Qimera: Data-free Quantization with Synthetic Boundary Supporting Samples

Kanghyun Choi[1], Deokki Hong[2], Noseong Park[1,2], Youngsok Kim[1,2], Jinho Lee[1,2,+]

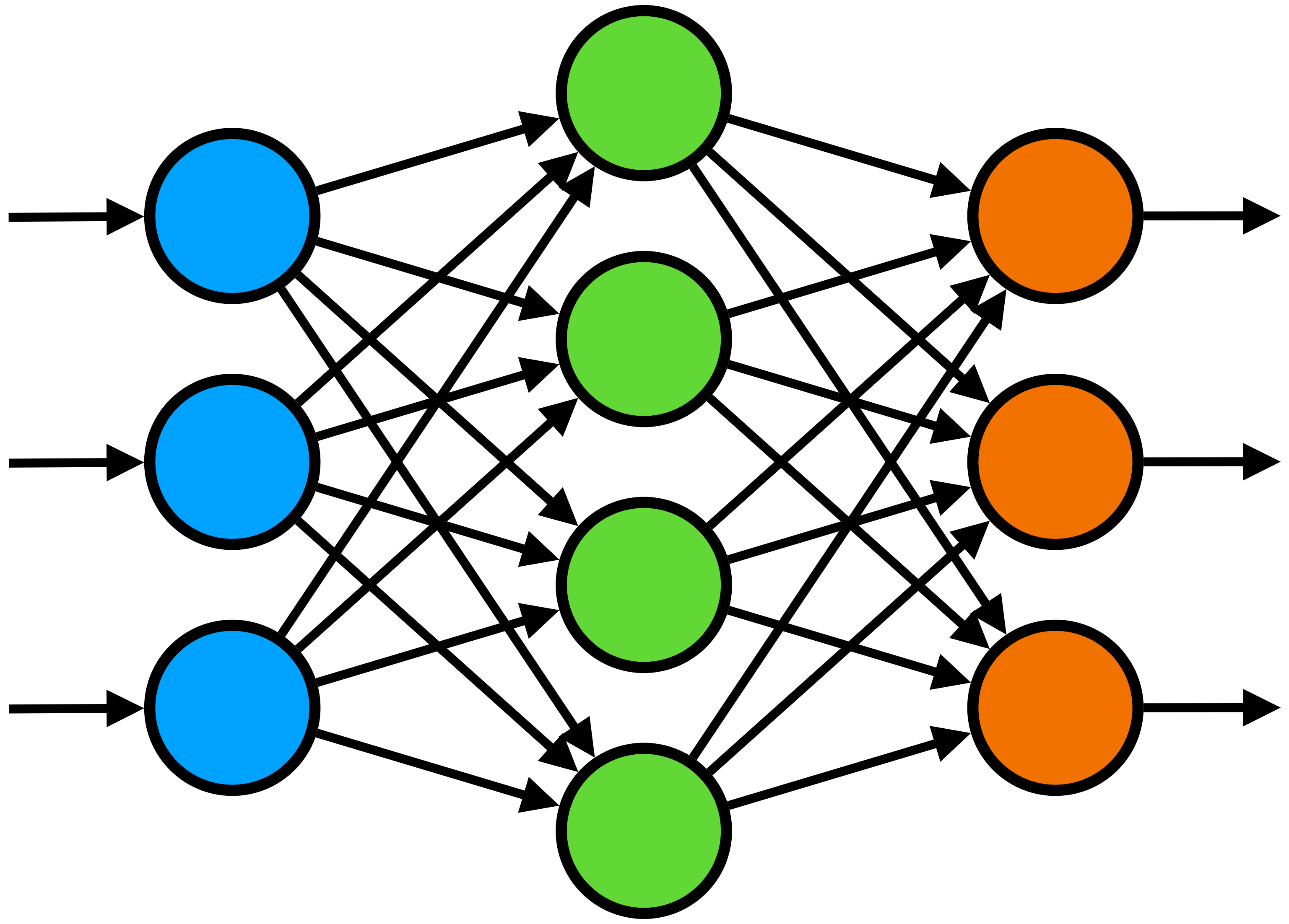
[1] Department of Computer Science, Yonsei University
[2] Department of Artificial Intelligence, Yonsei University

NeurIPS 2021 Presentation

[+] Corresponding author

Neural Network Compression

Background



Deep Neural Network

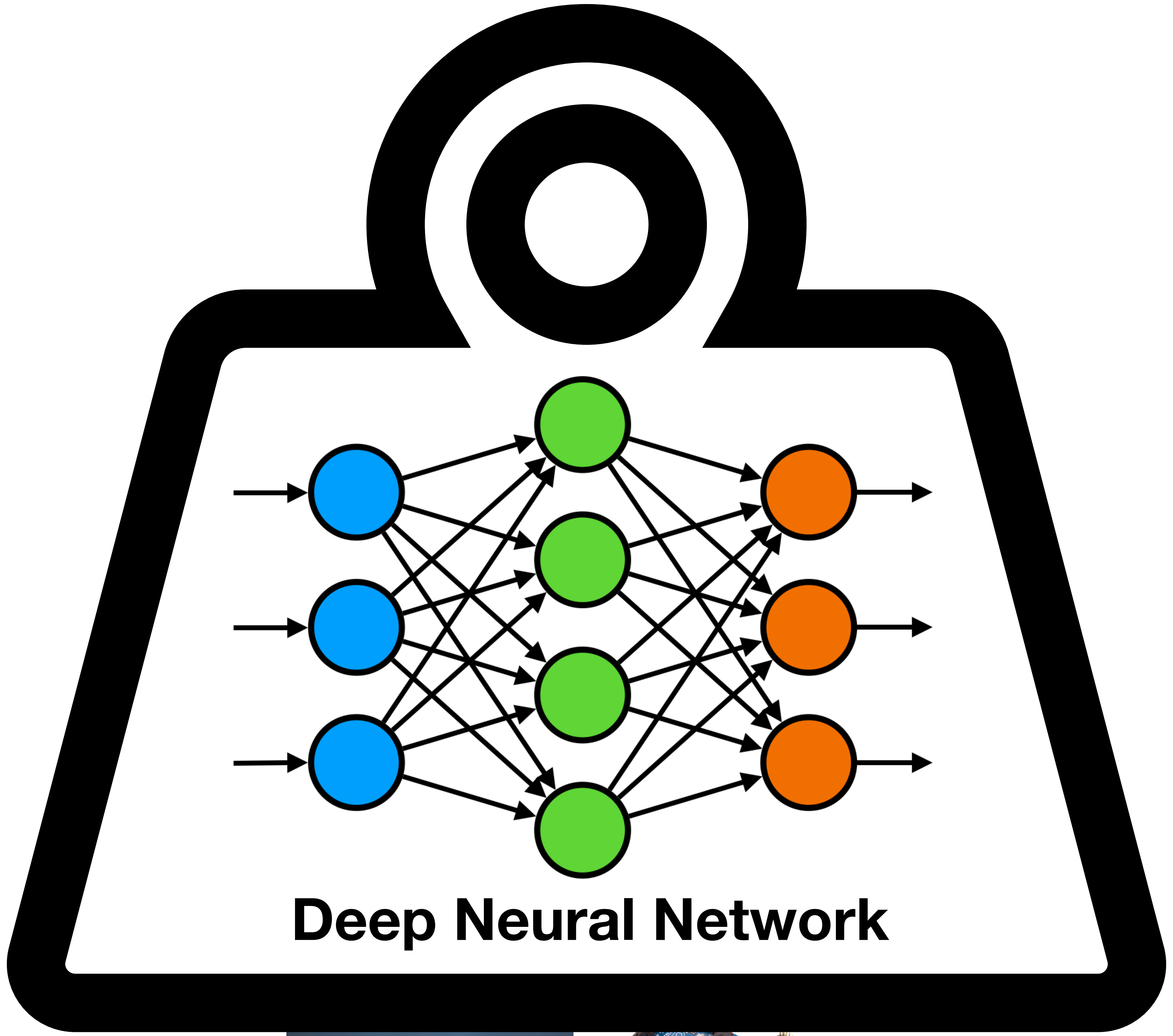
Low-power Edge Devices

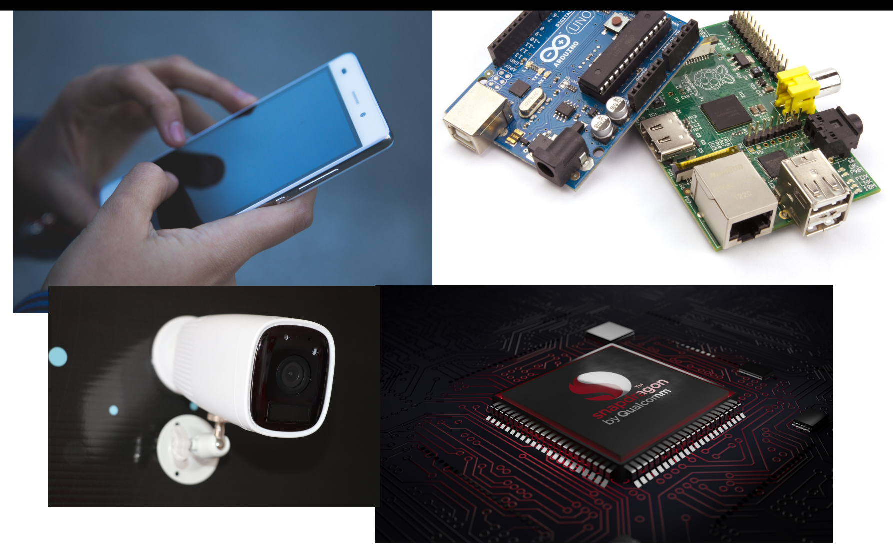
Neural Network Compression

Background

Existing DNN

Too Heavy Computation Cost





Low-power Edge Devices

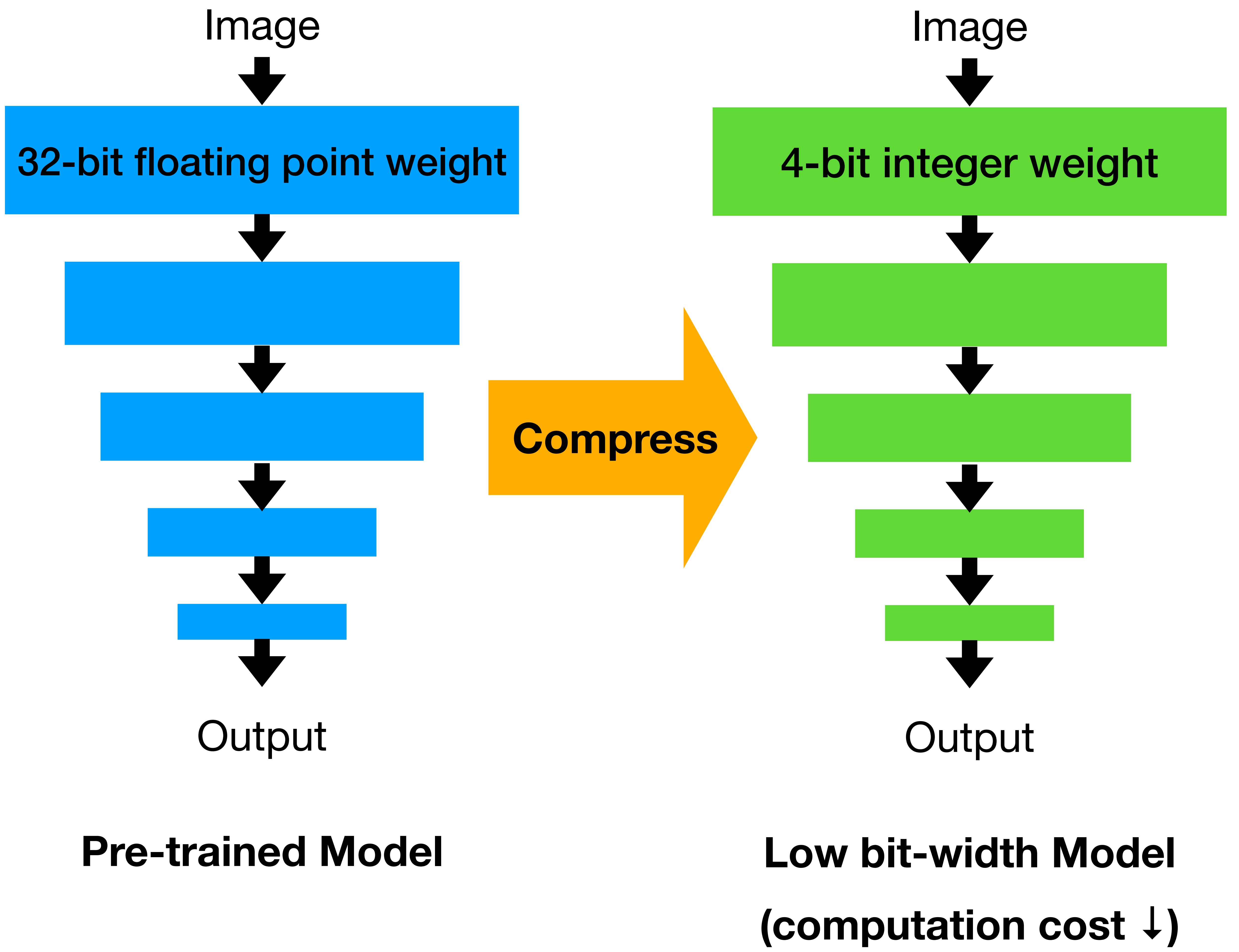
Need

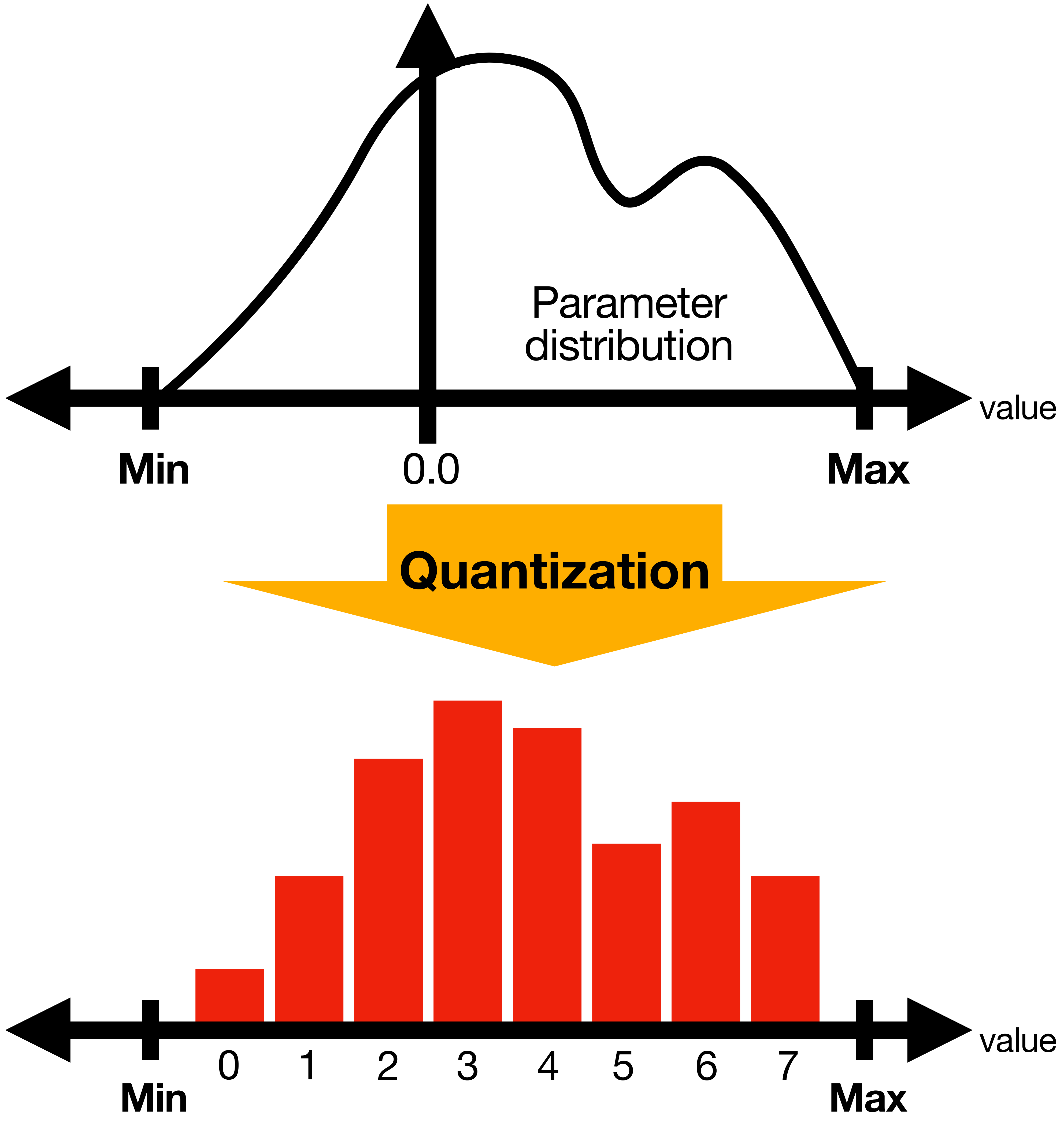
Powerful
Lightweight
Efficient

Neural Networks

Neural Network Compression

Related Work - Neural Network Quantization

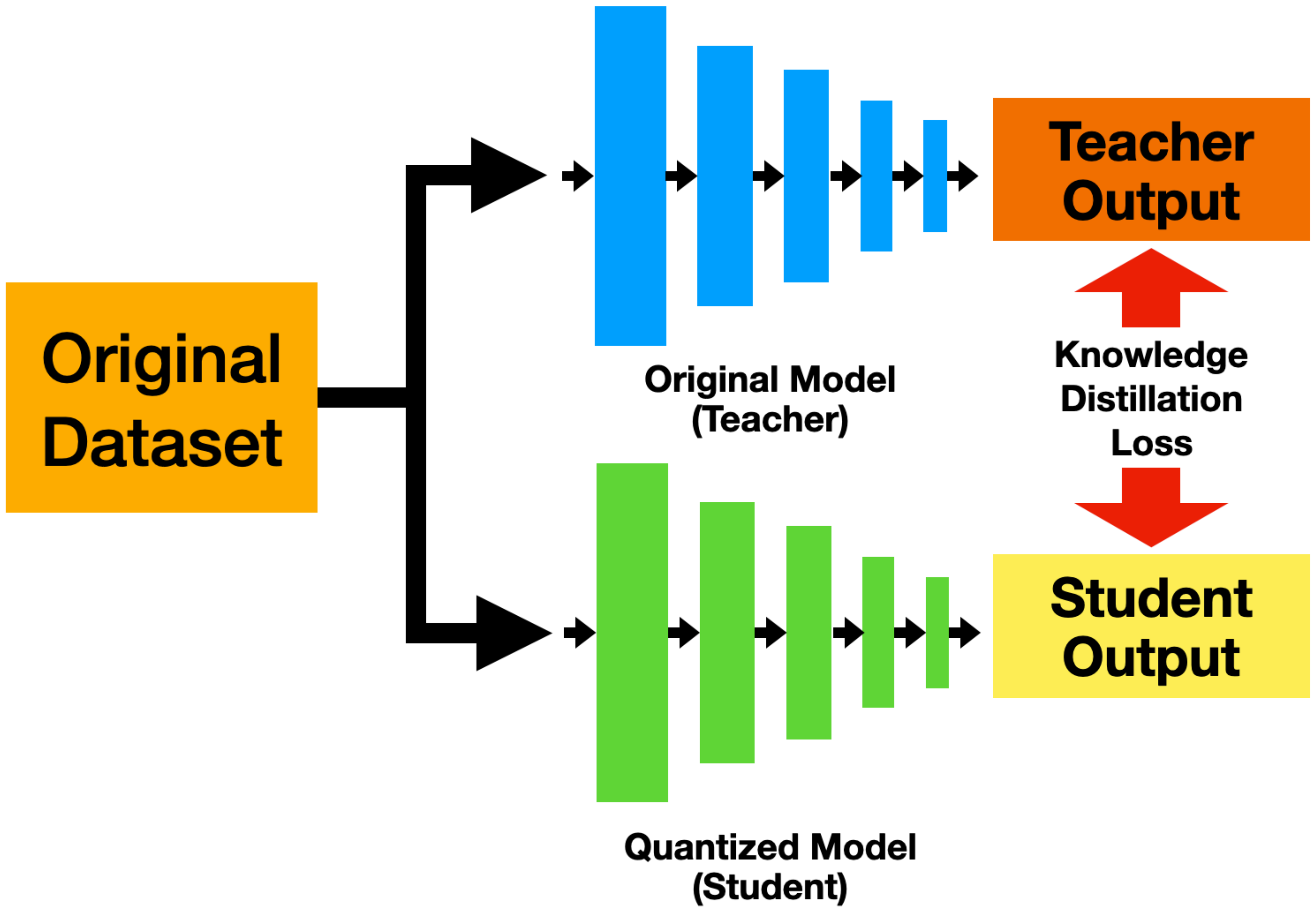




Neural Network Compression

Related Work - Neural Network Quantization

- Quantization suffers from accuracy degradation
- Use Teacher-Student knowledge distillation method to fine-tune quantized model
- Fine-tuning stage needs **the original train dataset**

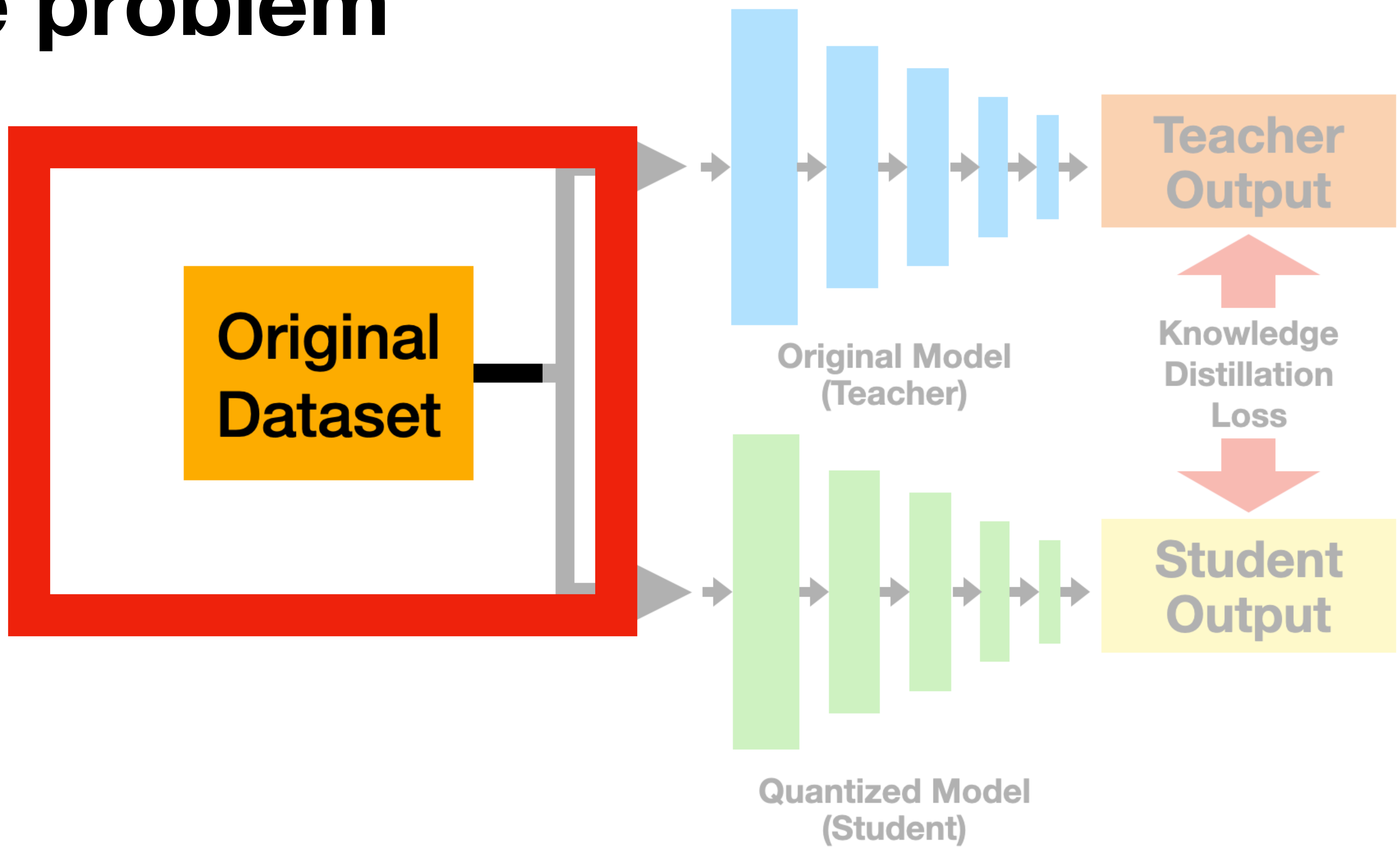


Data-free Neural Network Compression

Related Work - Data-free Compression

Original Dataset itself is the problem

- Copyright
- Privacy
- No public use
- Too large



Data-free Neural Network Compression

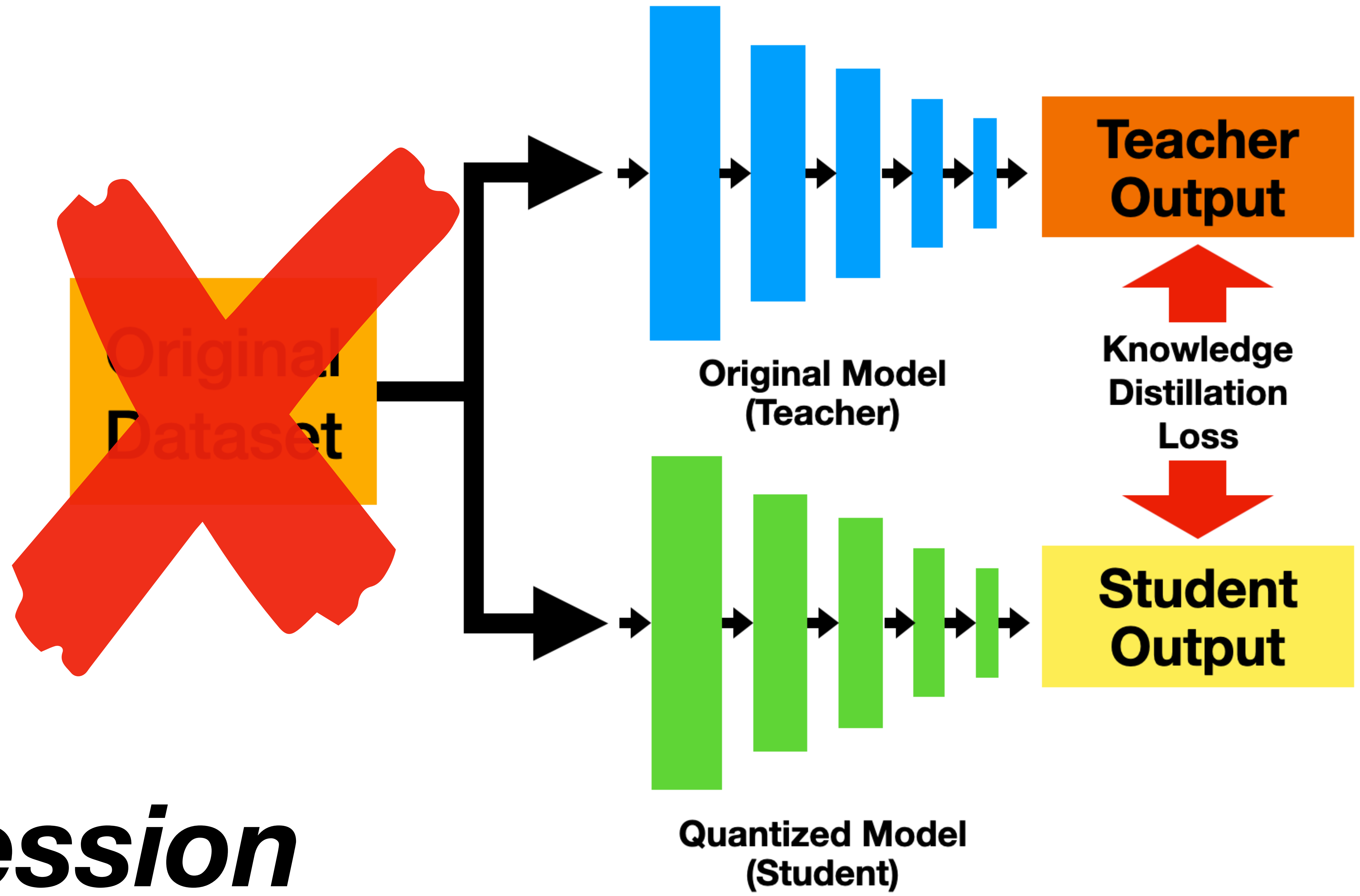
Related Work - Data-free Compression

Compression method

without original data

a.k.a.

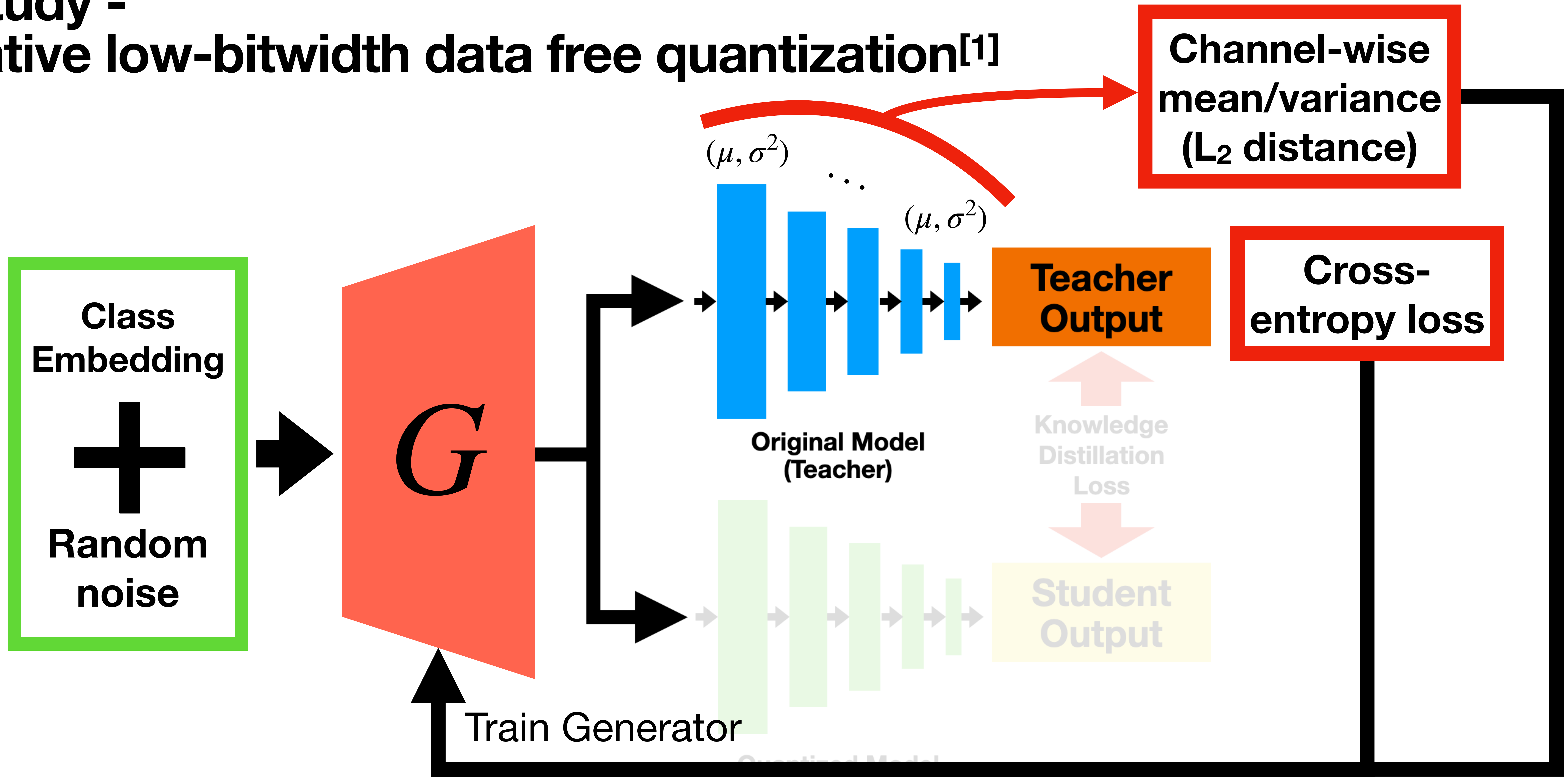
Data-free
Neural Network Compression



Data-free Neural Network Compression

Related Work - Generative Data-free Compression

Prior study -
Generative low-bitwidth data free quantization[1]

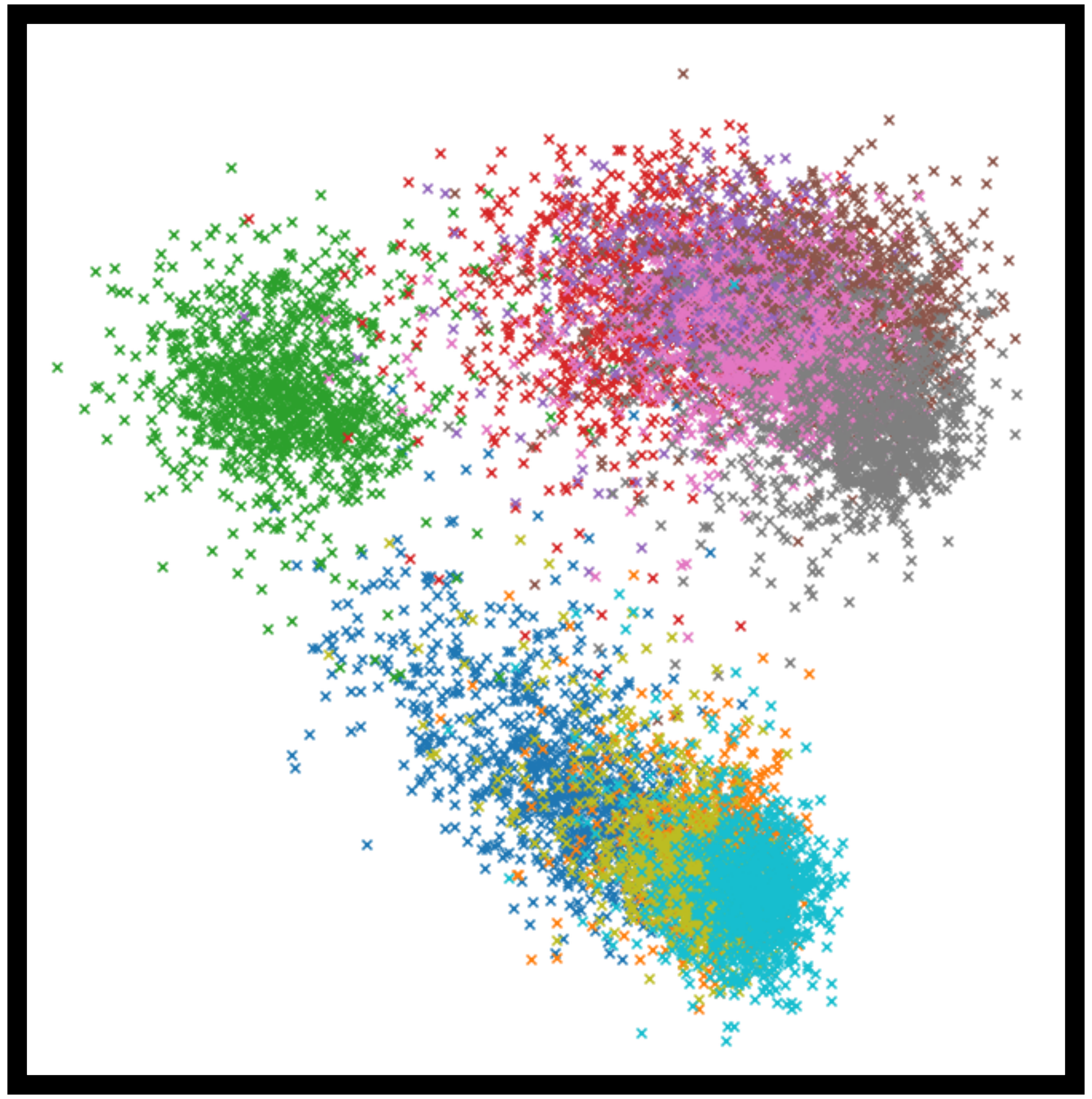


[1] Shoukai Xu et al. "Generative low-bitwidth data free quantization". In: *European Conference on Computer Vision*. 2020.

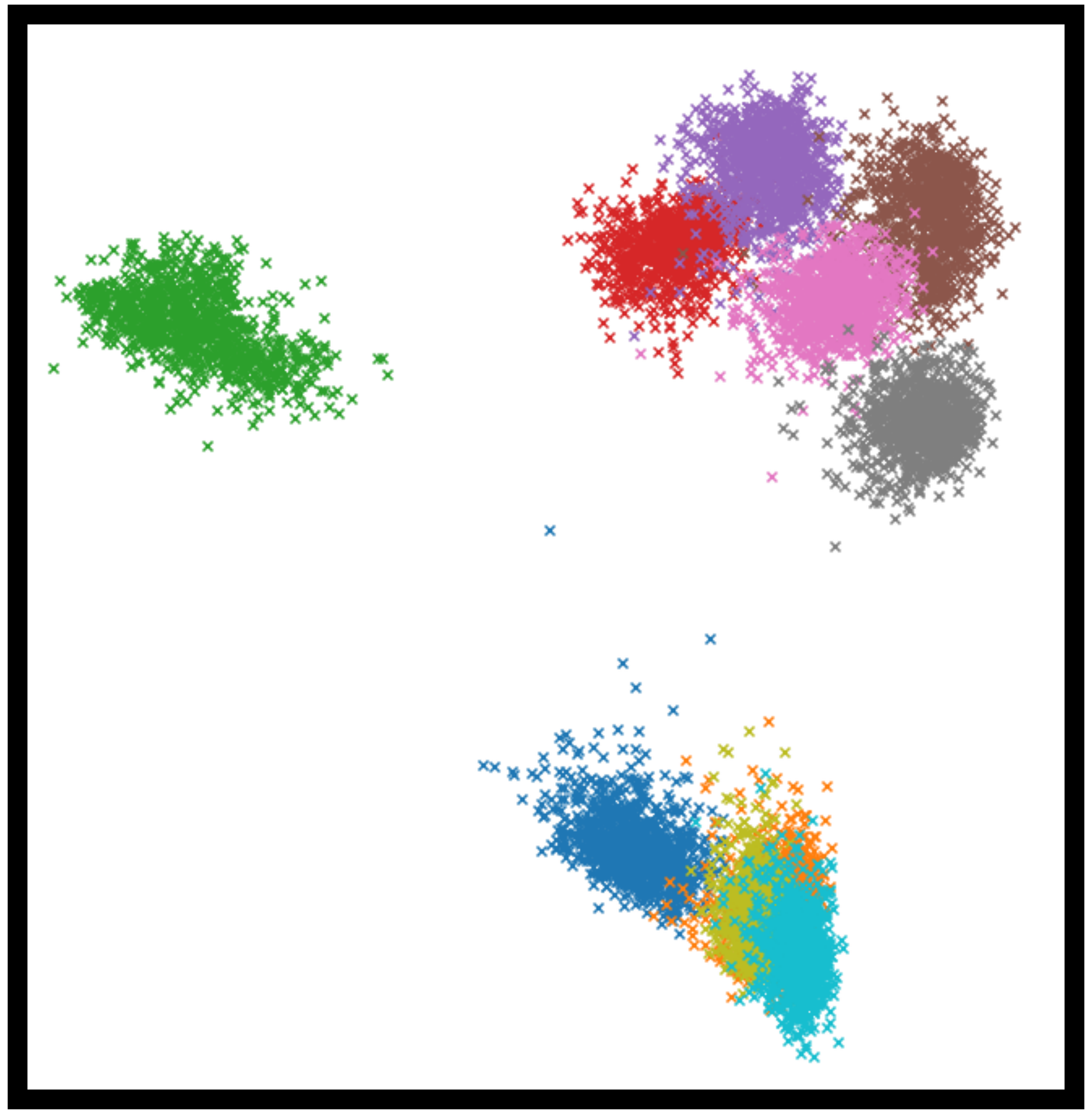
Data-free Neural Network Compression

Motivational Experiment

- Feature space visualization by dimension reduction using PCA



CIFAR-10 Original Distribution

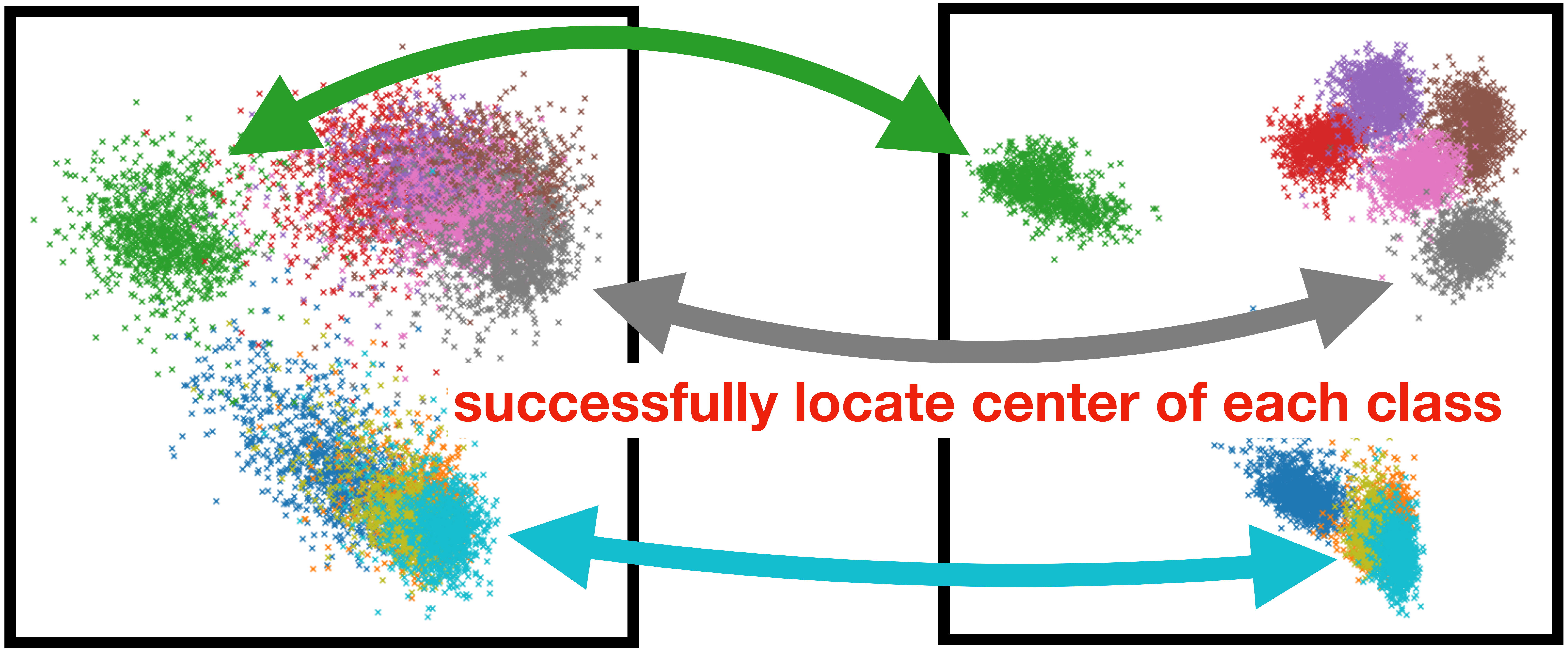


Synthetic Sample Distribution

Data-free Neural Network Compression

Motivational Experiment

- Feature space visualization by dimension reduction using PCA



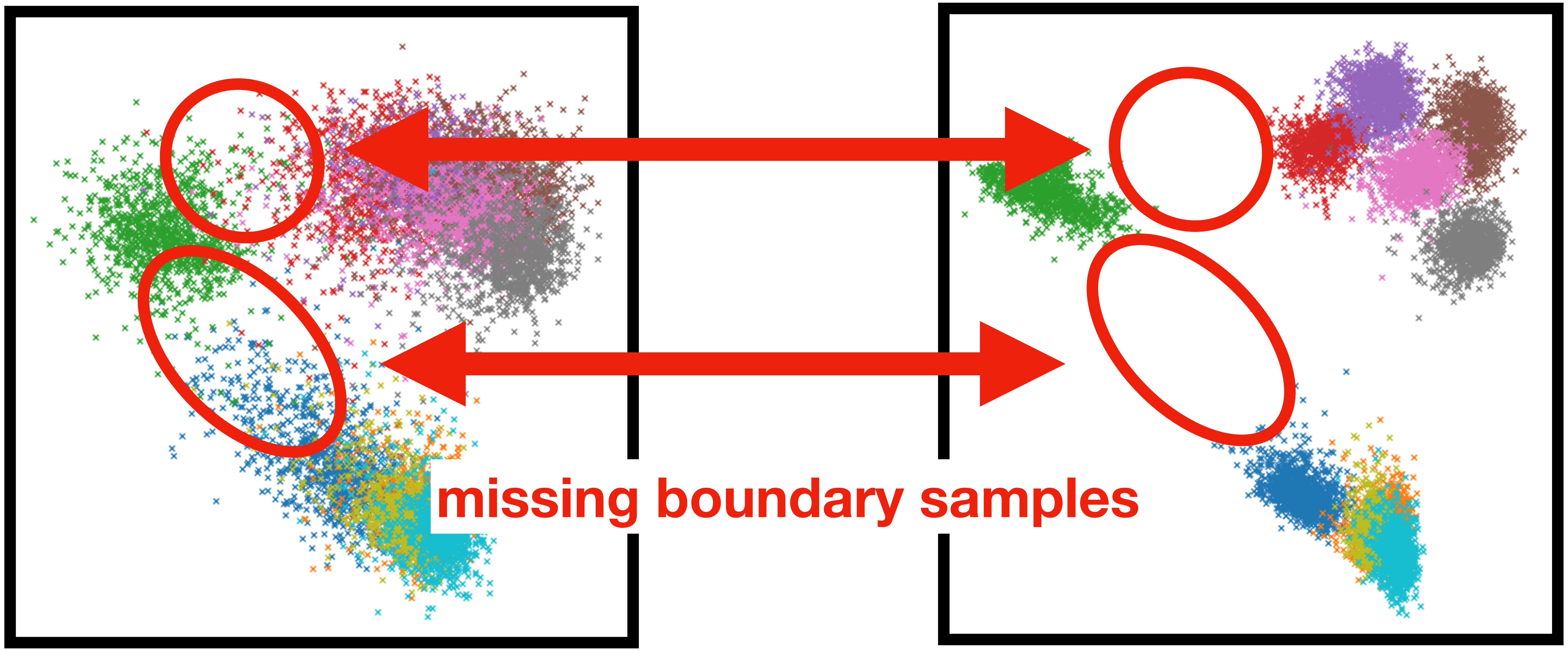
CIFAR-10 Original Distribution

Synthetic Sample Distribution

Data-free Neural Network Compression

Motivational Experiment

- Feature space visualization by dimension reduction using PCA



CIFAR-10 Original Distribution

Synthetic Sample Distribution

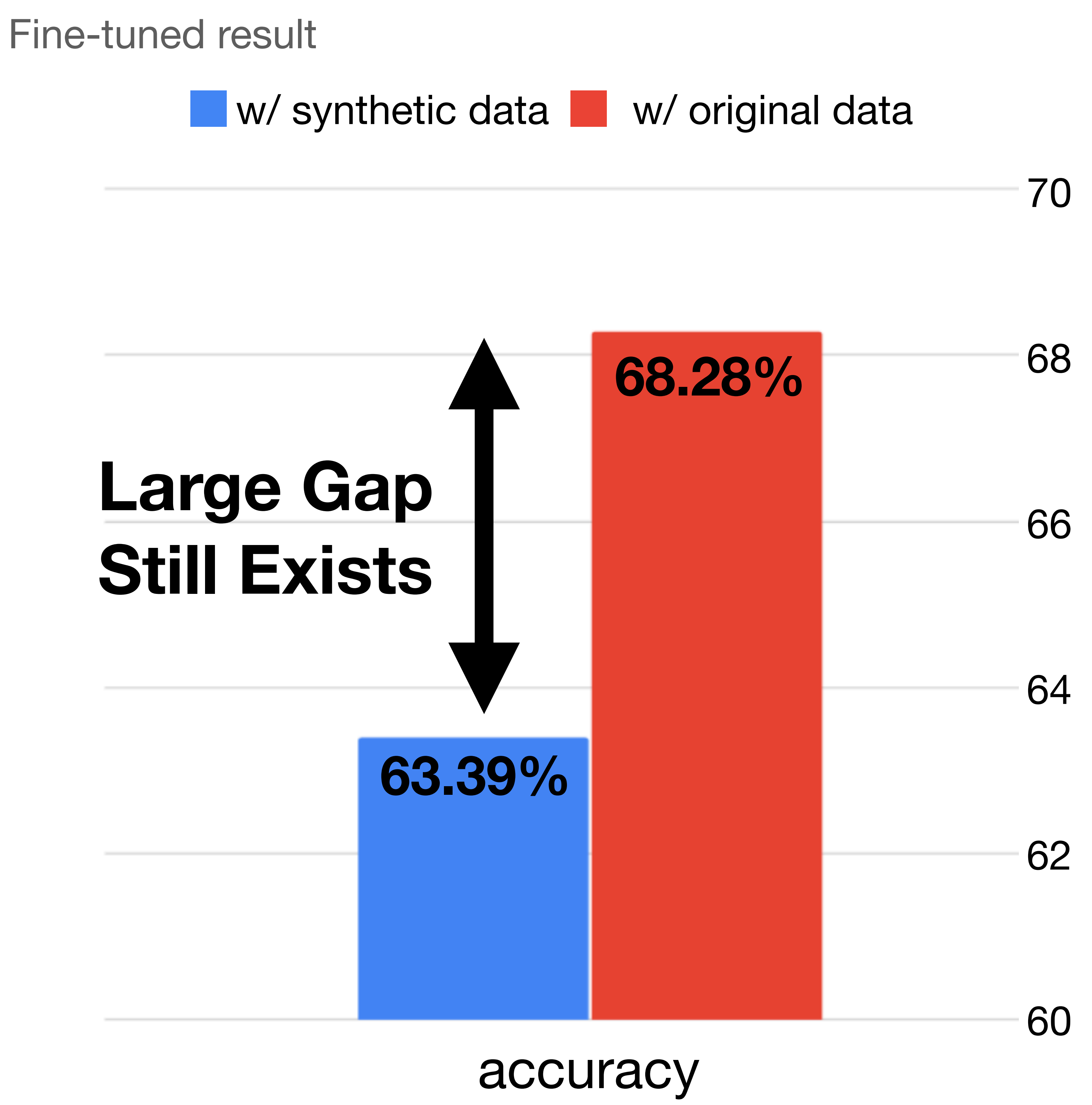
Data-free Neural Network Compression

Motivational Experiment - Quantitative Analysis

- Experiment on ResNet-20, CIFAR-100, 4w4a quantization
- Generative method still has considerable gap between original data

Hypothesis

The lack of boundary supporting samples cause accuracy degradation



Data-free Neural Network Compression

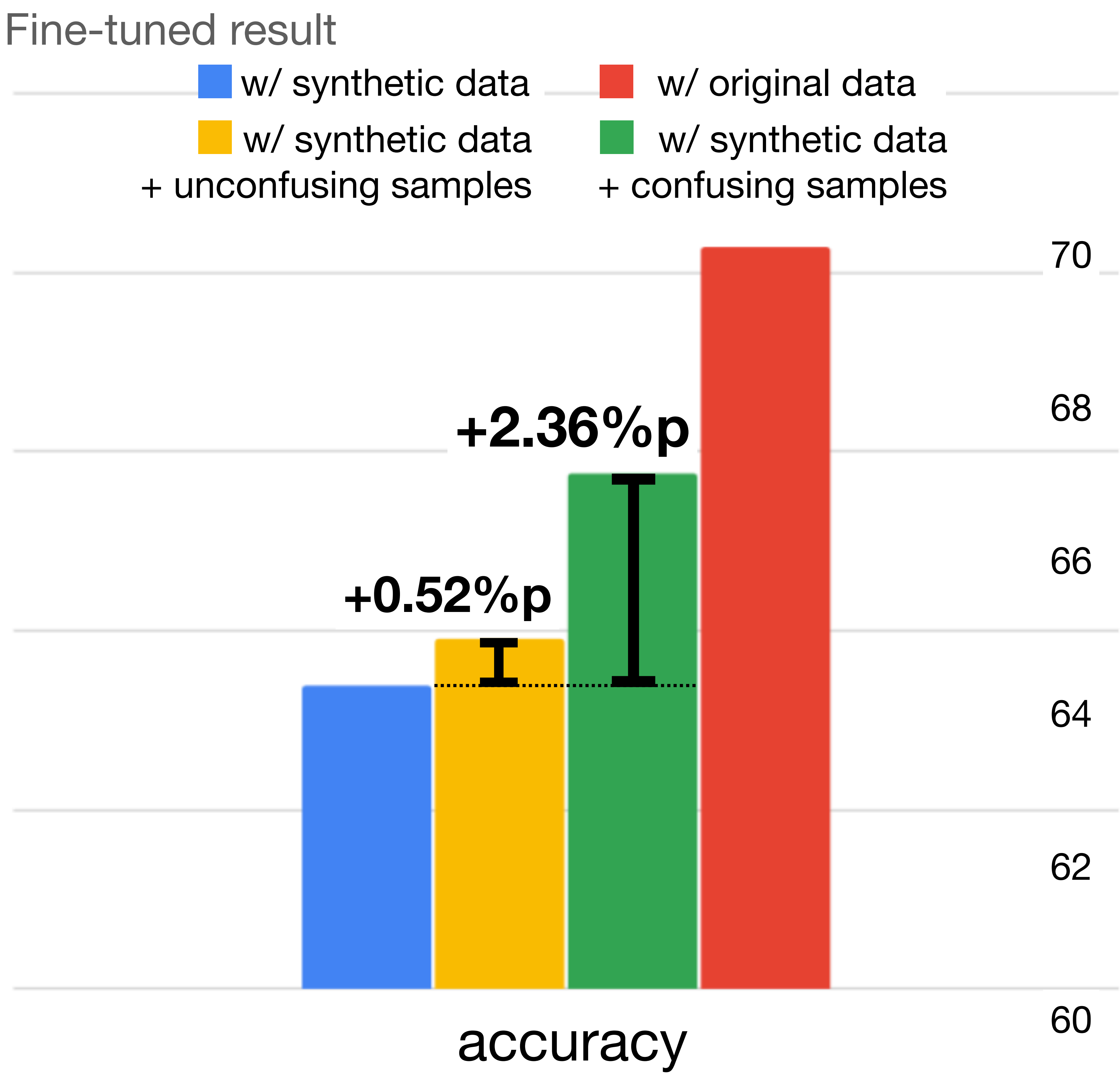
Motivational Experiment - Quantitative Analysis

- Add 15 real samples per class to synthetic data

    1. Unconfusing real samples that have high confidence from teacher

    2. Confusing real samples that have low confidence (**boundary samples**)

Experiment results show boundary supporting samples can help to reduce quantization error



Qimera

Data-free Quantization with Synthetic Boundary Supporting Samples

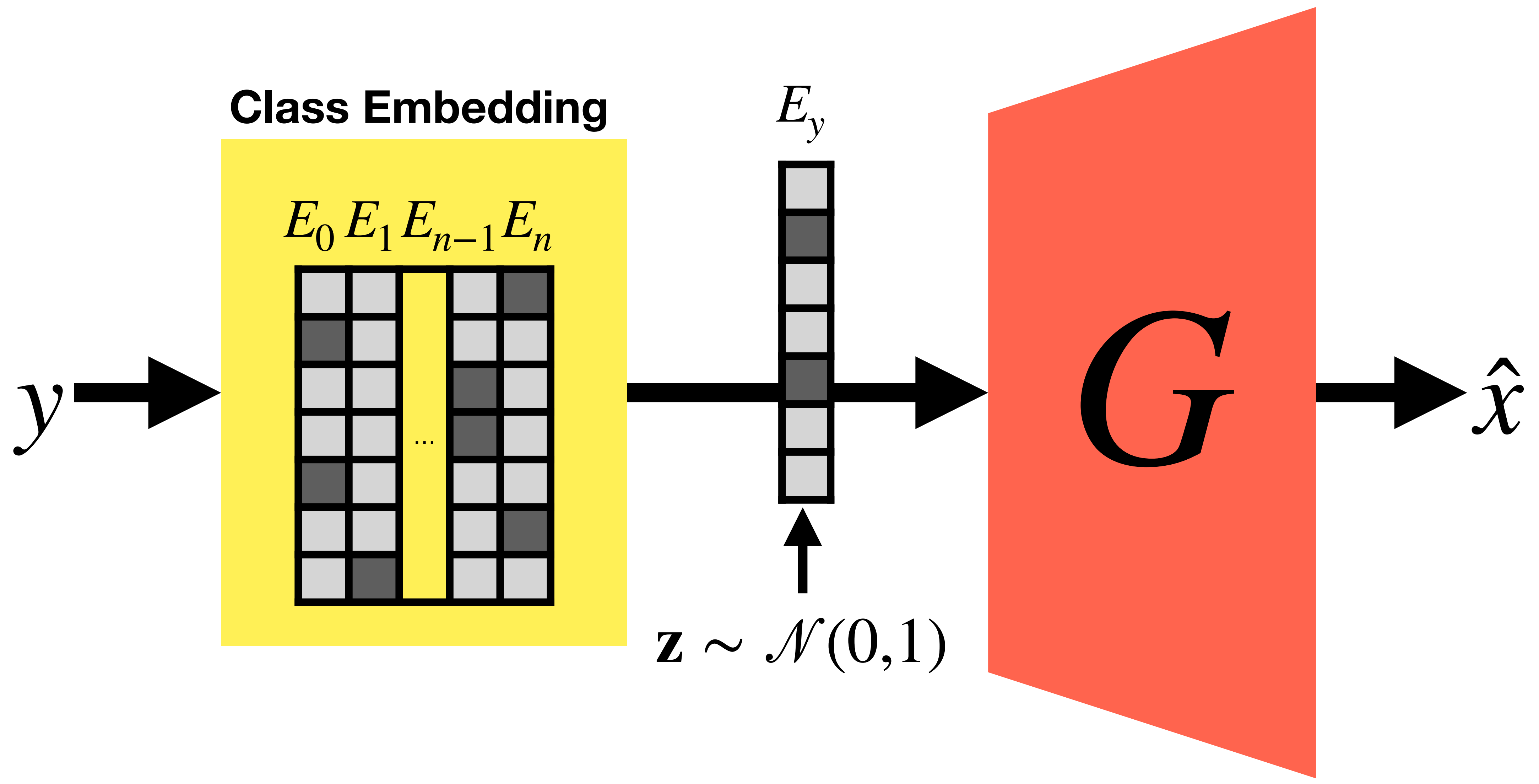
**Generative data-free quantization method,
focuses on synthesizing boundary supporting samples**

Three main methods,

1. Superposed Embedding (SE)
2. Disentanglement Mapping (DM)
3. Extracted Embedding Information (EEI)

Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 1 : Superposed Embedding (SE)

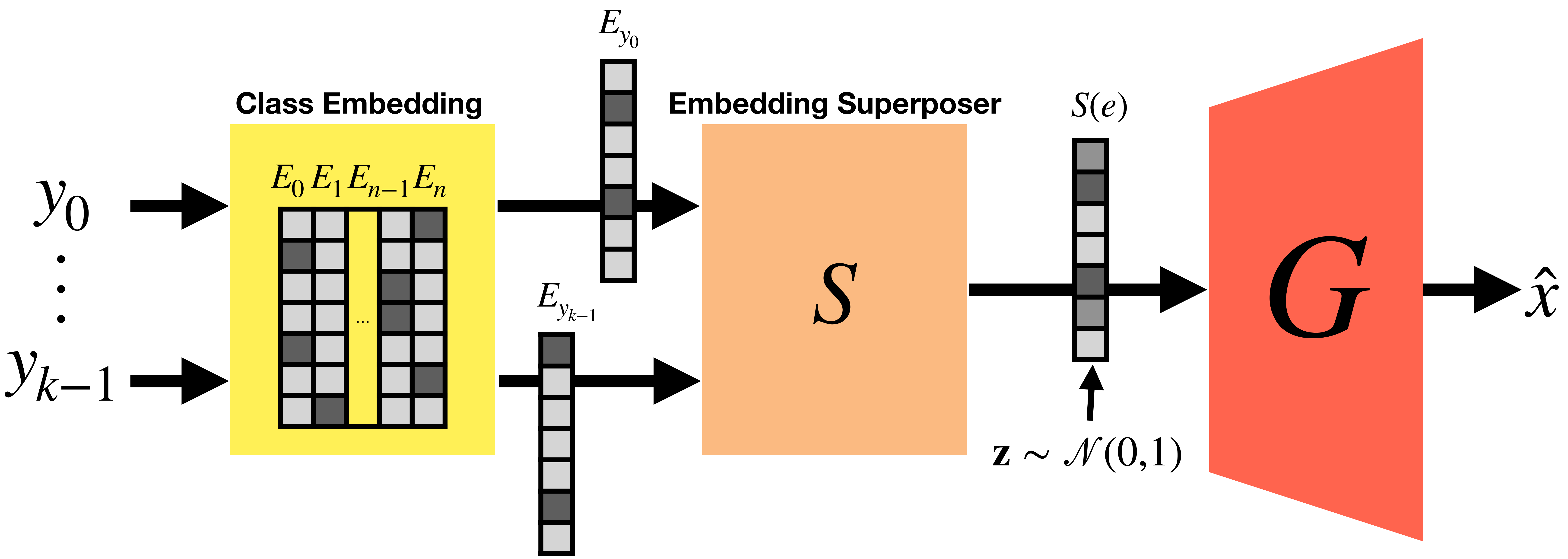


Synthetic image generation : $\hat{x} = G(\mathbf{z} + E_y)$ **,** $\mathbf{z} \sim \mathcal{N}(0,1)$,

where generator G, class embedding vector E_y, and random noise $\mathbf{z}$.

Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 1 : Superposed Embedding (SE)



Superposed Embedding (SE) : $S(e) = \mathbf{z} + \sum_{k}^{K} \lambda_k e_k$

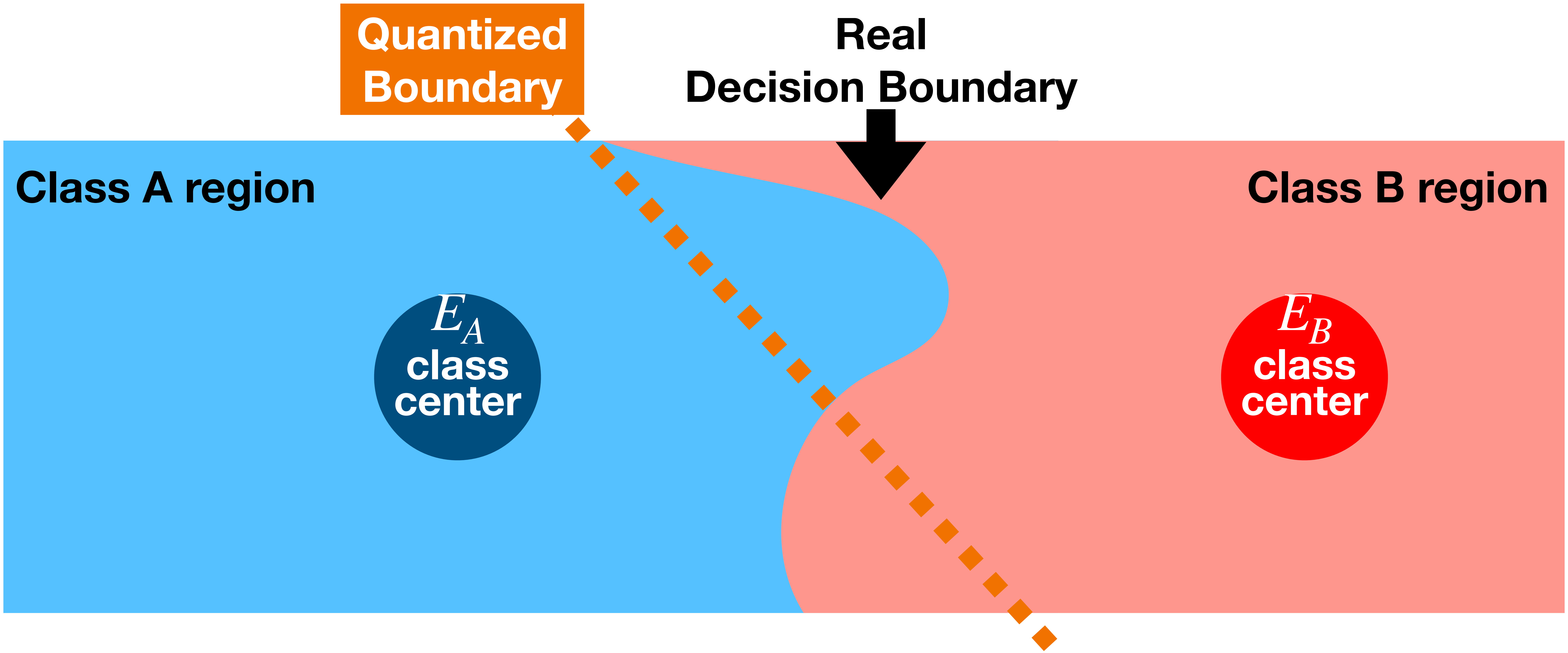
Boundary supporting samples from SE : $(\hat{x}', \hat{y}') = \left(G\big(S(e)\big), \sum_{k}^{K} \lambda_k y_k\right)$

$\lambda_i = Softmax(p_i)$

$p_i \sim \mathcal{N}(0,1)$

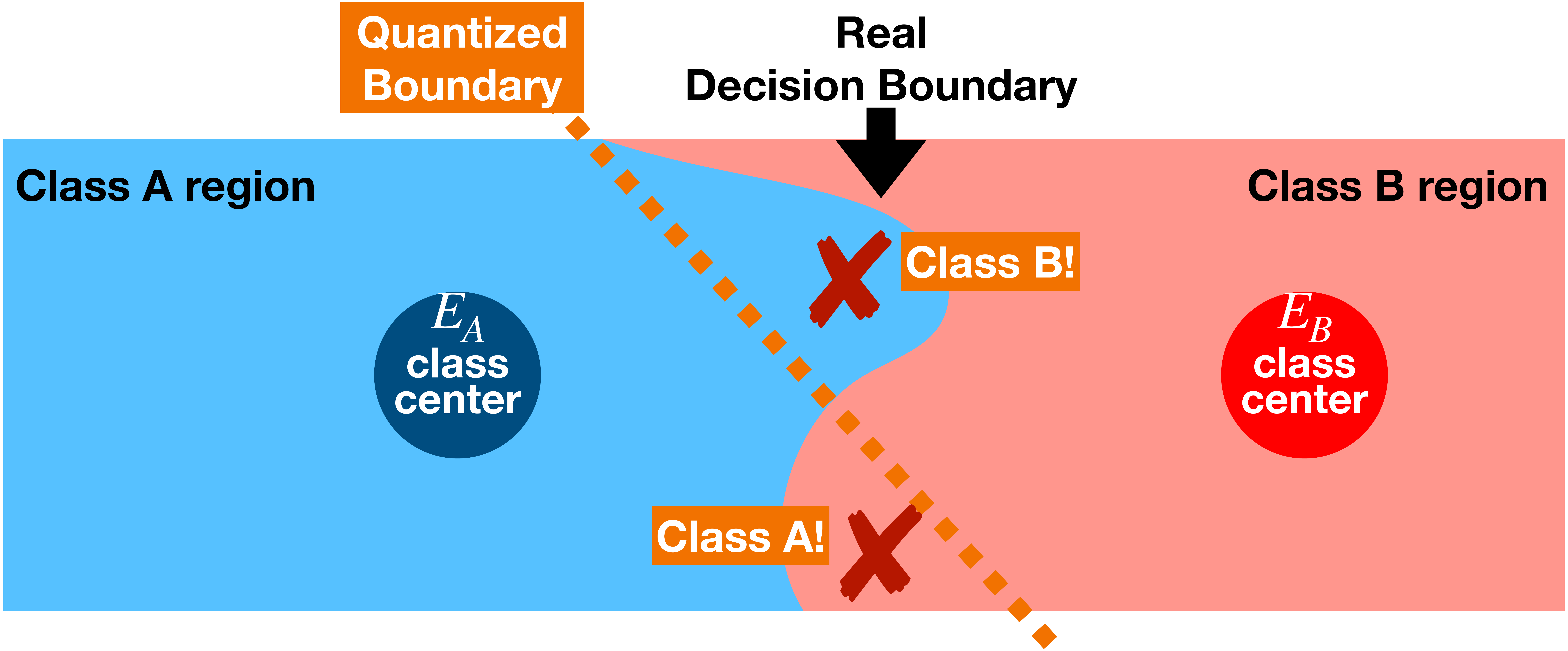
Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 1 : Superposed Embedding (SE)



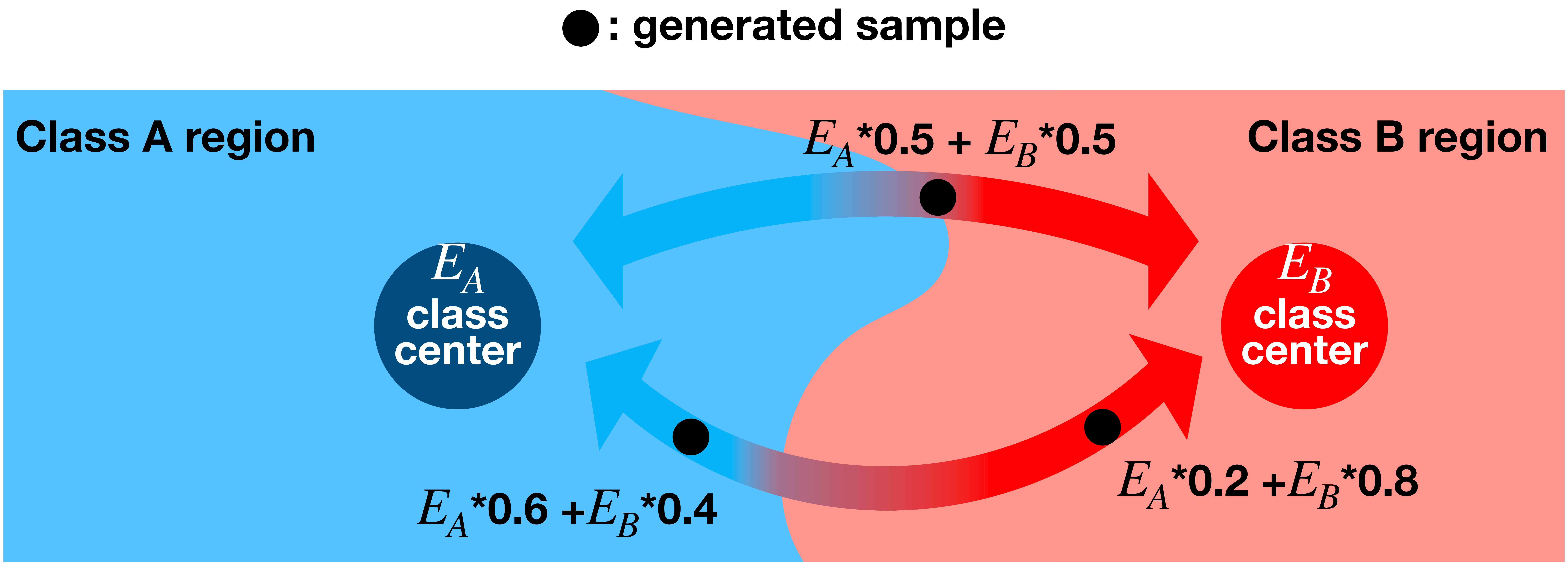
Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 1 : Superposed Embedding (SE)



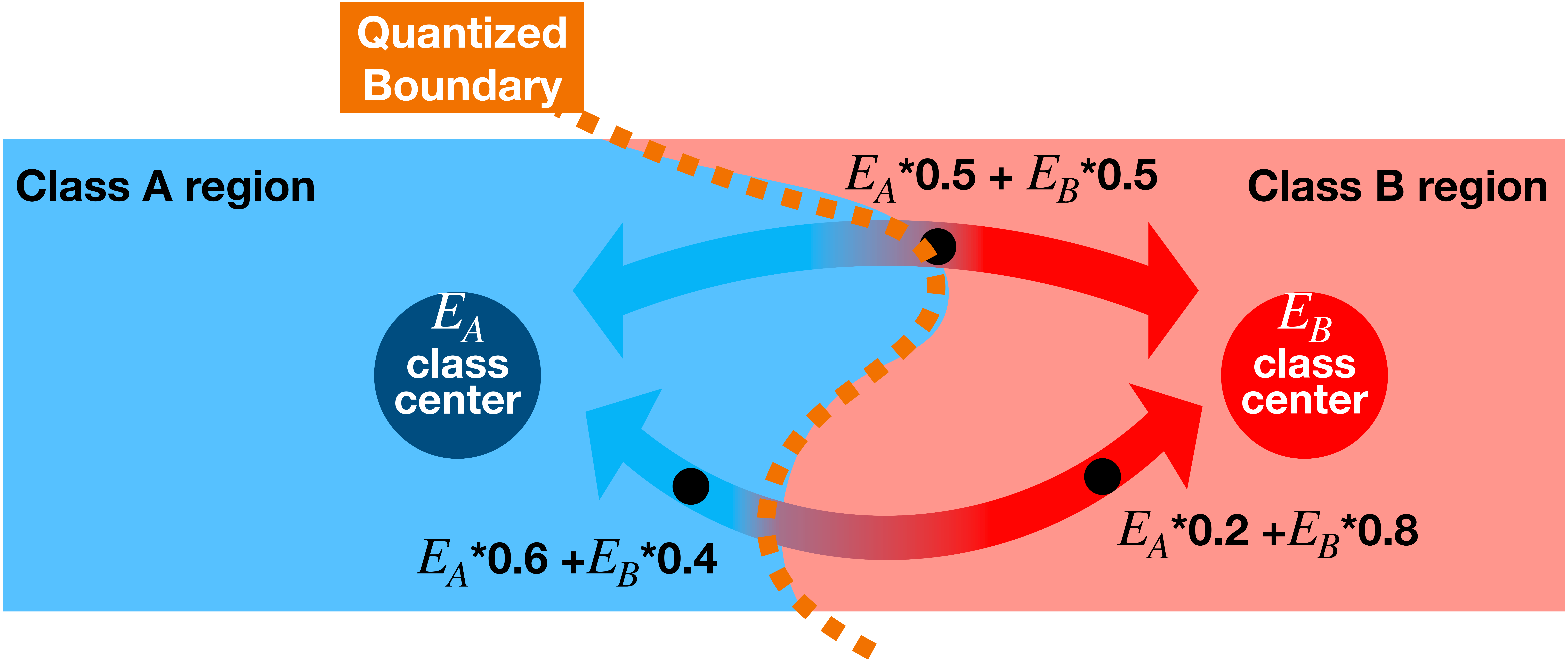
Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 1 : Superposed Embedding (SE)



Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 1 : Superposed Embedding (SE)



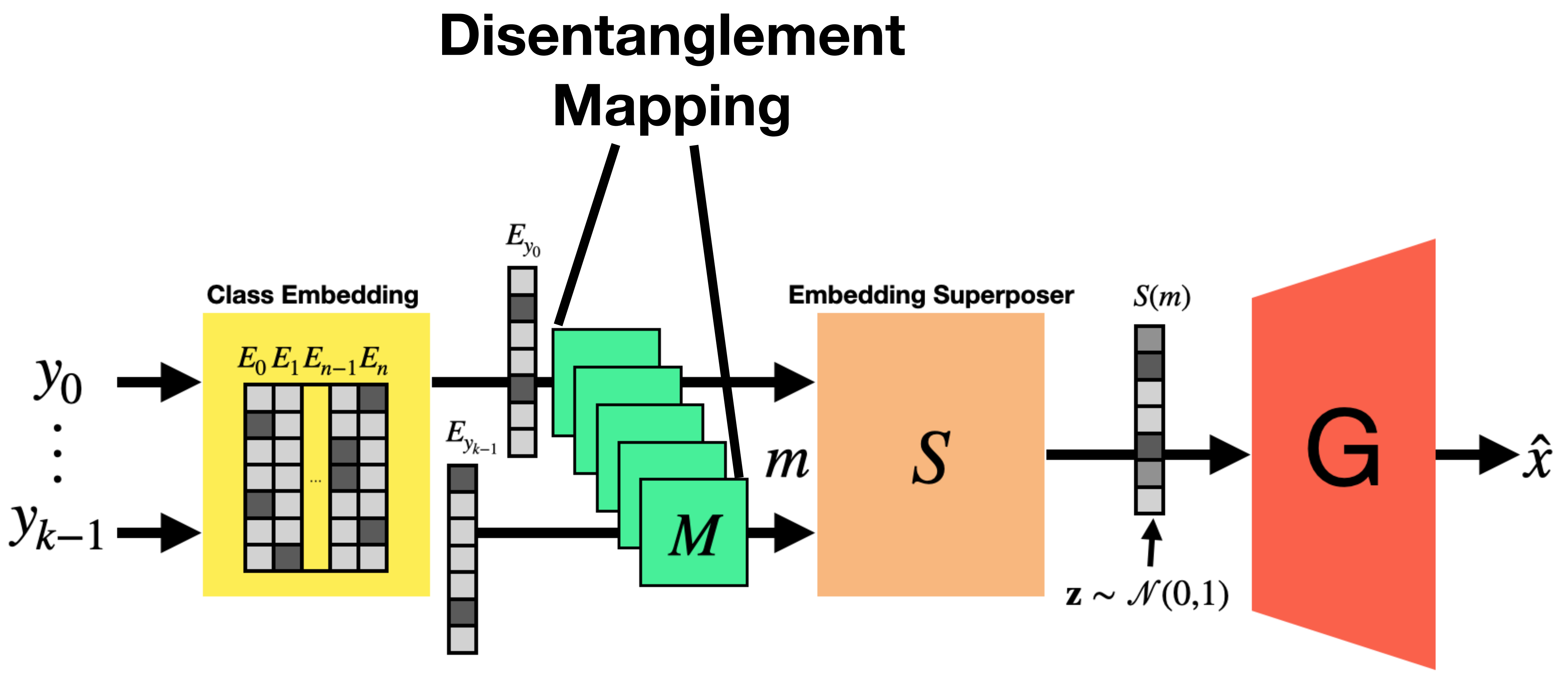
Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 2 : Disentanglement Mapping (DM)

Learnable mapping function $M : \mathbb{R}^D \rightarrow \mathbb{R}^d$

$$S(e) = \mathbf{z} + \sum_{k}^{K} \lambda_k M(e_k)$$

Implemented as single-layer perceptron



Inspired by StyleGAN[2] paper,

DM disentangles class embedding E_i from E_k, where $k \neq i$.

[2] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 2 : Disentanglement Mapping (DM)

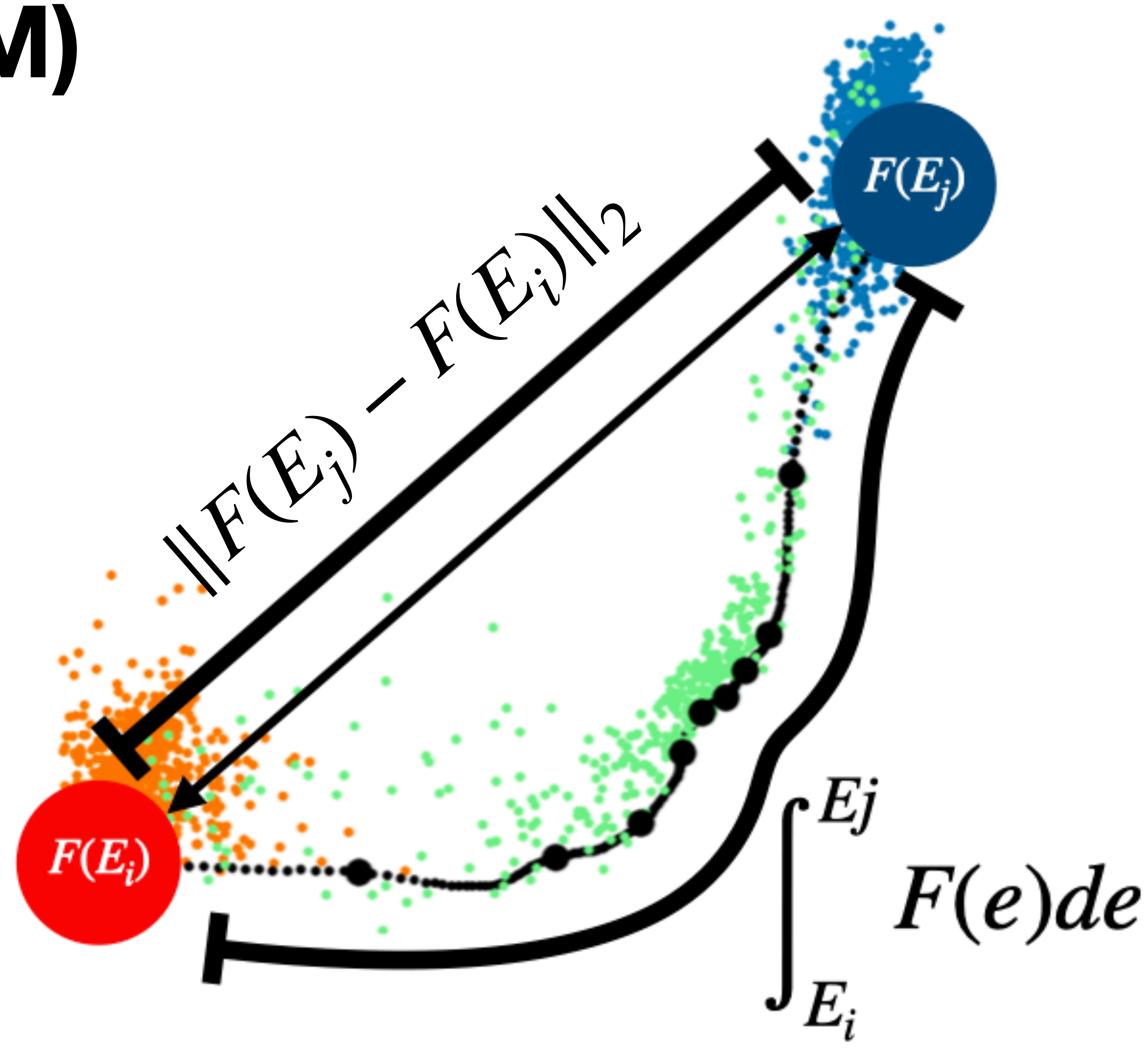
Analyzing the effect of DM

We measured the ratio of perceptual distance divided by Euclidean distance,

$$\text{s.t. } \frac{\int_{E_i}^{Ej} F(e)de}{\|F(E_j) - F(E_i)\|_2},$$

$$\text{where } \int_{E_i}^{Ej} F(e)de \sim \sum_{k}^{K} F(\frac{k}{K}E_i + (1-\frac{k}{K})E_j),$$

$F(e)$ is feature representation extracted from the teacher network.
K is set to 1000.



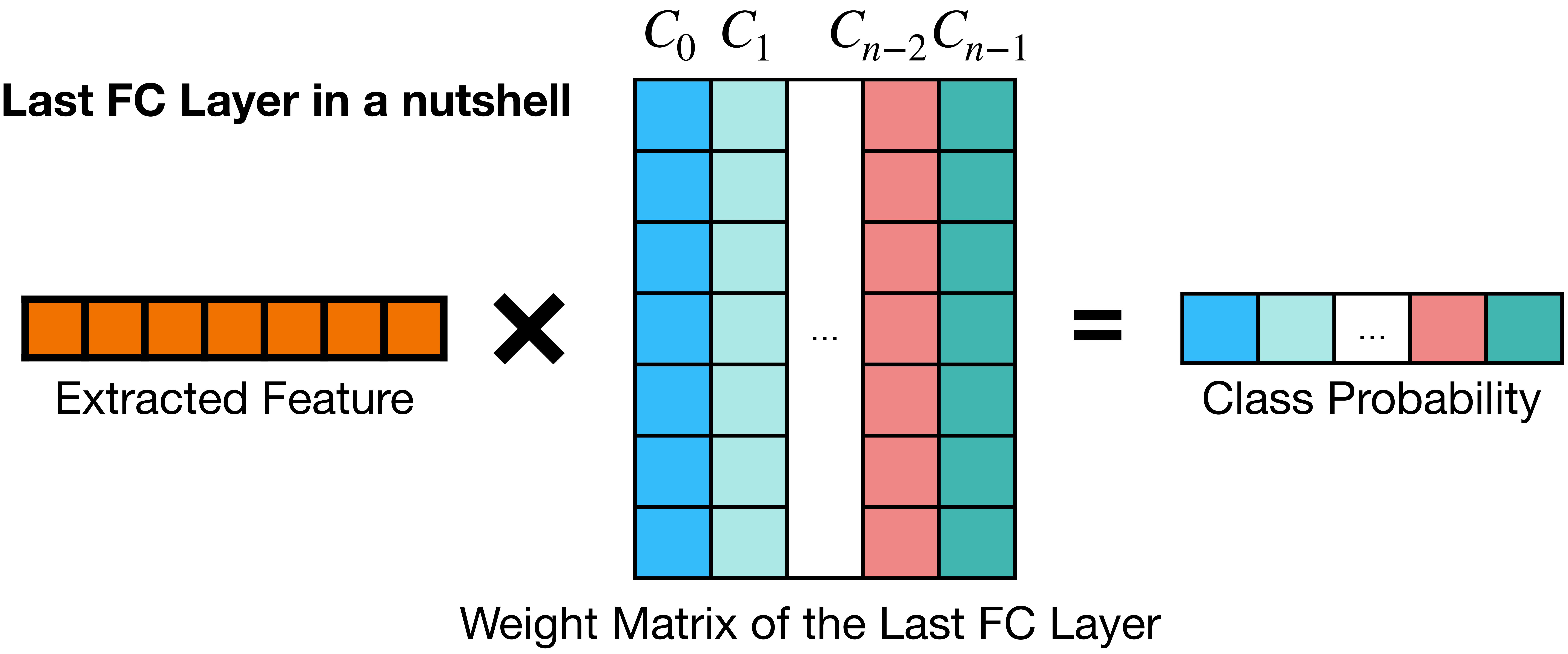
	Dist. Ratio
SE Only	1.64
SE + DM	1.59 (-0.05)

ResNet20, CIFAR100

Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

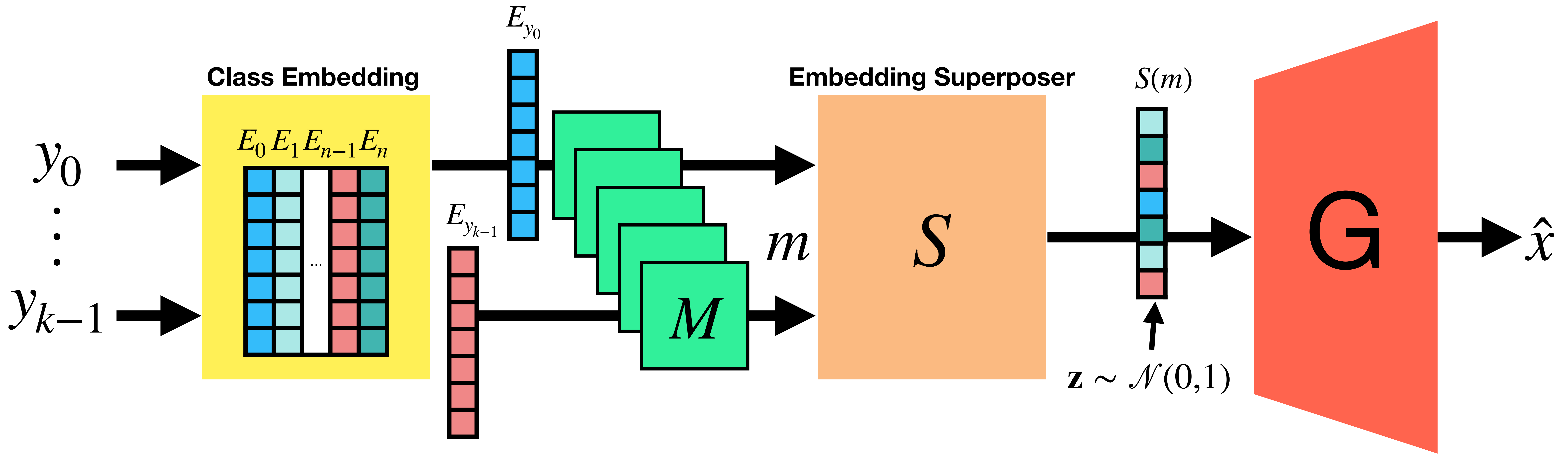
Method 3 : Extracted Embedding Information (EEI)

Last FC Layer in a nutshell



Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Method 3 : Extracted Embedding Information (EEI)



Corresponding column of the weight matrix represents class information

e.g. Distance between classes,Class similarity, etc.

Use corresponding column vectors as initialization of class embedding vectors

	Dist. Ratio
SE Only	1.64
SE + DM	1.59 (-0.05)
SE + DM + EEI	1.52 (-0.07)

ResNet20, CIFAR100

Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Experiment Results : Accuracy Improvement

Experiment results of Qimera show that,

- achieves superior performance in **most settings**
- significant improvement on **low-bit settings**
- robust increase on **large-scale dataset**
- higher accuracy gain on **deeper network** e.g. over 13%p improvement on ResNet-50

Dataset	Model	Bit	Qimera (%p improvement)
CIFAR-10	ResNet-20 (93.89)	4w4a	91.26 +- 0.49 (-0.87)
		5w5a	93.46 +- 0.03 (+0.08)
CIFAR-100	ResNet-20 (70.33)	4w4a	65.10 +- 0.33 (+1.71)
		5w5a	69.02 +- 0.22 (+0.32)
ImageNet	ResNet-18 (71.47)	4w4a	63.84 +- 0.30 (+3.24)
		5w5a	69.29 +- 0.16 (+0.89)
	ResNet-50 (77.73)	4w4a	66.25 +- 0.90 (+13.23)
		5w5a	75.32 +- 0.09 (+1.94)
	MobileNetV2 (73.03)	4w4a	61.62 +- 0.39 (+2.19)
		5w5a	70.45 +- 0.07 (+2.34)

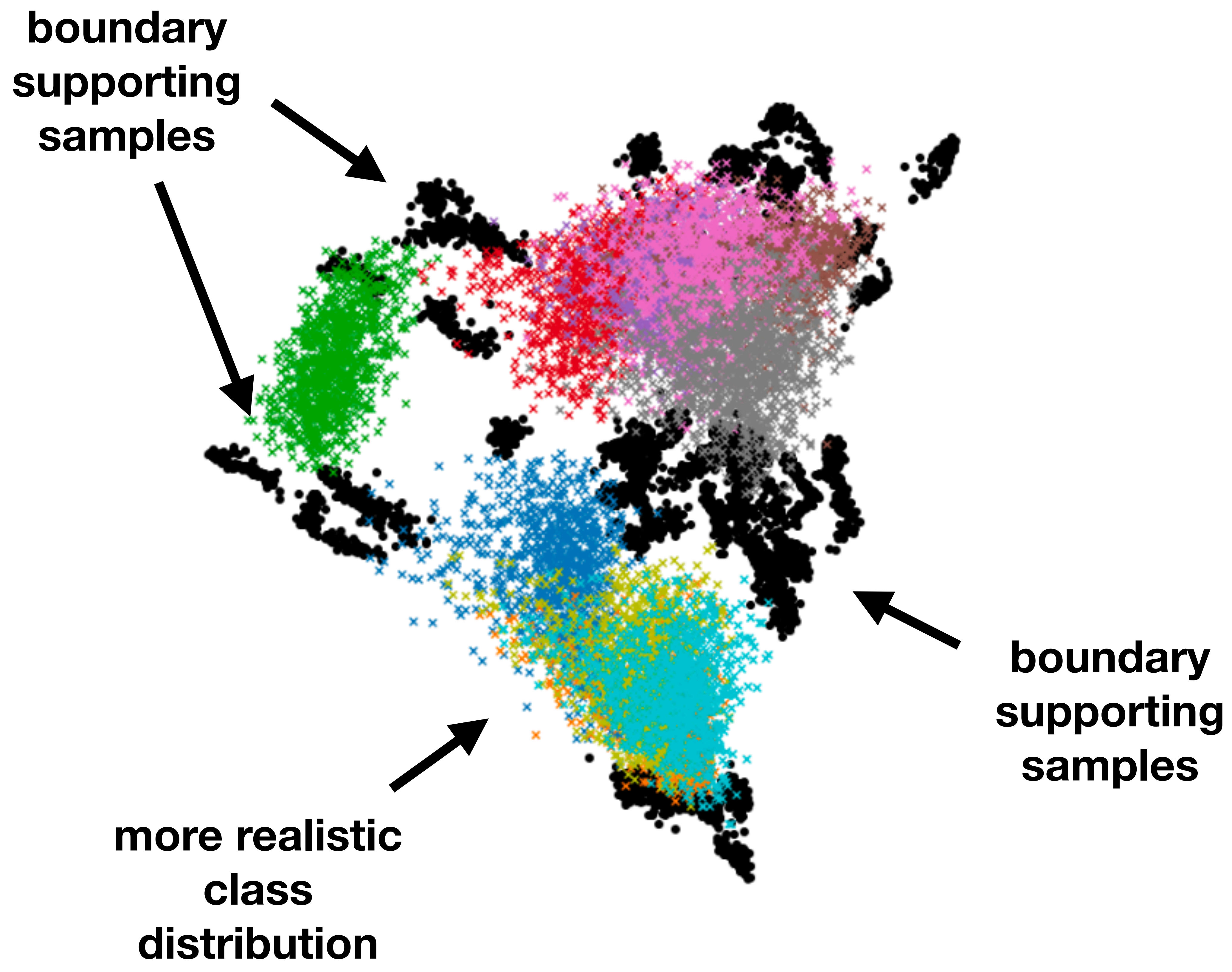
Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Experiment Results : Visualization

PCA visualization from motivational experiment

Feature space visualization of the Qimera shows that

- successfully generated boundary supporting samples
- more realistic class distribution



Qimera (Ours)

Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Ablation Study

Conducted ablation study upon SE, DM, EEI

- SE Only shows significant accuracy gain
- Both DM and EEI further improve accuracy by disentangling embedding space
- Using DM and EEI simultaneously with SE, overall 14%p improvement has gain

Dataset	Method	Accuracy
ImageNet (ResNet-50)	Baseline	52.12
	SE Only	64.09 (+11.98)
	SE + DM	66.06 (+13.94)
	SE + EEI	64.44 (+12.32)
	SE + DM + EEI	**66.25 (+14.13)**

Qimera : Data-free Quantization with Synthetic Boundary Supporting Samples

Conclusion

- The data-free quantization is a promising way to compress neural networks even without the original train dataset.
- We conducted an experiment that shows existing methods lack boundary supporting samples, which cause accuracy degradation.
- We propose a simple yet effective method to generate boundary supporting samples for data-free quantization.
- The extensive experiments show our method achieved SOTA performance in many settings.