are included.

Spatial-based Parameterization
: Cooperation between variable and transformation



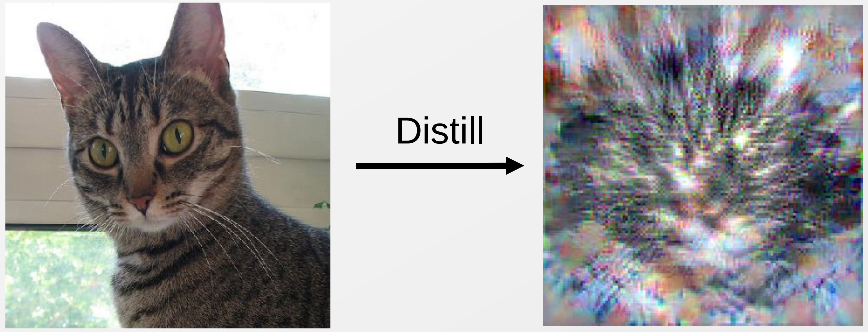
- Distortion in the spectral distribution of natural images
- Require training with an auxiliary network

Cazenavette et. al., Dataset distillation by matching training trajectories, CVPR, 2022
Kim et. al., Dataset condensation via efficient synthetic-data parameterization, ICML, 2022
Liu et. al., Dataset distillation via Factorization, NeurIPS, 2022

Q) How do we model S?

Input-sized variable

: Same size as the original instance

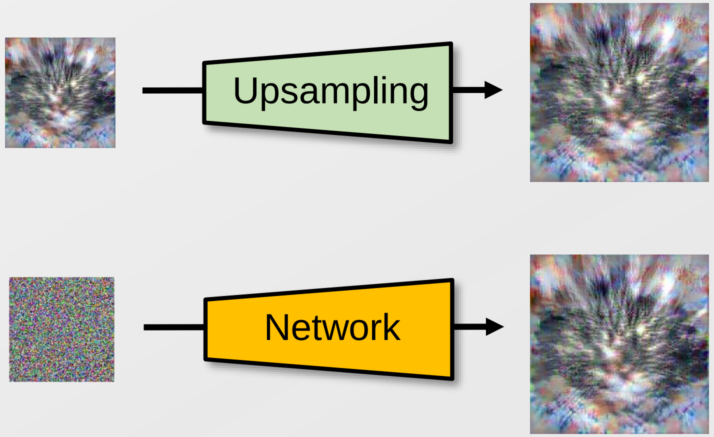


$C \times H \times W$ $\quad$ $C \times H \times W$

- Difficulty in specifying the importance of spatial dimension
  ➔ Superfluous dimensions are included.

Spatial-based Parameterization

: Cooperation between variable and transformation



- Distortion in the spectral distribution of natural images
- Require training with an auxiliary network

Utilize the frequency domain dimensions which are crucial for instance and dataset formation.

Cazenavette et. al., Dataset distillation by matching training trajectories, CVPR, 2022
Kim et. al., Dataset condensation via efficient synthetic-data parameterization, ICML, 2022
Liu et. al., Dataset distillation via Factorization, NeurIPS, 2022

Spatial domain

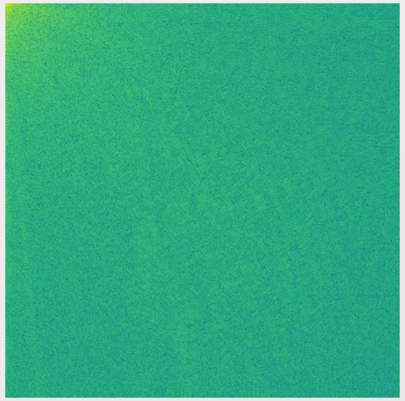
: deal with the pixel intensity

Frequency domain

: deal with the rate of pixel intensity changing

Frequency Transform $\mathcal{F}$

Inverse Frequency Transform $\mathcal{F}^{-1}$



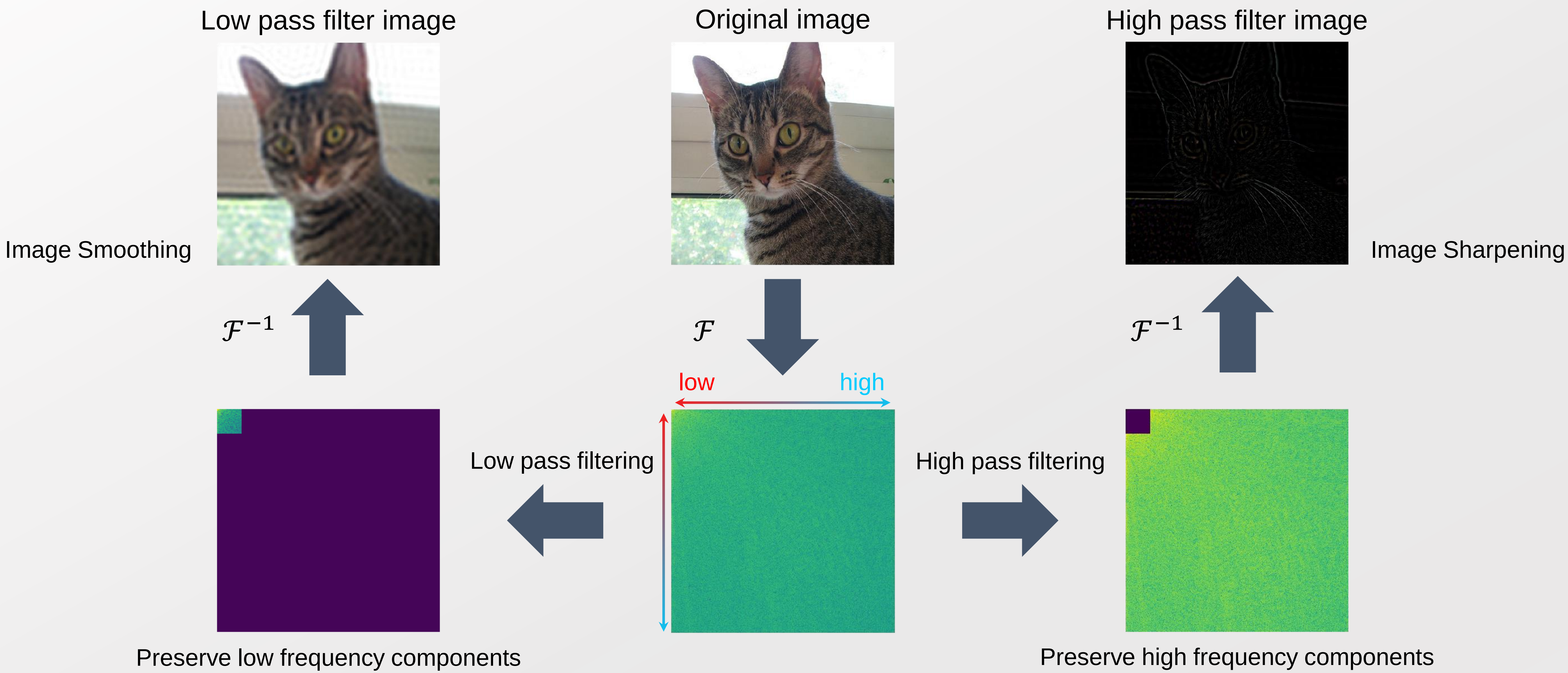
- (Inverse) Frequency transform is…
  - Differentiable
  - Static (No require an additional parameters)
  - Efficient and fast operation

Appropriate properties for dataset distillation

We used DCT as a default for the examples.
For better visualization, we show the single channel frequency representation.

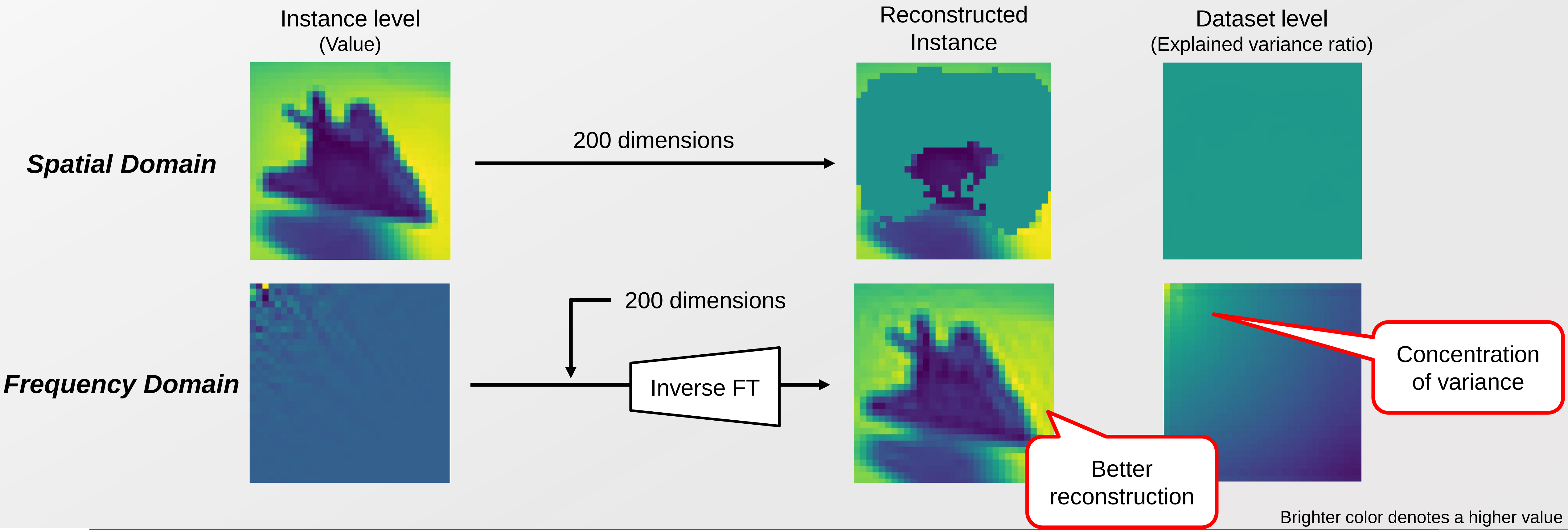
Preliminary Frequency Domain

- Each frequency dimension has characteristic.



Energy compaction property of frequency domain improves the efficiency of dataset distillation while preserves the essential information.

- Frequency domain compacts the **spatial domain information** on a few frequency dimensions.

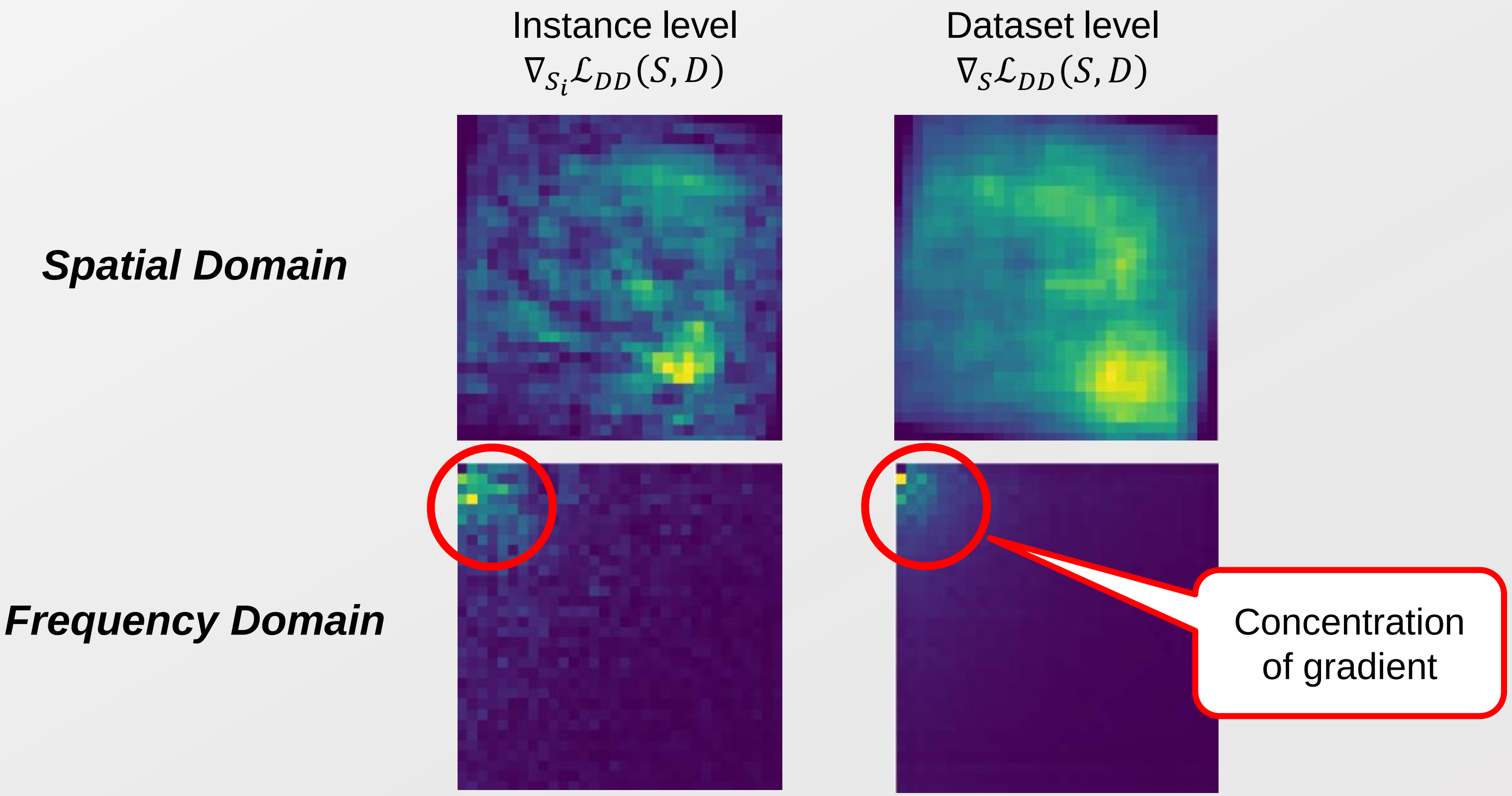


Brighter color denotes a higher value

Methodology Motivation

Energy compaction property of frequency domain improves the efficiency of dataset distillation while preserves the essential information.

- Frequency domain compacts the spatial domain information on a few frequency dimensions.
- Only a few specific frequency dimensions are required for the **gradient** update in dataset distillation.



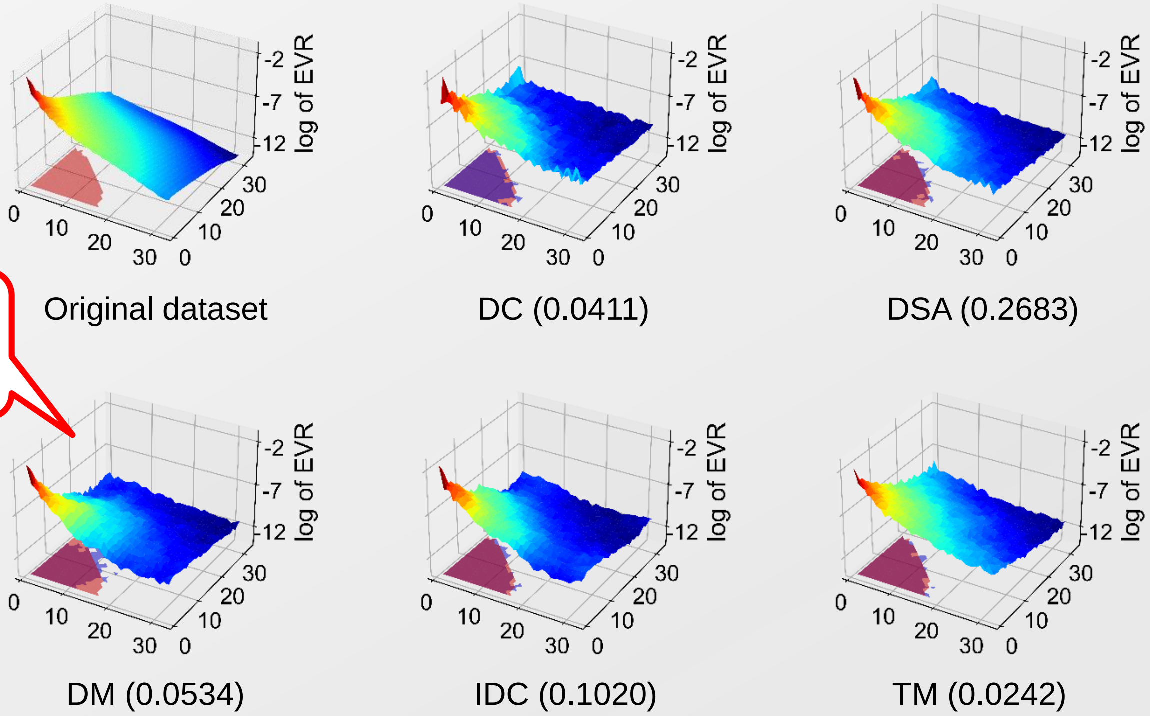
Brighter color denotes a higher value

Methodology Motivation

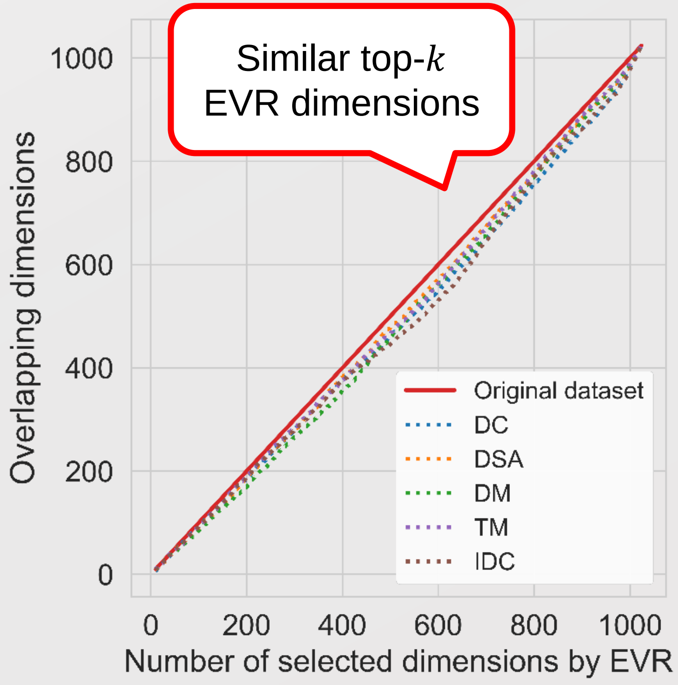
Energy compaction property of frequency domain improves the efficiency of dataset distillation while preserves the essential information.

- Frequency domain compacts the spatial domain information on a few frequency dimensions.
- Only a few specific frequency dimensions are required for the gradient update in dataset distillation.
- Trained synthetic dataset has similar **explained variance ratio features of original dataset**.

Similar EVR distribution



Similar top-k EVR dimensions



Methodology Motivation

Energy compaction property of frequency domain improves the efficiency of dataset distillation while preserves the essential information.

- Frequency domain compacts the spatial domain information on a few frequency dimensions.
- Only a few specific frequency dimensions are required for the gradient update in dataset distillation.
- Trained synthetic dataset has similar explained variance ratio features of original dataset.

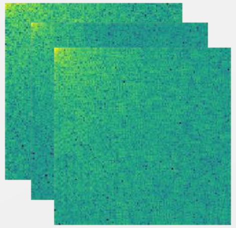
↓

Utilize the frequency domain dimensions which are crucial for instance and dataset formation.

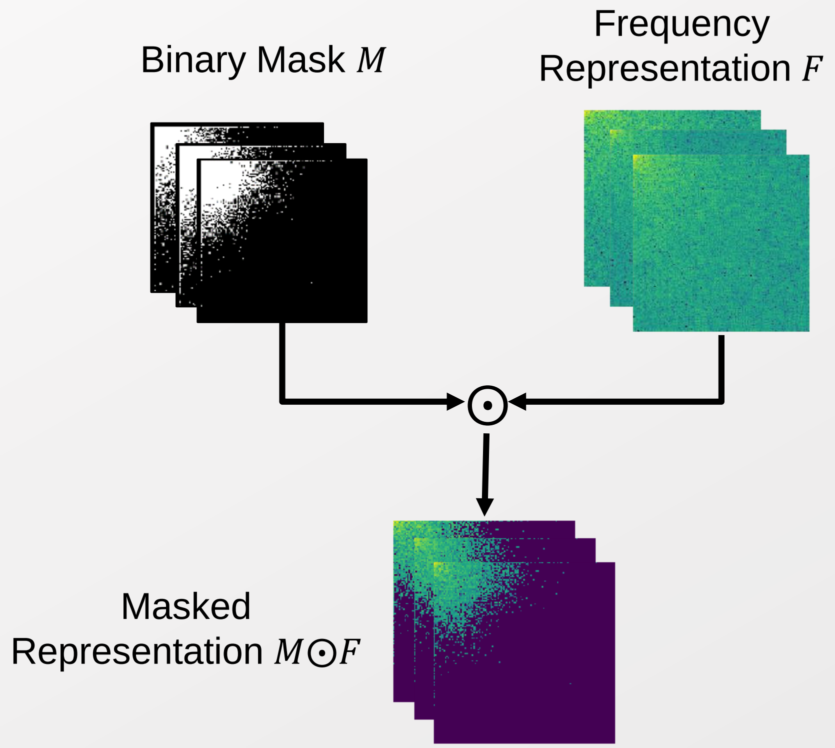
- Retain the task-relevant information of original dataset.
- Reduce the necessary budgets for each instance.
  ➔ The remaining budget can be utilized to increase the number of distilled instances.

- FreD consists of three main components:
  1. Synthetic Frequency Memory F
     - Optimization target
     - Initialized through the frequency representation of randomly sampled

Frequency
Representation F

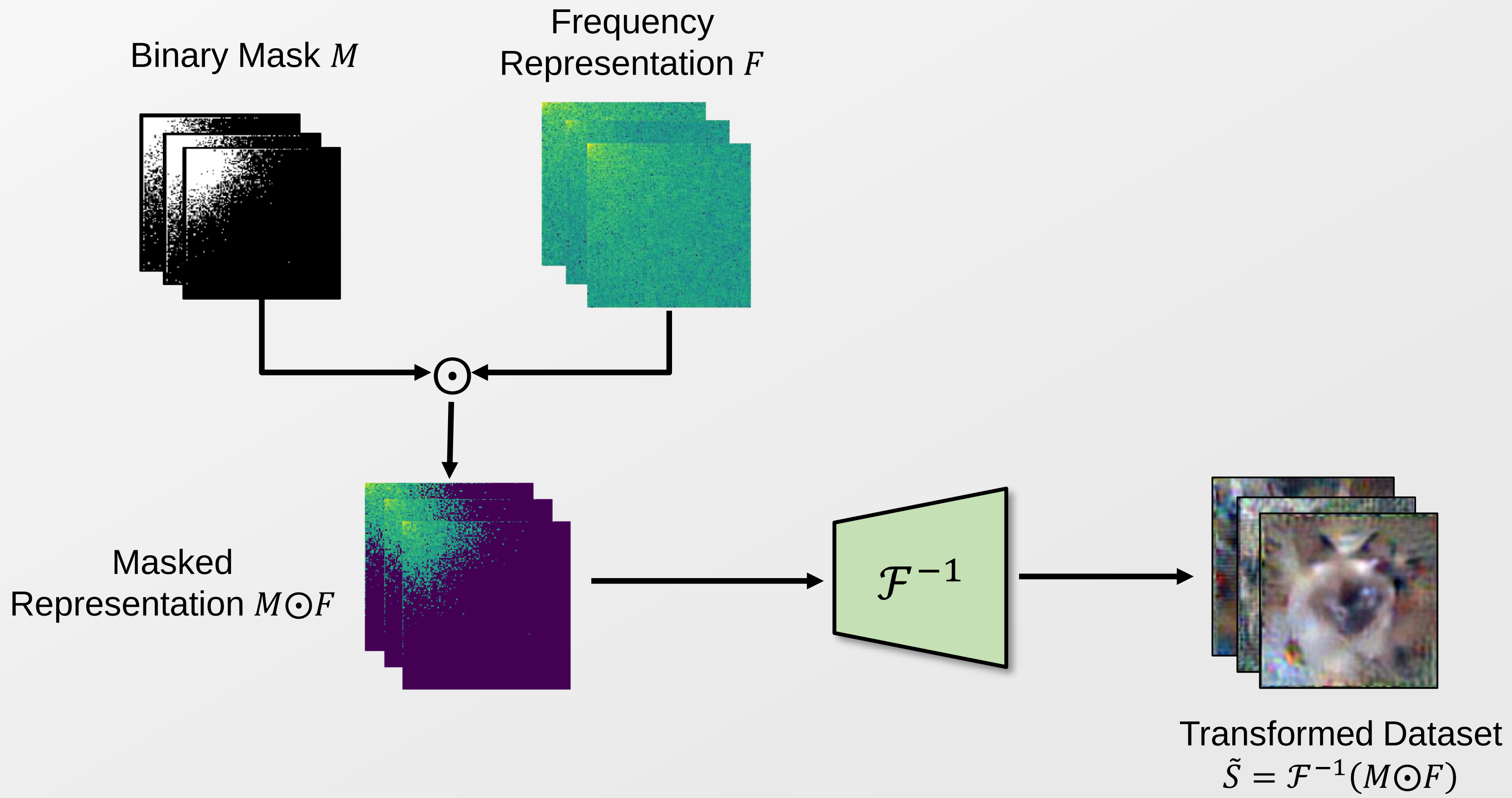


- FreD consists of three main components:
  1. Synthetic Frequency Memory F
  2. Binary Mask Memory M
     - To filter out uninformative dimensions in the frequency domain
     - Utilize the top-k dimensions based on Explained Variance Ratio (EVR) to maximally preserve the class-wise variance



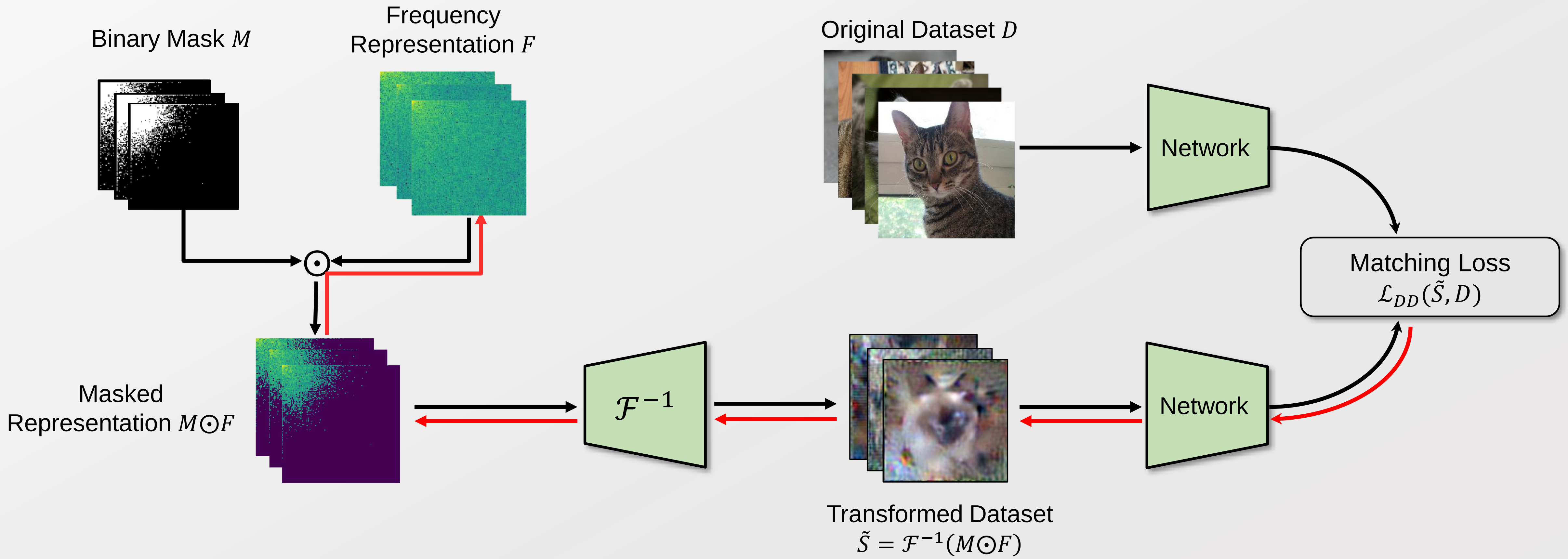
Methodology Frequency Domain-based Dataset Distillation (FreD)

- FreD consists of three main components:
  1. Synthetic Frequency Memory F
  2. Binary Mask Memory M
  3. Inverse Frequency Transform $\mathcal{F}^{-1}$
     - To transform the inferred frequency representation to the corresponding instance on the spatial domain



- FreD consists of three main components:
  1. Synthetic Frequency Memory F
  2. Binary Mask Memory M
  3. Inverse Frequency Transform $\mathcal{F}^{-1}$

$$F^* = \underset{F}{\operatorname{argmin}}\, \mathcal{L}_{DD}(\tilde{S}, D) \text{ where } \tilde{S} = \mathcal{F}^{-1}(M \odot F) \qquad \text{(Training)}$$

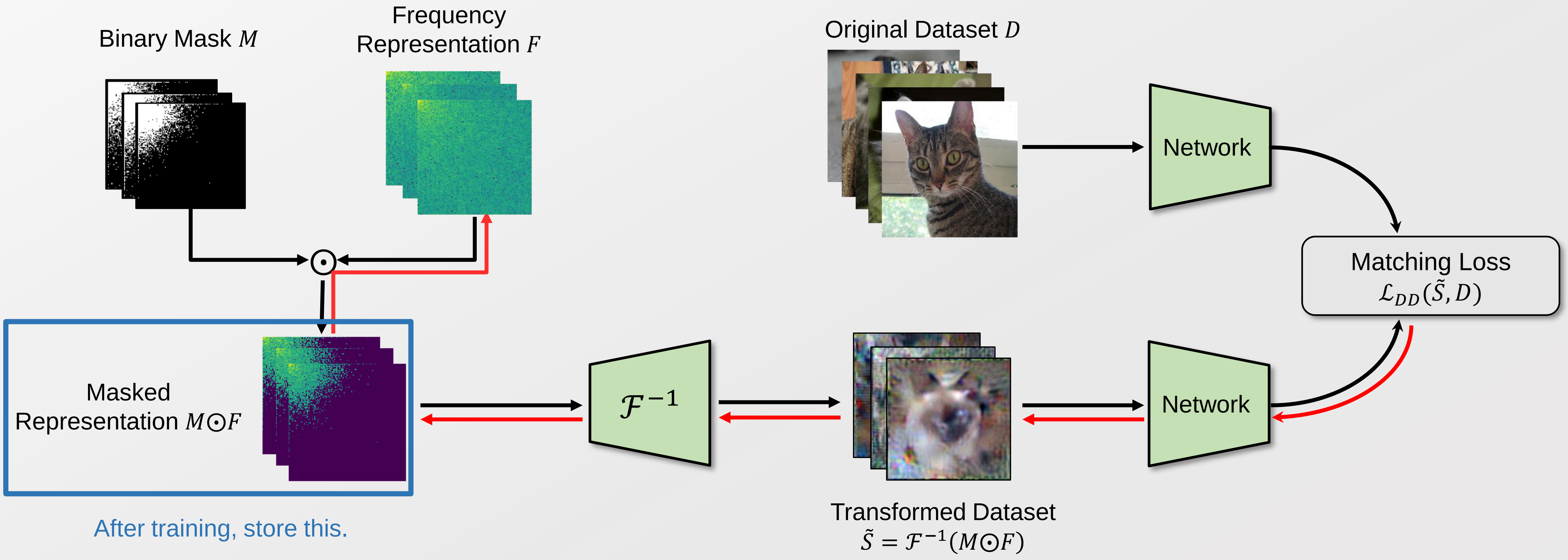


Frequency Domain-based Dataset Distillation (FreD)

- FreD consists of three main components:
  1. Synthetic Frequency Memory F
  2. Binary Mask Memory M
  3. Inverse Frequency Transform $\mathcal{F}^{-1}$

$$F^* = \underset{F}{\operatorname{argmin}}\, \mathcal{L}_{DD}(\tilde{S}, D) \text{ where } \tilde{S} = \mathcal{F}^{-1}(M \odot F) \qquad \text{(Training)}$$

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\tilde{S}^*; \theta) \text{ where } \tilde{S}^* = \mathcal{F}^{-1}(M \odot F^*) \qquad \text{(Evaluation)}$$



APPLIED ARTIFICIAL INTELLIGENCE LAB

Test accuracies (%) on SVHN, CIFAR-10, CIFAR-100

		SVHN			CIFAR-10			CIFAR-100		
	IPC	1	10	50	1	10	50	1	10	50
	#Params	30.72k	307.2k	1536k	30.72k	307.2k	1536k	30.72k	307.2k	1536k
Coreset	Random	14.6 ±1.6	35.1 ±4.1	70.9 ±0.9	14.4 ±2.0	26.0 ±1.2	43.4 ±1.0	4.2 ±0.3	14.6 ±0.5	30.0 ±0.4
	Herding	20.9 ±1.3	50.5 ±3.3	72.6 ±0.8	21.5 ±1.3	31.6 ±0.7	40.4 ±0.6	8.4 ±0.3	17.3 ±0.3	33.7 ±0.5
Input-sized parameterization	DC	31.2 ±1.4	76.1 ±0.6	82.3 ±0.3	28.3 ±0.5	44.9 ±0.5	53.9 ±0.5	12.8 ±0.3	25.2 ±0.3	-
	DSA	27.5 ±1.4	79.2 ±0.5	84.4 ±0.4	28.8 ±0.7	52.1 ±0.5	60.6 ±0.5	13.9 ±0.3	32.3 ±0.3	42.8 ±0.4
	DM	-	-	-	26.0 ±0.8	48.9 ±0.6	63.0 ±0.4	11.4 ±0.2	29.7 ±0.2	43.6 ±0.4
	CAFE+DSA	42.9 ±3.0	77.9 ±0.6	82.3 ±0.4	31.6 ±0.8	50.9 ±0.5	62.3 ±0.4	14.0 ±0.2	31.5 ±0.2	42.9 ±0.2
	TM	58.5 ±1.4	70.8 ±1.8	85.7 ±0.1	46.3 ±0.8	65.3 ±0.7	71.6 ±0.2	24.3 ±0.2	40.1 ±0.4	47.7 ±0.2
	KIP	57.3 ±0.1	75.0 ±0.1	80.5 ±0.1	49.9 ±0.2	62.7 ±0.3	68.6 ±0.2	15.7 ±0.2	28.3 ±0.1	-
	FRePo	-	-	-	46.8 ±0.7	65.5 ±0.4	71.7 ±0.2	28.7 ±0.1	42.5 ±0.2	44.3 ±0.2
Parameterization	IDC	68.1 ±0.1	87.3 ±0.2	90.2 ±0.1	50.0 ±0.4	67.5 ±0.5	74.5 ±0.1	-	-	-
	HaBa	69.8 ±1.3	83.2 ±0.4	88.3 ±0.1	48.3 ±0.8	69.9 ±0.4	74.0 ±0.2	33.4 ±0.4	40.2 ±0.2	47.0 ±0.2
	FreD	**82.2** ±0.6	**89.5** ±0.1	**90.3** ±0.3	**60.6** ±0.8	**70.3** ±0.3	**75.8** ±0.1	**34.6** ±0.4	**42.7** ±0.2	**47.8** ±0.1
Entire original dataset			95.4 ±0.1			84.8 ±0.1			56.2 ±0.3	
Increment of decoded instances	IDC	×5	×5	×5	×5	×5	×5	-	-	-
	HaBa	×5	×5	×5	×5	×5	×5	×5	×5	×5
	FreD	×16	×8	×4	×16	×6.4	×4	×8	×2.56	×2.56

- **Best result**
- Second-best result

- **High performance** on various benchmark datasets.
- Significant performance gap when the **limited budget is extreme**.
  - 12.4%p in SVHN / 10.6%p in CIFAR-10

High performance

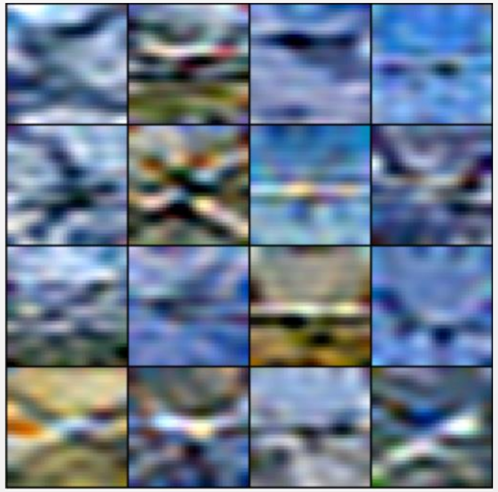
Experimental Results Qualitative Analysis

Class: Airplane

Binary Mask

Baseline (TM)

Ours (FreD)

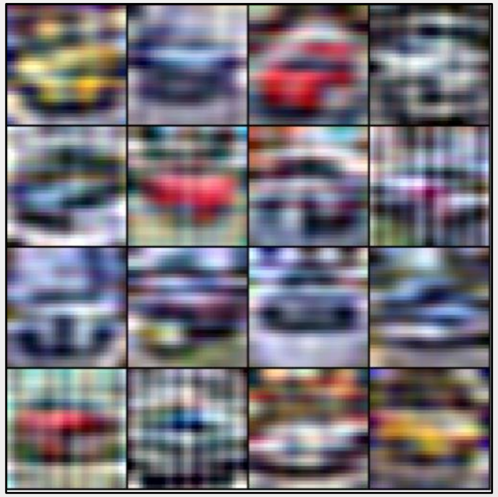


Class: Automobile

Binary Mask

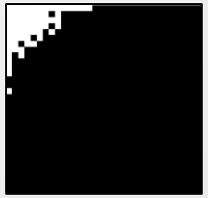
Baseline (TM)

Ours (FreD)



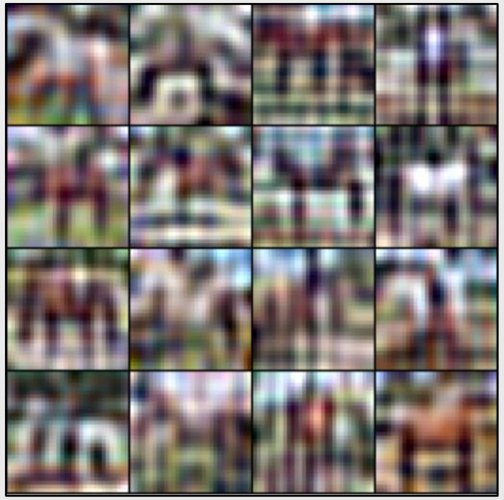
Class: Horse

Binary Mask



Baseline (TM)

Ours (FreD)



- **More instances** under the same budget.
  - Up to 16x more instances
- Intra-class diversity and inter-class discriminative features.
- Focus on low-frequency but slightly different binary mask for each class.

Efficient budget utilization

Experimental Results Compatibility of Parameterization

APPLIED ARTIFICIAL INTELLIGENCE LAB

- Parameterization methods should show consistent performance improvement across different distillation loss and test network architectures.

Test accuracies (%) on CIFAR-10

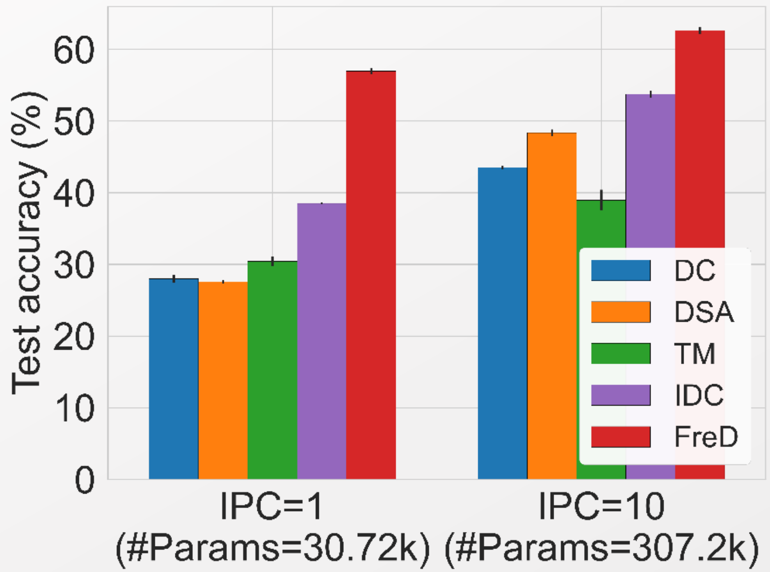
DC/DM/TM: gradient/feature/trajectory matching

		DC			DM			TM		
	IPC #Params	2 61.44k	11 337.92k	51 1566.72k	2 61.44k	11 337.92k	51 1566.72k	2 61.44k	11 337.92k	51 1566.72k
ConvNet	Vanilla	31.4 ±0.2	45.3 ±0.3	54.2 ±0.6	34.6 ±0.5	50.4 ±0.4	62.0 ±0.3	50.6 ±1.0	63.9 ±0.3	69.8 ±0.5
	w/ IDC	<u>35.2</u> ±0.5	<u>53.8</u> ±0.4	56.4 ±0.4	<u>45.1</u> ±0.5	<u>59.3</u> ±0.4	<u>64.6</u> ±0.3	56.1 ±0.4	60.9 ±0.4	71.1 ±0.4
	w/ HaBa	34.1 ±0.5	49.9 ±0.5	<u>58.9</u> ±0.2	37.3 ±0.1	56.8 ±0.1	64.4 ±0.4	<u>56.8</u> ±0.4	<u>69.5</u> ±0.3	<u>73.3</u> ±0.2
	w/ FreD	**45.3** ±0.5	**55.8** ±0.4	**59.8** ±0.5	**55.9** ±0.4	**61.3** ±0.8	**66.6** ±0.6	**61.4** ±0.3	**70.7** ±0.5	**75.5** ±0.2
Average of Cross-Architectures (AlexNet, VGG11, ResNet18)	Vanilla	22.0 ±0.9	29.2 ±0.9	34.1 ±0.6	21.5 ±2.2	39.5 ±1.1	52.6 ±0.7	33.1 ±1.1	43.9 ±1.4	55.0 ±1.0
	w/ IDC	<u>28.7</u> ±1.2	<u>35.4</u> ±0.6	<u>40.2</u> ±0.7	<u>37.3</u> ±1.1	<u>50.5</u> ±0.6	<u>61.3</u> ±0.5	42.5 ±1.5	48.7 ±1.8	61.5 ±1.0
	w/ HaBa	25.4 ±0.9	31.4 ±0.7	35.5 ±0.9	30.1 ±0.6	47.0 ±0.5	60.1 ±0.6	<u>46.4</u> ±1.0	<u>55.8</u> ±1.8	<u>64.0</u> ±0.9
	w/ FreD	**37.3** ±0.9	**37.4** ±0.7	**42.7** ±0.8	**48.1** ±0.7	**57.3** ±0.8	**65.0** ±0.7	**49.7** ±1.0	**60.1** ±0.7	**69.1** ±0.7

- **Highest performance improvement** for all experimental combinations.
  - From 1.2%p to 10.8%p in training architecture.
  - From 2.0%p to 10.6%p in cross-architecture generalization.

Better compatibility

Test dataset: CIFAR-10-C



Test accuracies (%) based on different corruptions in CIFAR-10-C

	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr	Elastic	Pixel	JPEG	Avg.
DC	28.2	28.3	28.0	28.5	28.3	28.2	27.8	28.0	28.2	24.3	28.4	28.6	28.1	28.5	28.5	28.0
DSA	27.8	27.8	27.5	28.1	27.8	27.8	27.3	27.6	27.8	23.7	27.9	28.8	27.6	28.2	28.0	27.6
TM	30.8	31.1	29.9	30.5	29.0	29.5	29.6	31.0	30.5	28.0	32.3	32.4	29.5	31.1	31.5	30.4
IDC	37.4	37.8	36.3	39.7	38.2	39.0	39.0	38.9	38.6	35.7	39.3	40.4	38.5	39.4	39.1	38.5
FreD	**56.7**	**57.3**	**54.4**	**58.9**	**56.4**	**57.2**	**57.3**	**58.0**	**56.5**	**53.6**	**59.2**	**53.5**	**57.2**	**59.6**	**58.9**	**57.0**

Test dataset: ImageNet-Subset-C

#Params	Model	ImgNette-C	ImgWoof-C	ImgFruit-C	ImgYellow-C	ImgMeow-C	ImgSquawk-C
491520 (IPC=1)	TM	38.0 ±1.6	23.8 ±1.0	22.7 ±1.1	35.6 ±1.7	23.3 ±1.1	-
	w/ IDC	34.5 ±0.6	18.7 ±0.4	28.5 ±0.9	36.8 ±1.4	22.2 ±1.2	26.8 ±0.5
	w/ FreD	**51.2** ±0.6	**31.0** ±0.9	**32.3** ±1.4	**48.2** ±1.0	**30.3** ±0.3	**45.9** ±0.6
4915200 (IPC=10)	TM	50.9 ±0.7	30.9 ±0.7	32.3 ±0.8	45.6 ±1.0	30.1 ±0.5	44.4 ±1.8
	w/ IDC	40.4 ±1.0	21.9 ±0.3	32.2 ±0.7	39.6 ±0.5	23.9 ±0.8	40.5 ±0.7
	w/ FreD	**55.2** ±0.8	**33.8** ±0.8	**35.7** ±0.6	**47.9** ±0.4	**31.3** ±0.9	**52.5** ±0.8

- **Best performance** with a significant gap over the baselines.
- Superior robustness regardless of corruption type.
- Performance improvement regardless of whether the test dataset is corrupted.

Superior cross-domain generalization

Conclusion

- We demonstrate the energy compaction property of frequency domain is efficient for dataset distillation.

- We propose a new parameterization method for dataset distillation, coined FreD.
  - Select a set of frequency dimensions based on explained variance ratio.
  - Optimize the frequency representations of the selected dimensions.
  - Utilize the inverse frequency transform which is highly suitable choice for dataset distillation.

- We show the efficacy of utilizing the frequency domain for dataset distillation through the various experiments.
  - High performance on the extensive datasets.
  - Significant reduction in the required budget for synthesizing an instance.
  - Consistent performance improvement regardless the dataset distillation objective and test network architecture.
  - Superior robustness against the corruption.

  - More and detail results can be found in the paper and supplementary material.

Poster
Wed 13 Dec 10:45-12:45
Great Hall & Hall B1+B2 #518

Thank you!

Contact: tlsehdgur0@kaist.ac.kr

Code: https://github.com/sdh0818/FreD