JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

THU-AM-374

DaFKD : Domain-aware Federated Knowledge Distillation

Haozhao Wang[1], Yichen Li[1,3], Wenchao Xu[4], Ruixuan Li[1]*, Yufeng Zhan[5] and Zhigang Zeng[2]

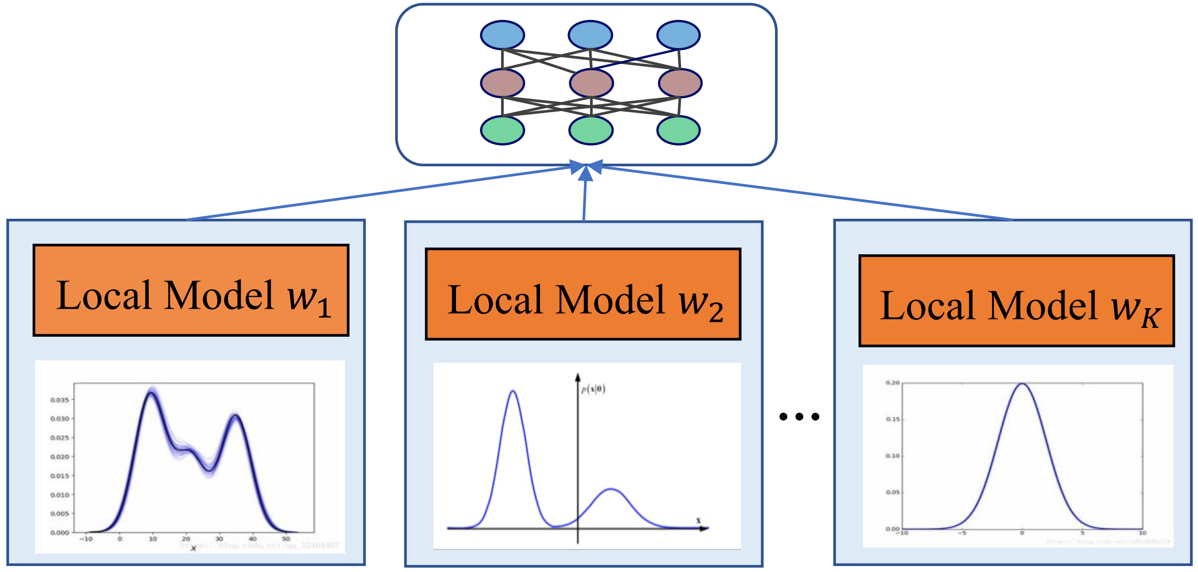
{[1]School of Computer Science and Technology, [2]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology}, Wuhan, China

[3]Soochow University, Suzhou, China, and [4]The Hong Kong Polytechnic University, Hong Kong

[5]Beijing Institute of Technology, Beijing, China

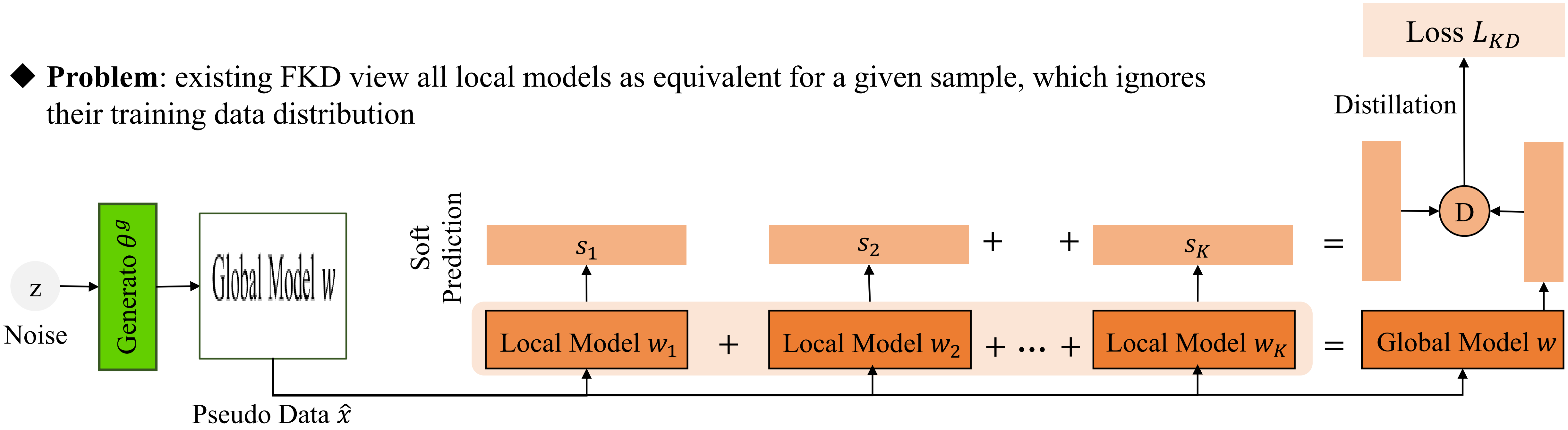
Problem: NonIID Issue in Federated Knowledge Distillation

- **Federated Learning** collaboratively trains a model from NonIID data across multiple clients



- **Federated Knowledge Distillation (FKD)**: global model is obtained by the ensemble distillation of multiple local models for a given sample

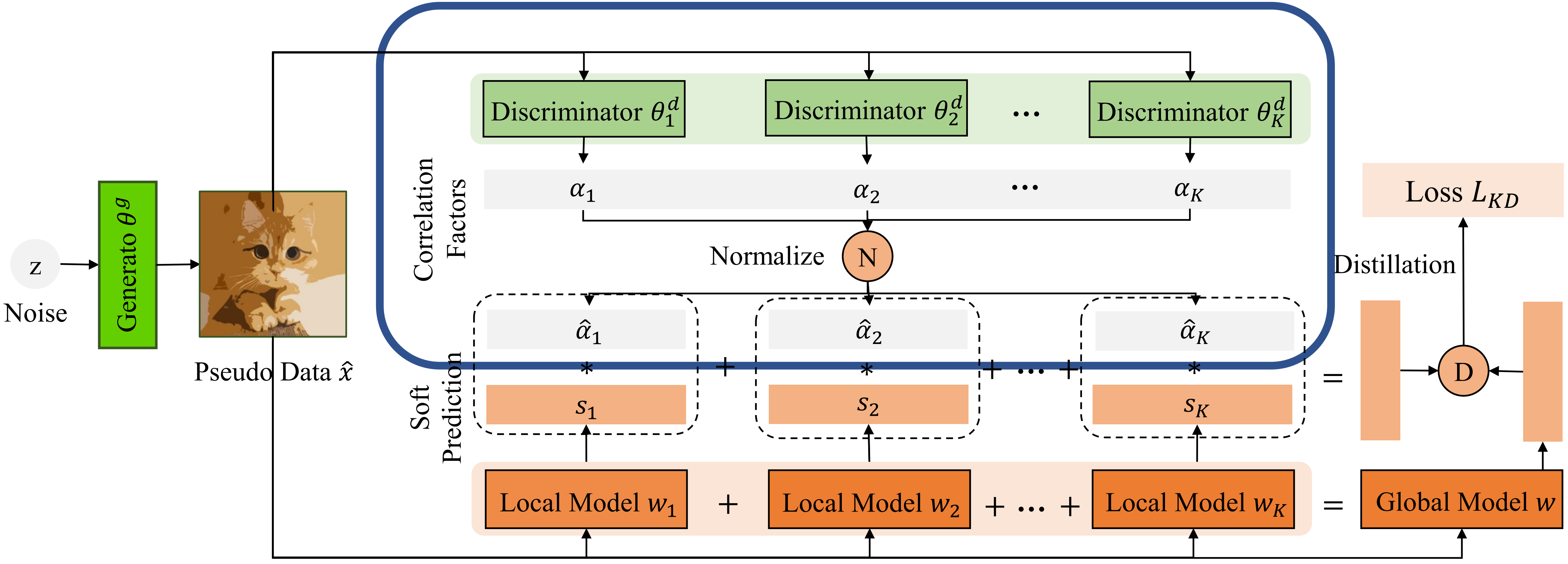
- **Problem**: existing FKD view all local models as equivalent for a given sample, which ignores their training data distribution



Method: Domain-aware Federated Knowledge Distillation

◆ **Domain-aware FKD**: takes the NonIID data into account when making distillation for the global model

- Adaptive factor: for any distillation sample, endow each local model with a specific importance factor
- Discriminator: based on the generator, train a discriminator for each client to produce the importance factor



Results: Theoretical and Empirical Improvement

◆ **Theory**: upper bound is not related to the degree of NonIID

Our method

$$\mathcal{L}_p(\sum_{k=1}^{K}\hat{\alpha}_k(x)h_{\hat{p}_k})$$

$$\leq (K+1)\mathcal{L}_{\hat{p}}(h_{\hat{p}}) + (K+1)\sqrt{\frac{\sigma^2 log\frac{2K}{\delta}}{2m}}.$$

No items related to NonIID

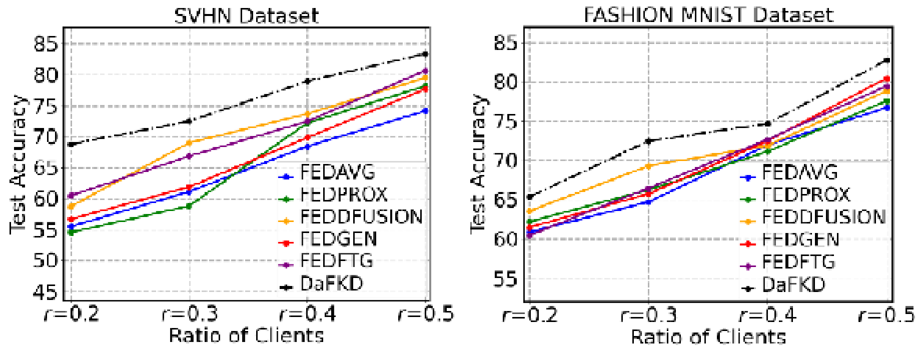
Baseline

$$\mathcal{L}_p(\sum_{k=1}^{K}\hat{\alpha}_k(x)h_{\hat{p}_k}) \leq \mathcal{L}_{\hat{p}}(h_{\hat{p}}) \tag{12}$$

$$+ \sqrt{\frac{4}{m'}(d\log\frac{2em}{d} + \log\frac{4K}{\delta})} + \boxed{\frac{1}{K}\sum_{i=1}^{K} d_{\mathcal{H}\Delta\mathcal{H}}(p'_k, p) + \lambda'_k}.$$

◆ **Experiment**: highest accuracy among all methods

Dataset	Setting	FEDAVG	FEDGEN	FEDFTG	DaFKD
MNIST, E = 20	$\alpha = 0.05$	69.11± 1.39	81.06± 1.09	80.95±1.06	**82.33±0.44**
	$\alpha = 0.1$	95.16± 0.79	94.98± 0.47	94.43±0.49	**95.56±0.41**
	$\alpha = 1$	98.11± 0.14	96.39± 0.90	98.47±0.21	**98.96±0.38**
SVHN, E = 20	$\alpha = 0.05$	33.01± 0.12	47.36± 0.42	48.69±1.87	**51.14±0.16**
	$\alpha = 0.1$	53.54± 0.21	60.03± 1.12	63.75±0.11	**72.80±0.11**
	$\alpha = 10$	81.44± 0.01	82.91± 0.73	83.49±1.32	**87.31±0.85**
FASHION MNIST, E = 20	$\alpha = 0.05$	30.01± 0.54	36.59± 0.98	34.84±0.77	37.85±0.24
	$\alpha = 0.1$	67.97± 0.03	67.29± 2.05	67.25±0.14	**70.81±0.21**
	$\alpha = 10$	82.37± 0.82	81.57± 1.96	81.96±1.86	**83.37±0.06**
EMNIST, E = 40	$\alpha = 0.05$	67.28± 0.14	**68.95±0.88**	67.08±0.97	67.64±1.86
	$\alpha = 0.1$	69.13± 0.23	72.15± 2.04	72.91±1.87	**74.96±0.91**
	$\alpha = 10$	81.35± 1.03	82.02± 1.19	82.65±1.04	**84.60±1.86**



(a) SVHN Dataset. (b) FASHION MNIST Dataset.

Federated Learning

I. Background and Challenge

II. Related Works and Limitations

III. Methodology and Theory

IV. Experimental Results

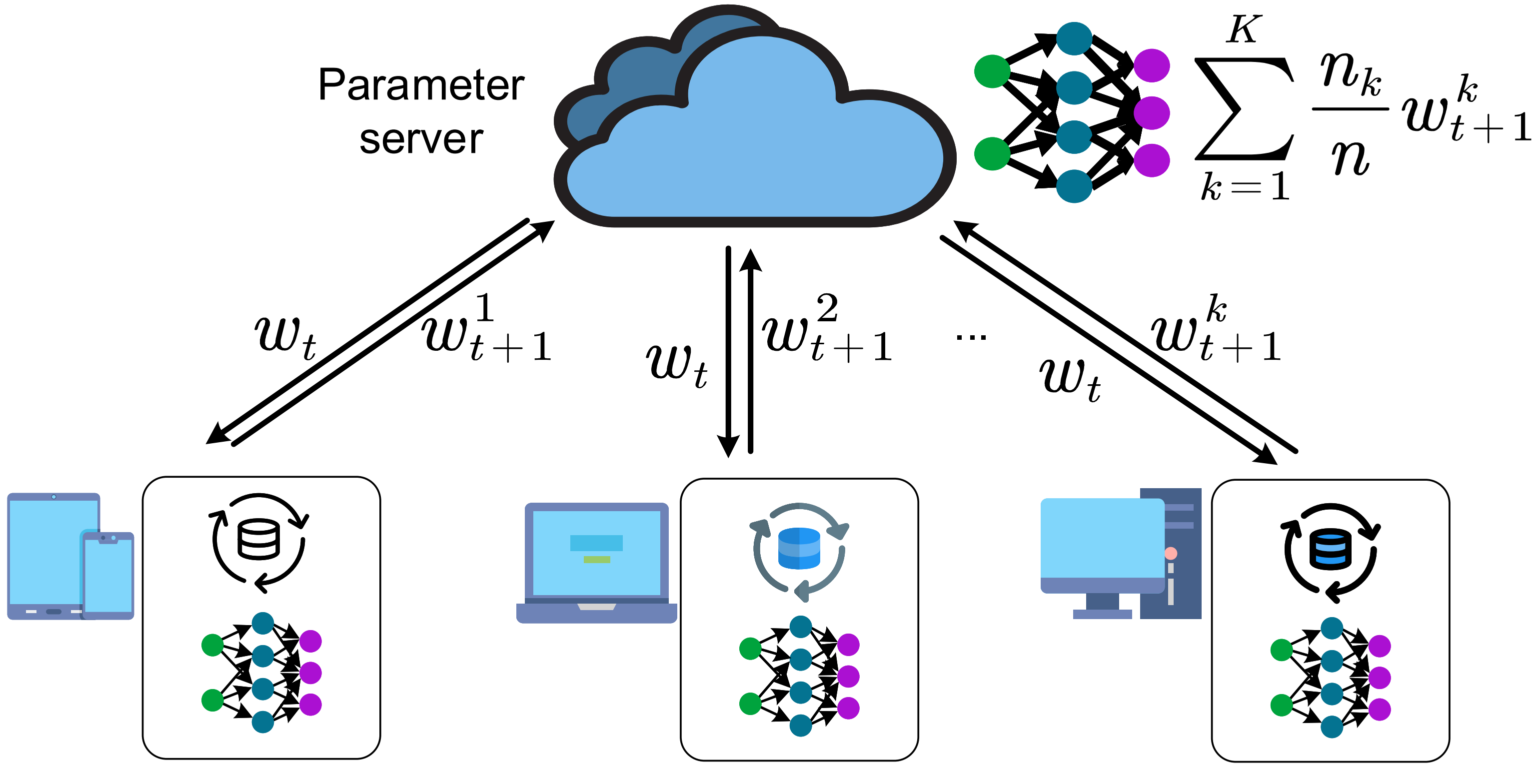
V. Conclusion

Pervasive Application of Federated Learning

Federated Learning has been deployed in a wide range of applications, such as medical analysis and intelligent industry

FedAvg: Federated Learning with Aggregating Parameters

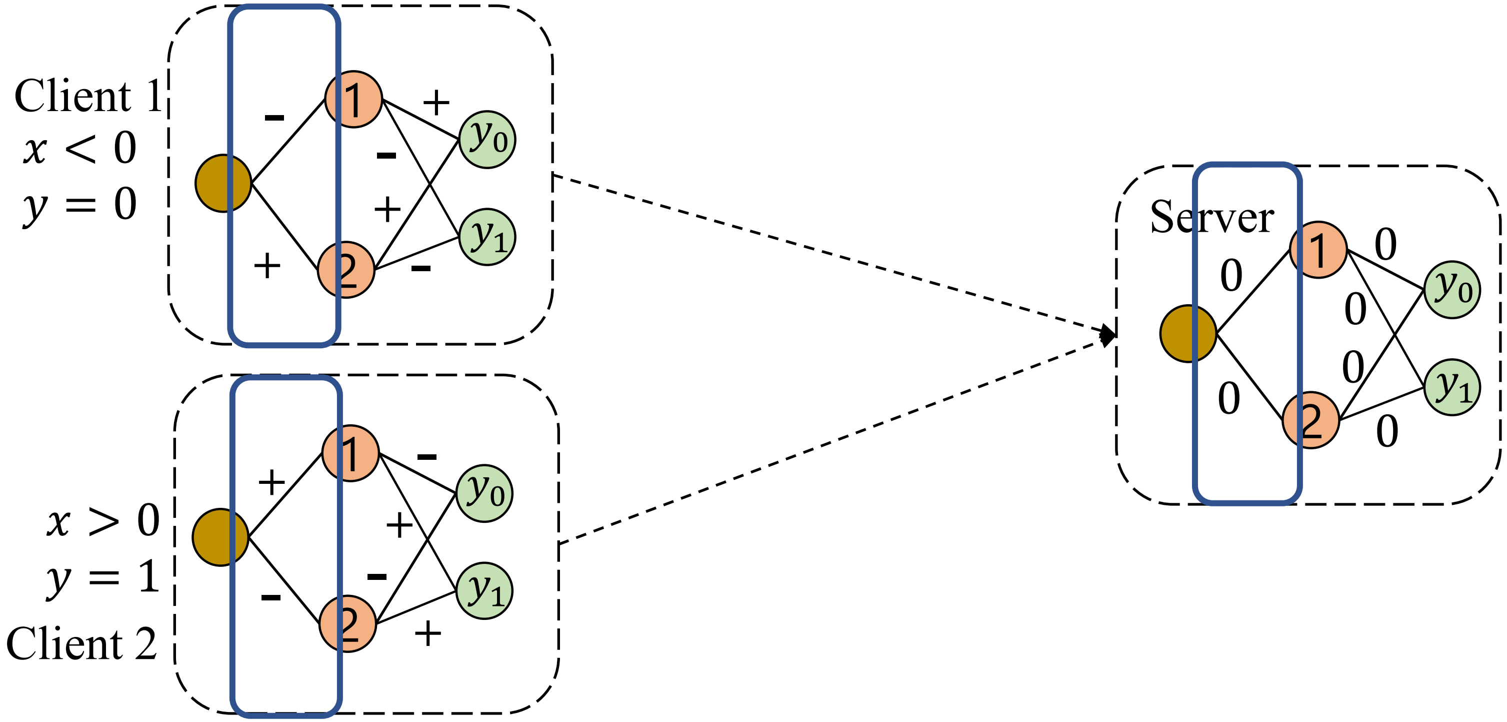
- **Federated Learning**: collaboratively train a model from data across multiple clients



FedAvg: Global model is obtained by computing the average of parameters of multiple local models

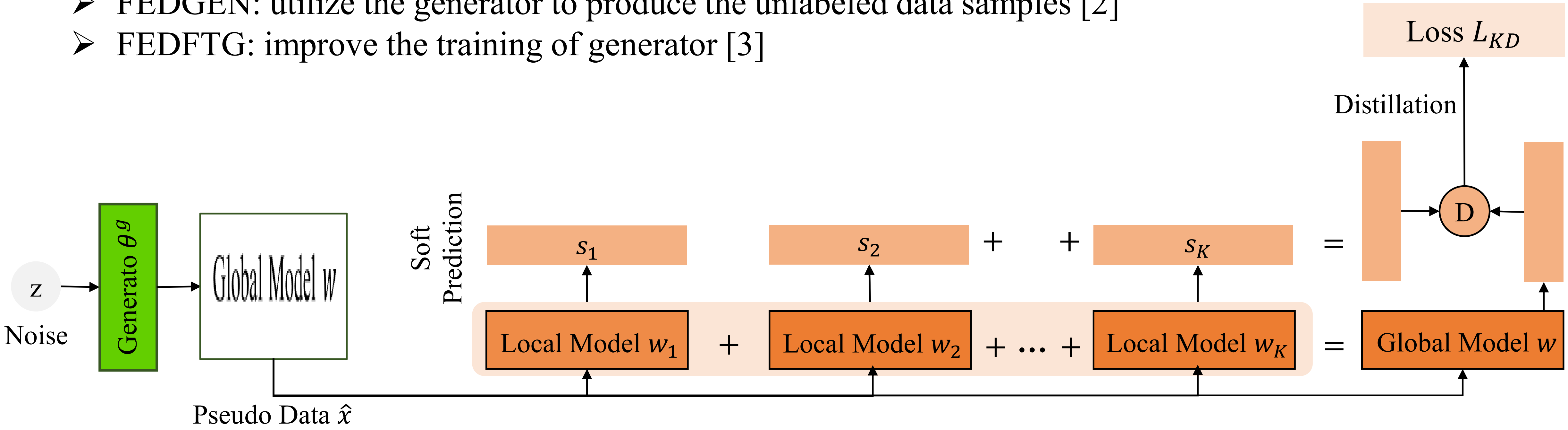
Limitations of Aggregating Parameters

- **Insight**: each local model contains the specific local knowledge in its parameters structure
- **Limitation**: the parameters structure will be broken when they are aggregated into the global model, losing the specific local knowledge
- **Example**: two clients with data samples (x<0, y=0) and (x>0, y=1) respectively. Their parameters may *cancel out* each other when they are aggregated in the server



Federated Knowledge Distillation

- **Federated Knowledge Distillation (FKD)**: global model is obtained by the ensemble distillation of multiple local models for a given sample
- **Existing methods**:
  - FEDDFUSION: utilize unlabeled training samples as the distillation dataset [1]
  - FEDGEN: utilize the generator to produce the unlabeled data samples [2]
  - FEDFTG: improve the training of generator [3]



[1] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in fed- erated learning. Advances in Neural Information Processing Systems, 33:2351–2363, 2020

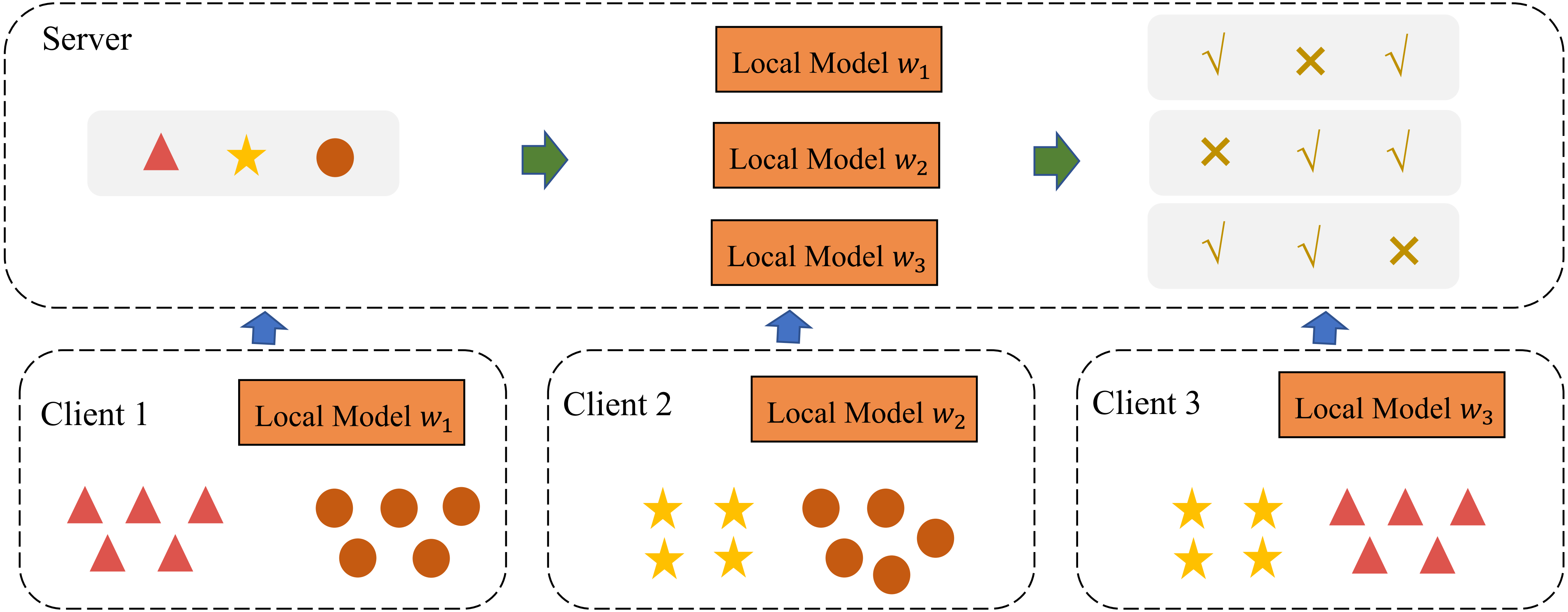
[2] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In International Conference on Machine Learning, pages 12878–12889. PMLR, 2021

[3] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling- Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 10164–10173

Limitations of Existing FKD Methods

- **NonIID**: data is non-identically and independently distributed (NonIID) across multiple clients
  - *Each local model may make mistakes when the input samples are far from its distribution*

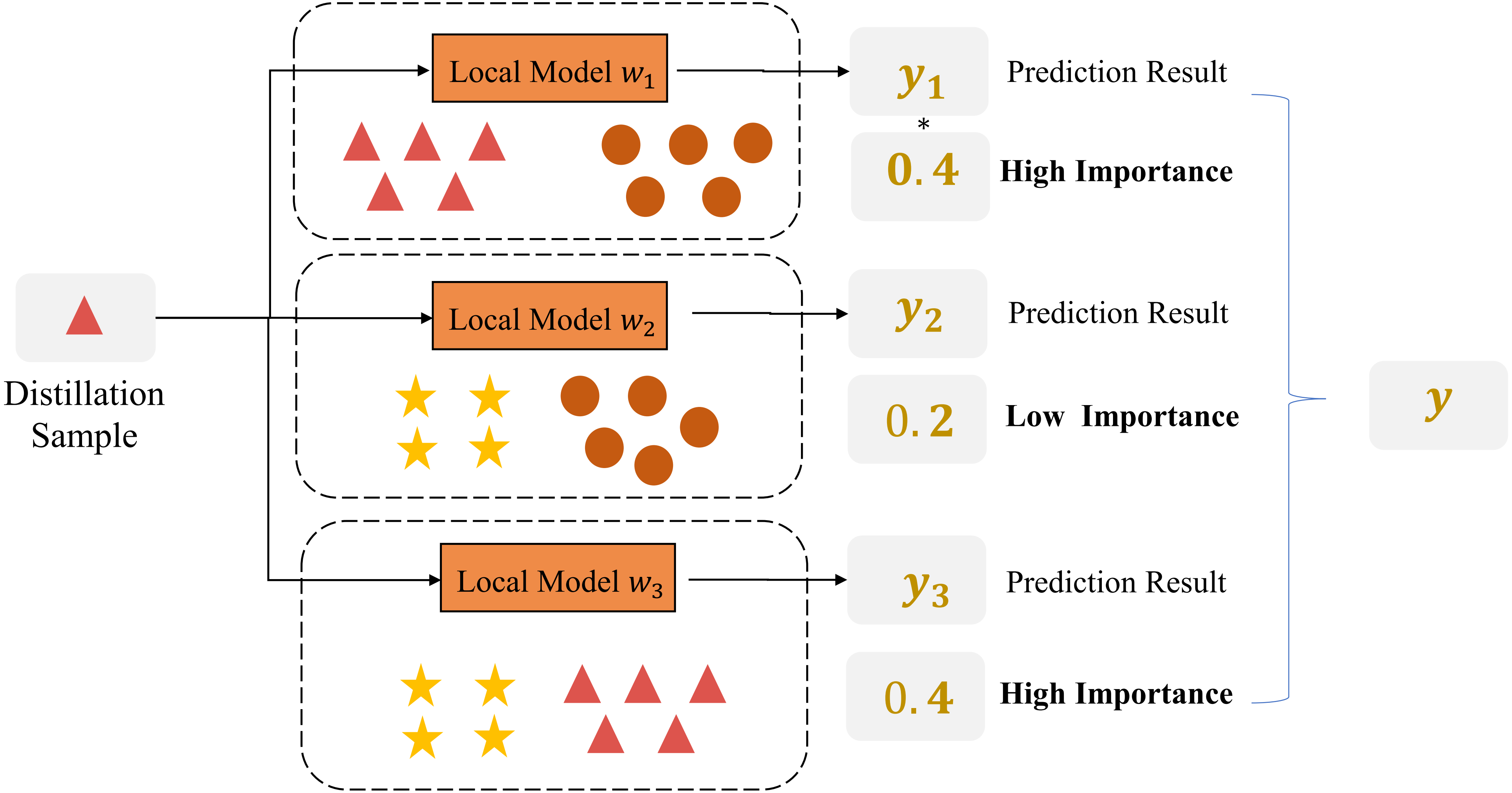
- **Limitation**: existing FKD methods view all local models as equivalent for a given sample, which ignores their training data distribution



Federated Learning

Method: Domain-aware Federated Knowledge Distillation

- **Domain-aware FKD (DaFKD)**: taking NonIID into account when making distillation for the global model
  - *Adaptive Importance*: for any distillation sample, endow each local model with a specific importance factor

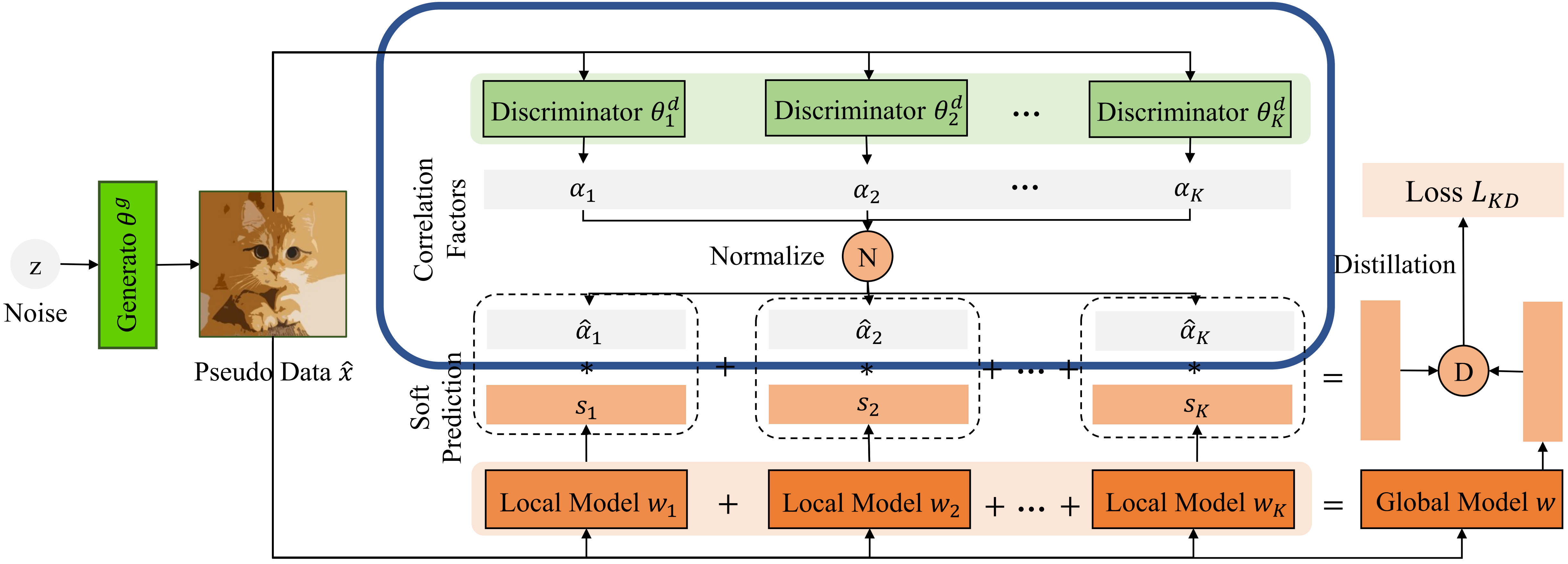


DaFKD: Identifying Importance with Domain Discriminators

- **Domain Discriminator** : identify whether a given sample is close to its local domain
  - **Generator**: *trained in a FedAvg manner*
  - **Discriminator***: personalize for each client and not average*

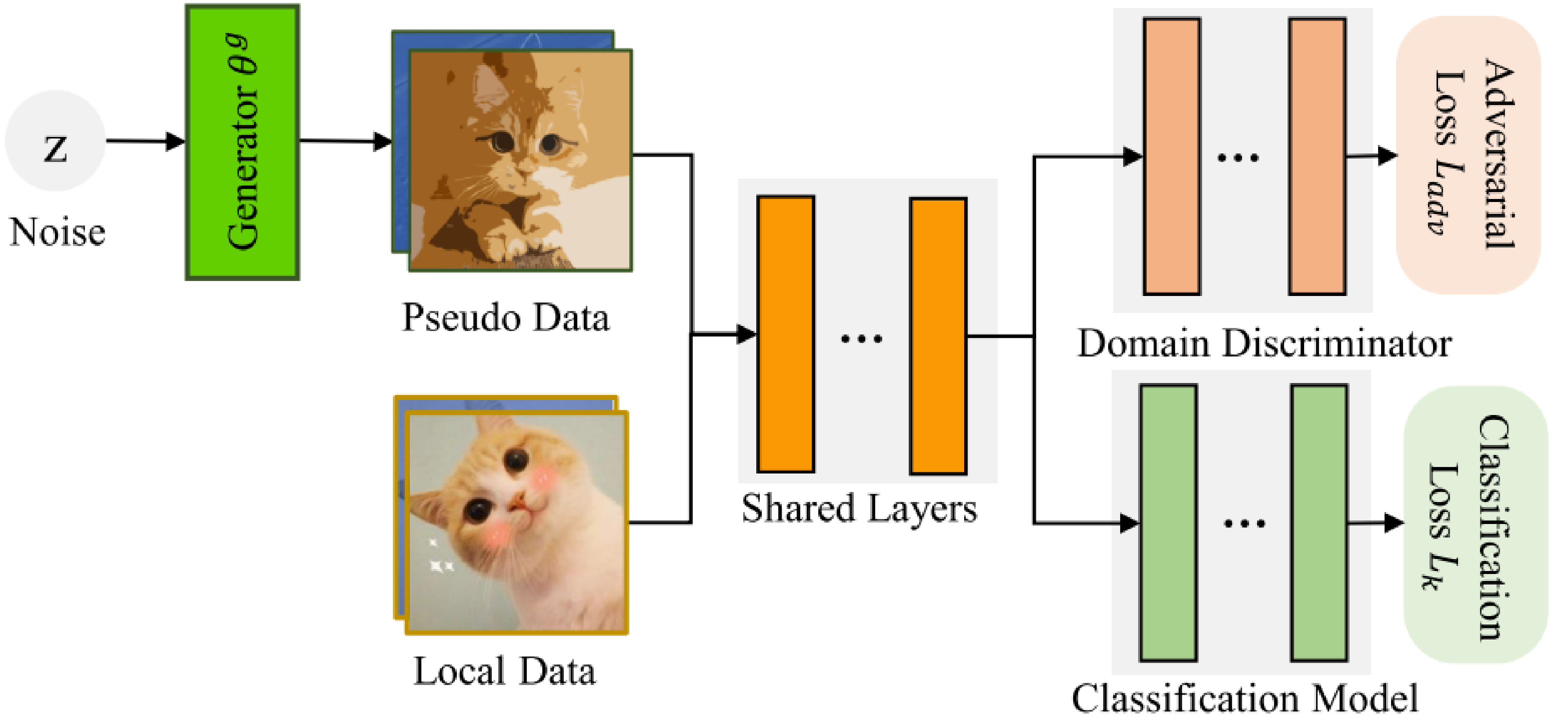
DaFKD: Domain-aware Federated Distillation

- **Importance**: the discriminator outputs an importance factor for each given distillation sample
- **Ensemble Distillation**: all logits are aggregated in a weighted manner with the importance factor as the weight



Partial Parameters Sharing between Discriminator and Target Model

- **Motivation**:
  - *Both the discriminator and the classification model seek to maximize the distinguishability of the samples*
  - *Communication cost can be reduced when the discriminator and classification model share partial layers*
- **Partial Parameters Sharing**: discriminator and classification model share partial shallow layers



Theoretical Guarantee

◆ **Theory**: the upper loss bound of our method is not related to the degree of NonIID

Our method

$$\mathcal{L}_p(\sum_{k=1}^{K} \hat{\alpha}_k(x) h_{\hat{p}_k}) \leq (K+1)\mathcal{L}_{\hat{p}}(h_{\hat{p}}) + (K+1)\sqrt{\frac{\sigma^2 log\frac{2K}{\delta}}{2m}}.$$

No items related to NonIID

Baseline

$$\mathcal{L}_p(\sum_{k=1}^{K} \hat{\alpha}_k(x) h_{\hat{p}_k}) \leq \mathcal{L}_{\hat{p}}(h_{\hat{p}}) \tag{12}$$

$$+\sqrt{\frac{4}{m'}(d\log\frac{2em}{d} + \log\frac{4K}{\delta})} + \boxed{\frac{1}{K}\sum_{i=1}^{K} d_{\mathcal{H}\Delta\mathcal{H}}(p'_k, p) + \lambda'_k}.$$

Two items related to NonIID

Empirical Results

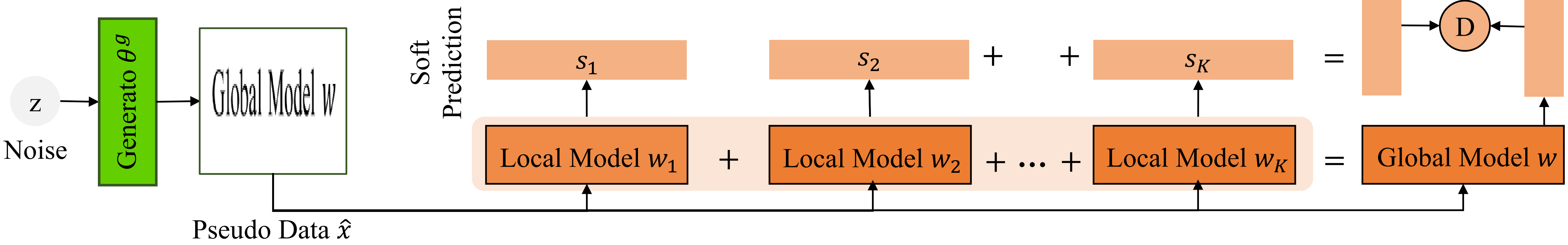
◆ **Results**: highest accuracy among all methods on various datasets and models (improve by up to 6%)

Dataset	Setting	FEDAVG	FEDGEN	FEDFTG	DaFKD
MNIST, E = 20	$\alpha = 0.05$	69.11± 1.39	81.06± 1.09	80.95±1.06	**82.33±0.44**
	$\alpha = 0.1$	95.16± 0.79	94.98± 0.47	94.43±0.49	**95.56±0.41**
	$\alpha = 1$	98.11± 0.14	96.39± 0.90	98.47±0.21	**98.96±0.38**
SVHN, E = 20	$\alpha = 0.05$	33.01± 0.12	47.36± 0.42	48.69±1.87	**51.14±0.16**
	$\alpha = 0.1$	53.54± 0.21	60.03± 1.12	63.75±0.11	**72.80±0.11**
	$\alpha = 10$	81.44± 0.01	82.91± 0.73	83.49±1.32	**87.31±0.85**
FASHION MNIST, E = 20	$\alpha = 0.05$	30.01± 0.54	36.59± 0.98	34.84±0.77	37.85±0.24
	$\alpha = 0.1$	67.97± 0.03	67.29± 2.05	67.25±0.14	**70.81±0.21**
	$\alpha = 10$	82.37± 0.82	81.57± 1.96	81.96±1.86	**83.37±0.06**
EMNIST, E = 40	$\alpha = 0.05$	67.28± 0.14	**68.95±0.88**	67.08±0.97	67.64±1.86
	$\alpha = 0.1$	69.13± 0.23	72.15± 2.04	72.91±1.87	**74.96±0.91**
	$\alpha = 10$	81.35± 1.03	82.02± 1.19	82.65±1.04	**84.60±1.86**

Summary

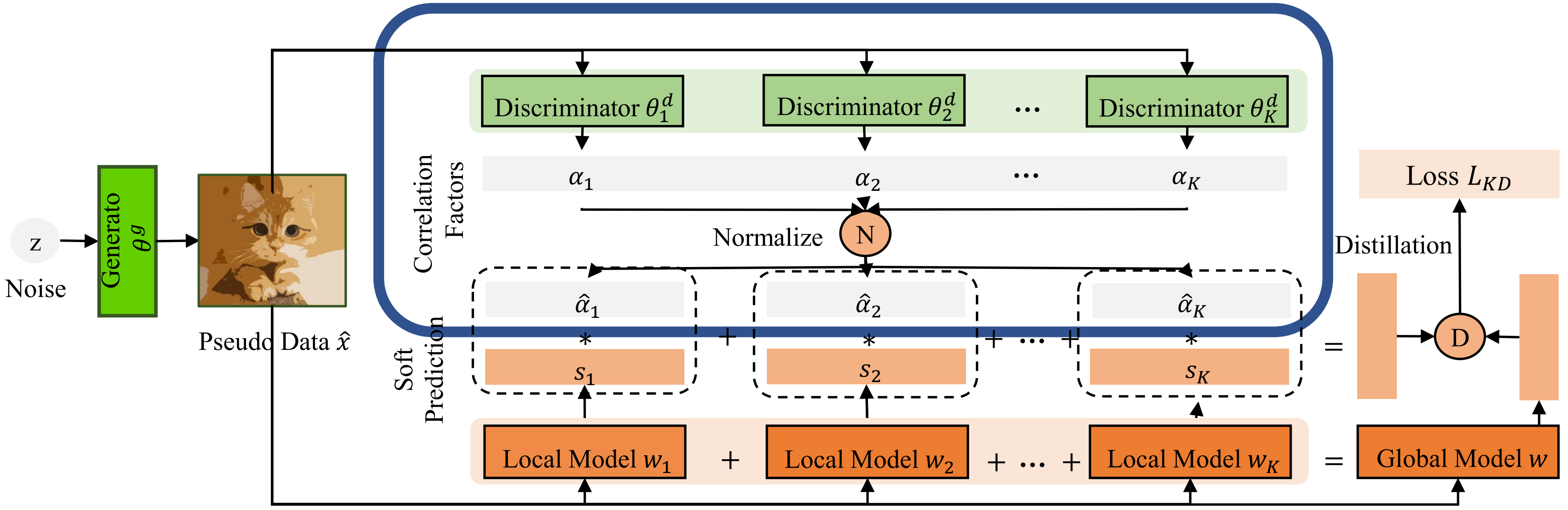
◆ **Challenge**: the data is NonIID in FL settings

◆ **Problem**: existing FKD methods view all local models as equivalent and ignore NonIID data distributions



◆ **Method**: utilize a discriminator to identify the importance of each local model for given distillation sample

◆ **Result**: theoretical guarantee and empirical improvements



JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank You

Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li,
Yufeng Zhan and Zhigang Zeng