



# Fed2PKD: Bridging Model Diversity in Federated Learning via Two-Pronged Knowledge Distillation

**Zaipeng Xie<sup>\*†</sup>, Han Xu<sup>†</sup>, Xing Gao<sup>†</sup>, Junchen Jiang<sup>†</sup>, and Ruiqian Han<sup>†‡</sup>**

*<sup>\*</sup>Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University*

*<sup>†</sup> College of Computer Science and Software Engineering, Hohai University, China*

*<sup>‡</sup> The Hong Kong University of Science and Technology (Guangzhou), China*



**IEEE**



**IEEE  
COMPUTER  
SOCIETY**



**IEEE COMPUTER SOCIETY  
TCSVC**

*Technical community on Services Computing*

# Outline

---

- Background
- Motivation
- Methodology
- Experiment
- Conclusion

## ■ Heterogeneous Federated Learning (HFL)

- ❑ a collaborative machine learning paradigm that enables training across clients with diverse model architectures and data distributions while preserving privacy, adapting to differences in computational capabilities and other factors among participants.

## ■ Challenges

- ❑ **Data heterogeneity:** Dealing with non-IID data distributions across clients, which can hinder the generalizability of global models..
- ❑ **Model heterogeneity:** Managing diverse model architectures and capacities among clients, which complicates model aggregation and knowledge transfer.
- ❑ **Convergence issues:** Ensuring model convergence despite heterogeneous update patterns and learning rates across clients.

## ■ **Prominent Approach:** Knowledge distillation, Personalization strategies, Ensemble and fusion methods, Prototype-based methods

## ■ Local Knowledge Skew

- ❑ The discrepancy between the knowledge learned by a client from its local data and the global knowledge that the federated model aims to capture.

## ■ Impact on Model Performance

- ❑ Undermines the global model's effectiveness
- ❑ Leads to poor convergence and suboptimal performance
- ❑ Causes misalignment with the overall objective of federated learning

## ■ Challenges

- ❑ harmonizing local models with the aggregate global knowledge to achieve consistent and balanced generalization across the network.
- ❑ This issue is particularly significant in heterogeneous federated learning environments where clients have diverse model architectures and data distributions.

# Motivation

---

## ■ Address challenges in Heterogeneous Federated Learning:

- ❑ Tackle model diversity across clients
- ❑ Handle non-IID data distributions

## ■ Enable personalization while preserving privacy:

- ❑ Allow each client to have a personalized model
- ❑ Adapt to unique local distributions and resources
- ❑ Facilitate collaborative learning without sharing raw data

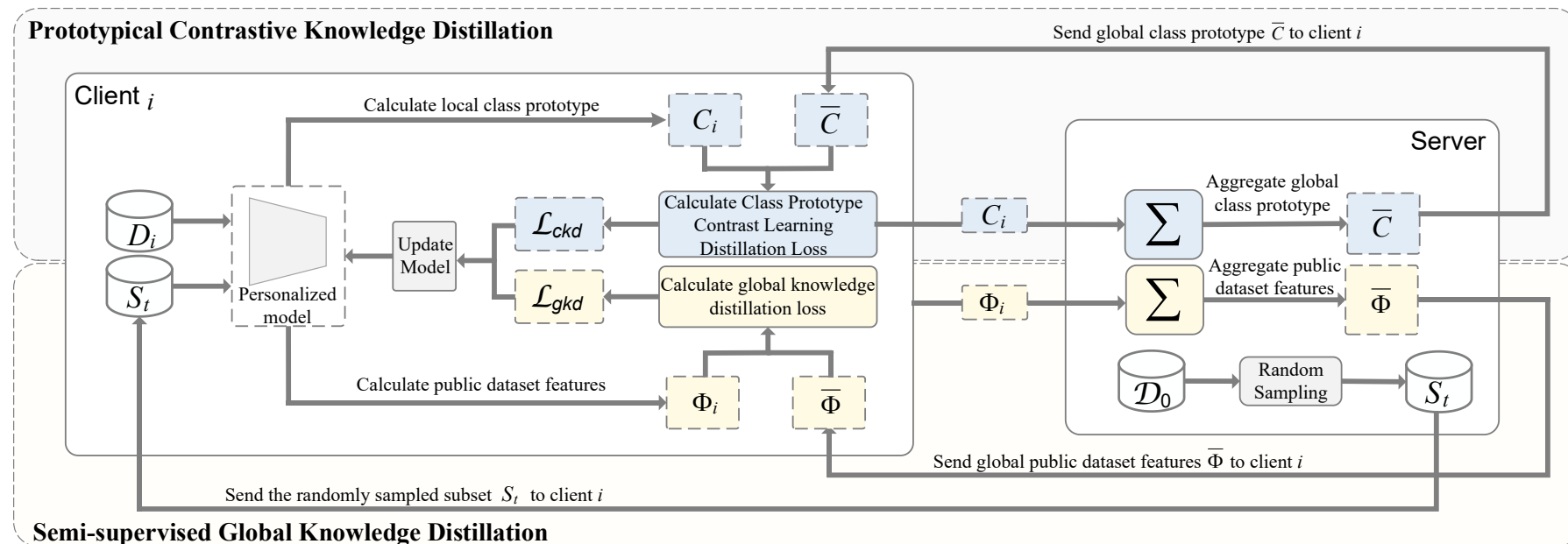
## ■ Improve performance

- ❑ Enhance both local and global model accuracy of HFL
- ❑ Achieve better generalization across diverse client data.

# Method - Overall

## ■ Fed2PKD (Federated Learning via Two-Pronged Knowledge Distillation)

- ❑ **Prototypical Contrastive Knowledge Distillation:** aligns client sample embeddings with corresponding global class prototypes while promoting divergence from other class prototypes, using contrastive loss to improve class differentiation and model performance.
- ❑ **Semi-supervised Global Knowledge Distillation:** aligns local features with global ones using global knowledge distillation, indirectly learning global prototypes not present in their local data.



# Method - Prototypical Contrastive Knowledge Distillation (PCKD)

■ **Goal:** Improve class differentiation by aligning local class prototypes with global ones using contrastive loss.

■ **Contrastive Knowledge Distillation Loss:**

$$\mathcal{L}_{\text{ckd}} = -\log \left( \frac{\exp \left( \text{sim} \left( z_i, \bar{C}^+ \right) / \tau \right)}{\sum_{j \in \mathcal{C}} \exp \left( \text{sim} \left( z_i, \bar{C}^j \right) / \tau \right)} \right)$$

- **Local prototypes**  $C_i^j$  : Centroids of client class embeddings  $\bar{C}^j = \frac{1}{|N_j|} \sum_{i \in N_j} \frac{|D_{i,j}|}{|N_j|} \cdot C_i^j$
- **Global prototypes**  $\bar{C}^j$  : Aggregated from local prototypes  $\text{sim} \left( z_i, \bar{C}^+ \right) = \frac{1}{2} \left( 1 + d_{\cosin} \left( z_i, \bar{C}^+ \right) \right)$
- **Contrastive distillation loss**  $\mathcal{L}_{\text{ckd}}$  : Align client sample embeddings with global class prototypes to improve classification performance.

# Method - Prototypical Contrastive Knowledge Distillation (PCKD) CLOUD 2024

## ■ Implementations

- ❑ **Calculate Local Prototypes:** Clients calculate local class prototypes.
- ❑ **Aggregate Global Prototypes:** Server aggregates local prototypes into global ones.
- ❑ **Knowledge Distillation:** Clients use global prototypes to compute contrastive distillation loss and update their models.

## ■ Benefits:

- ❑ **Enhanced Class Differentiation:** Aligns local embeddings with global prototypes, improving classification accuracy.
- ❑ **Privacy Preservation:** Prototypes are efficiently shared instead of raw data across heterogeneous models.
- ❑ **Balanced Learning:** Ensures embeddings are close to their class prototypes and distinct from others.



## Method - Semi-supervised Global Knowledge Distillation (SGKD)

---

■ **Goal:** Improve the global model's generalization by aligning local client features with global dataset characteristics using semi-supervised learning.

■ **Global Knowledge Distillation Loss:**

$$\mathcal{L}_{gkd} = \frac{1}{R} \sum_{r=1}^R \left( 1 - d_{\cosin} \left( \underset{\substack{\uparrow \\ \text{Local Embedding}}}{\Phi_i^r}, \underset{\substack{\uparrow \\ \text{Global Embedding}}}{\bar{\Phi}^r} \right) \right)$$

■ It aligns local models with global dataset characteristics, enhancing generalization and mitigating local knowledge skew in heterogeneous federated learning environments.

# Method - Semi-supervised Global Knowledge Distillation (SGKD)

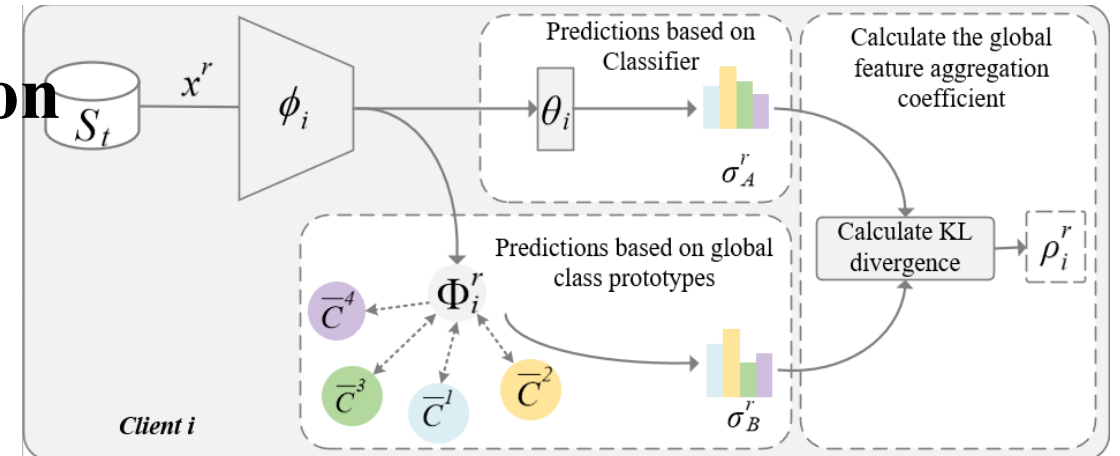
- **Subset Selection:** Server randomly selects a subset  $S_t$  from a public unlabeled dataset  $D_0$  and distributes it to clients.
- **Local Embedding Calculation:** Each client computes embeddings  $\Phi_i^r$  for the samples in  $S_t$ .
- **Global Embedding Aggregation:** Server aggregates these embeddings to form

a global representation  $\bar{\Phi}^r = \frac{1}{\sum_{i \in \{1, \dots, N\}} \rho_i^r} \sum_{i \in \{1, \dots, N\}} \rho_i^r \cdot \Phi_i^r$ .

- **Redistribution of Global Representation**

- **Calculation of Global Knowledge**

**Distillation Loss and model update**



# Method – Semi-supervised Global Knowledge Distillation (SGKD)

---

## ■ Public Dataset

- ❑ **Definition:** An unlabeled public dataset  $D_0$  shared among all clients.
- ❑ **Purpose:** Acts as a baseline to ensure consistent feature learning across clients.

## ■ Benefits:

- ❑ **Enhanced Generalization:** Aligns client models with global data characteristics.
- ❑ **Reduction of Local Knowledge Skew:** Mitigates discrepancies between local and global knowledge.
- ❑ **Improved Consistency:** Ensures model predictions are consistent across diverse data distributions.

## Method – Overall Training

---

■ Fed2PKD unifies model training by synchronizing individual prototypes and minimizing local and global losses, aiming to reduce local knowledge skew and improve global knowledge transfer. The composite objective is:

$$\arg \min_{\mathbf{w}} \mathcal{L} = \sum_{i=1}^N \frac{|D_i|}{|D|} \mathcal{L}_i,$$

$$\text{where } \mathcal{L}_i = \mathcal{L}_{ce} + \lambda \cdot \mathcal{L}_{ckd} + \mu \cdot \mathcal{L}_{gkd}.$$

■ Benefits:

- ❑ **Complementary Focus:** PCKD ensures class-specific feature alignment, while SGKD ensures global data distribution understanding.
- ❑ **Improved Generalization:** Combines local feature consistency with global knowledge to enhance overall model performance in heterogeneous federated learning.

## Method – Theoretical Analysis

■ **Theorem 1: Convergence:** Under the assumptions of bounded gradients and smooth loss functions, the Fed2PKD algorithm converges to a global optimum.

□ Bounded Gradients

$$\mathbb{E}[\mathcal{L}_i(w(t+1))] \leq \mathbb{E}[\mathcal{L}_i(w(t))]$$

□ Smooth Loss Functions

$$-\eta \left(1 - \frac{\eta}{2} L_s\right) \|\mathbb{E}[\nabla \mathcal{L}_i(w(t))]\|^2$$

□ Convexity of the Loss Function

■ **Theorem 2: Stability:** The Fed2PKD algorithm remains stable during training under the given assumptions, ensuring that the model monotonically converges.

□ Controlled Learning Rate

□ Bounded Variance of Updates

□ Client Synchronization

$$\sum_{i=1}^N \mathbb{E}[\mathcal{L}_i(w(t+1))] \leq \sum_{i=1}^N \mathbb{E}[\mathcal{L}_i(w(t))],$$

# Experiments - Setup

	Experimental Settings
Applications	CNN for MNIST handwritten digit recognition task, ResNet for CIFAR-10 and CIFAR-100 image classification task.
Dataset Non-IID processing	20 client nodes, the sample is unbalanced, the n-way k-shot dataset is divided, three data distributions are taken for each dataset, and each client is tested on the local and global data distribution
Model design	CNN: small (2 convolutional layers, 2 fully connected layers), medium (3 convolutional layers, 2 fully connected layers), or large (3 convolutional layers, 3 fully connected layers) ResNet: ResNet50, ResNet101, or ResNet152
Experimental Conditions	n-way k-shot Setting: Few-shot learning setting where n is the number of classes and k is the number of samples per class.
Training rounds	300 rounds for Global and Local tasks

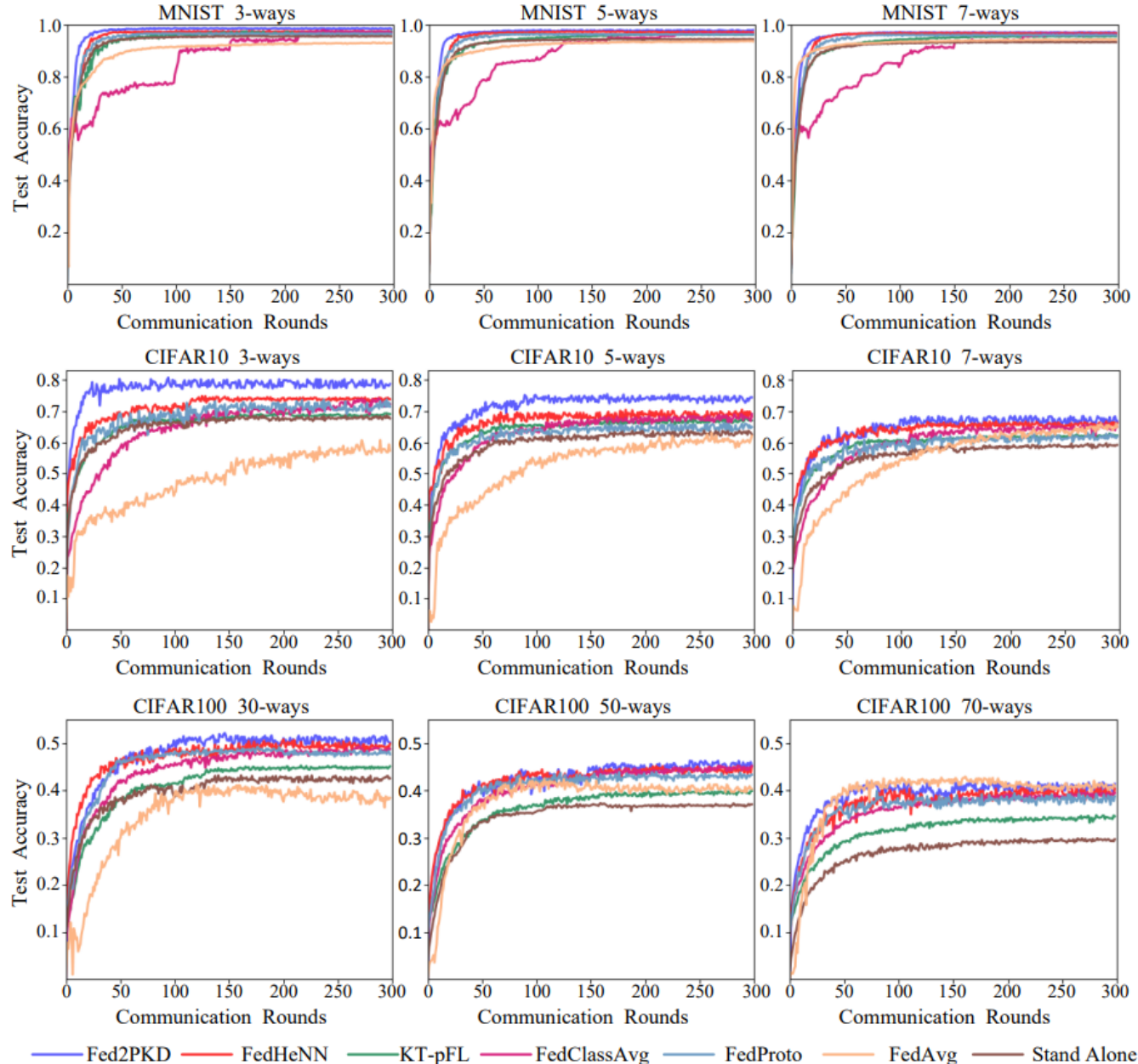
# Experiments – Baselines

---

- **FedAvg:** A pioneering method that aggregates model updates from distributed clients in a non-heterogeneous setting to enhance the global model.
- **FedHeNN:** Designed to aggregate learning from clients with heterogeneous model architectures, FedHeNN represents an advanced solution for managing architectural diversity in federated learning systems.
- **FedClassAvg:** It improves personalization by averaging different clients' classifier layers, enhancing local learning and performance without additional datasets, and ensuring efficient communication.
- **KT-pFL:** It employs parameterized models to adapt to varying client data distributions, thus enhancing learning across different datasets and promoting personalized model training in federated learning settings.
- **FedProto:** This method aims to tackle model heterogeneity challenges by shifting from traditional data or model sharing to the exchange of data distribution prototypes



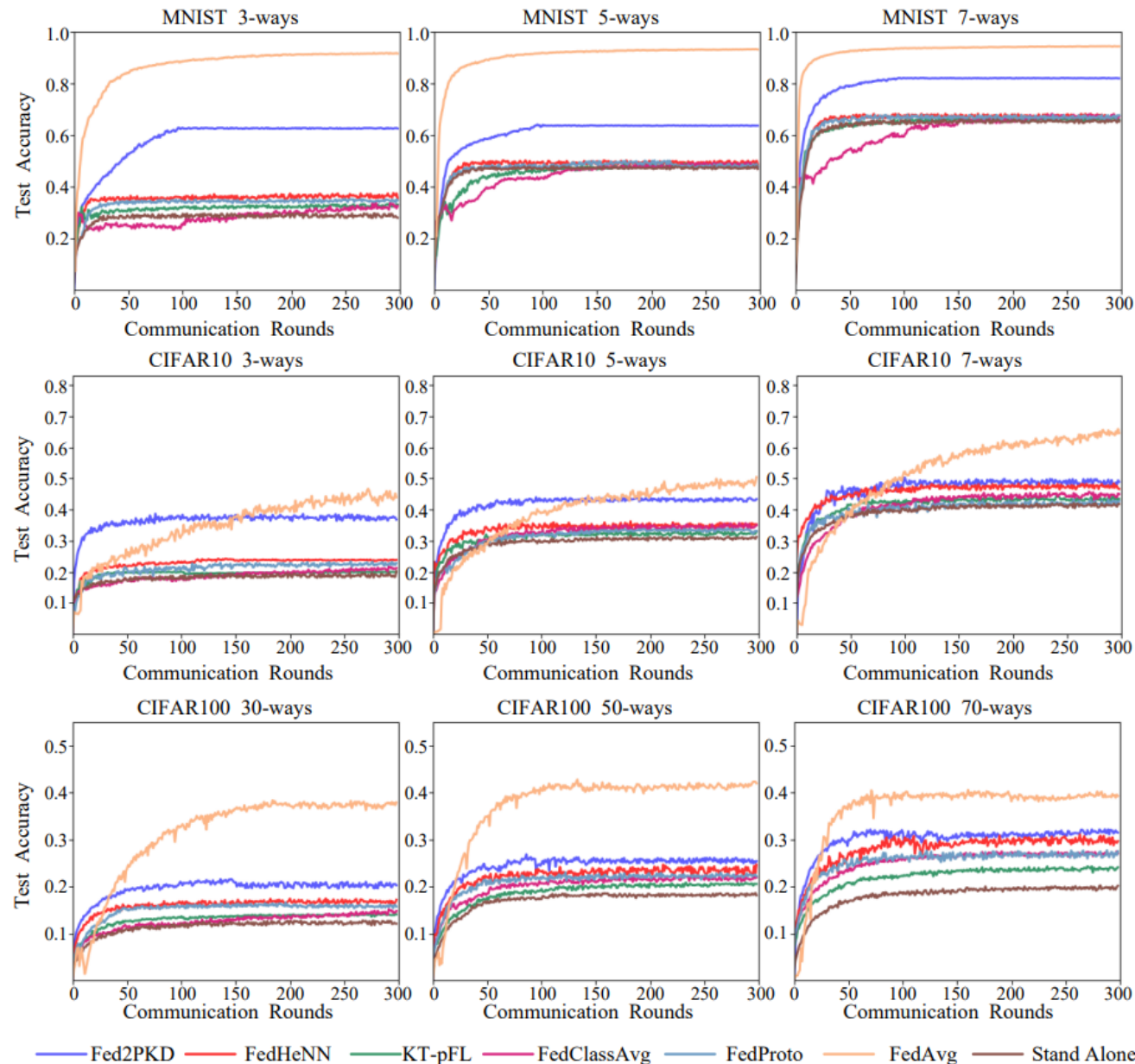
# Experimental Results – Local tasks



- **Consistent Performance:** Fed2PKD consistently outperforms baselines across all datasets.
- **Superior Accuracy:** Significant improvements over the top baselines
- **Adaptability:** Handles diverse local data distributions effectively.



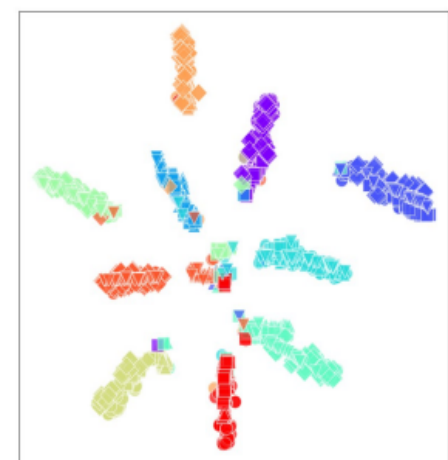
# Experimental Results – Global Tasks



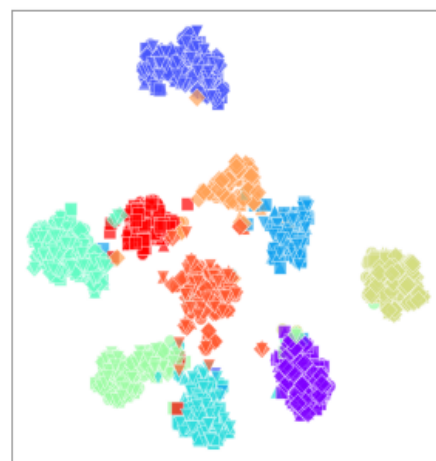
- **Accuracy Improvement:** Fed2PKD shows notable accuracy improvements in global tasks over all baseline methods except FedAVG.
- **Scalability:** Performance gains are consistent with increasing complexity and heterogeneity
- **Better Generalization:** Fed2PKD effectively manages global data characteristics, leading to improved generalization.

# Experimental Results – tSNE

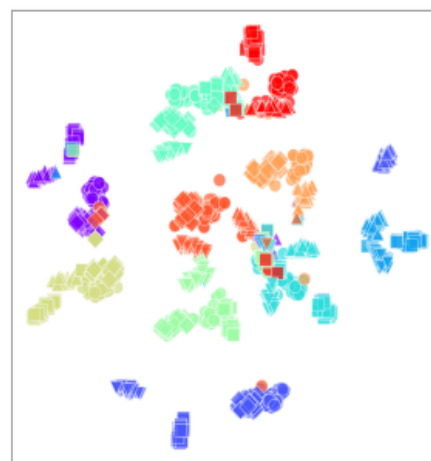
## ■ t-Distributed Stochastic Neighbor Embedding



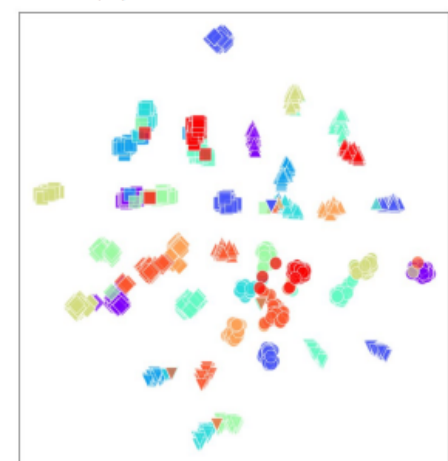
(a) Fed2PKD



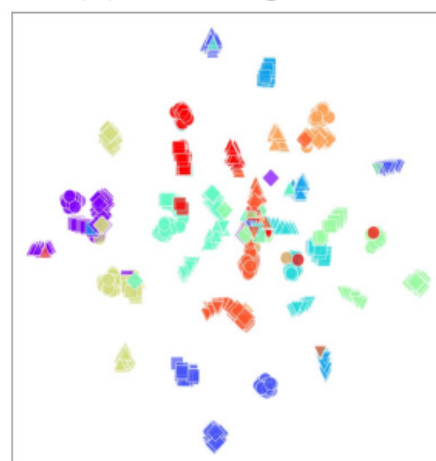
(b) FedAvg



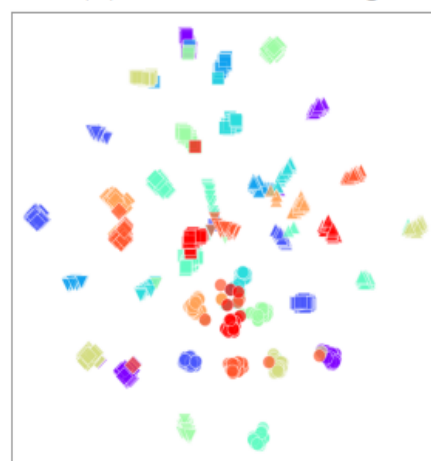
(c) FedClassAvg



(d) FedHeNN



(e) FedProto



(f) KT-pFL

- Class 0
- Class 1
- Class 2
- Class 3
- Class 4
- Class 5
- Class 6
- Class 7
- Class 8
- Class 9
- Client 0
- ▲ Client 1
- Client 2
- ◆ Client 3
- ▼ Client 4

## ■ Feature Embeddings:

Fed2PKD are more distinct and well-clustered compared to baseline methods.

## ■ Class Separation: Highlight the improved separation of classes by Fed2PKD

## ■ Consistency Across Clients: Emphasize that the consistent feature representation across clients leads to better generalization and model performance.

# Experimental Results – Ablation Study

---

## ■ Impact of Prototypical Contrastive Knowledge Distillation (PCKD)

- ❑ Removing PCKD significantly decreases model accuracy, demonstrating its crucial role in aligning local and global prototypes.
- ❑ Accuracy **drops by 15%** on average for global tasks when PCKD is removed.

## ■ Impact of Semi-supervised Global Knowledge Distillation (SGKD)

- ❑ Removing SGKD leads to a significant drop in performance, highlighting its importance in capturing global data characteristics.
- ❑ Accuracy **drops by 25%** on average for global tasks when SGKD is removed.

## ■ Both PCKD and SGKD are critical for the effectiveness of Fed2PKD.

# Experimental Results – Ablation Study

---

- Temperature Parameter ( $\tau$ ): Optimal range: 0.05 - 0.1
  - Too high or too low values lead to reduced performance.
- Contrastive Loss Weight ( $\lambda$ ): Optimal value: 0.9
  - Lower values reduce the impact of contrastive loss, leading to poorer alignment of local embeddings with global prototypes.
- Global Knowledge Distillation Loss Weight ( $\mu$ ): Optimal value: 0.9
  - Ensures a balanced contribution of global knowledge distillation in the overall loss function.
- Public Dataset Subset Size ( $R$ ): Optimal range: 20-50 samples
  - Smaller subsets fail to provide sufficient global knowledge, while larger subsets do not significantly improve performance and increase computation.
- non-IID-ness among client datasets ( $\alpha$ ): Optimal range: 5-10
  - This approach allows the researchers to analyze the impact of varying degrees of distribution drift between public and client datasets on the performance of both local and global models.

# Discussion

---

## ■ Strengths of Fed2PKD

- **Effective Knowledge Transfer:** Combines local and global knowledge through prototypical contrastive and semi-supervised global knowledge distillation.
- **Privacy Preservation:** Achieves high performance without sharing raw data, maintaining data privacy across clients.
- **Scalability:** Scalable to heterogeneous model architectures and data distributions.

## ■ Challenges and Limitations

- **Computational Overhead:** Potential increased computational overhead due to the additional distillation processes.
- **Hyperparameter Sensitivity:** Performance is sensitive to the choice of hyperparameters ( $\tau$ ,  $\lambda$ ,  $\mu$ , and  $R$ ), requiring careful tuning.
- **Public Dataset Requirement:** Dependence on a suitable public unlabeled dataset for semi-supervised global knowledge distillation.

# Conclusions

---

- This study introduces Fed2PKD, a novel approach in federated learning designed to address model diversity with data heterogeneity, offering personalized client architectures.
- Fed2PKD incorporates a dual knowledge distillation mechanism to foster collaborative learning among diverse devices.
- Fed2PKD overcomes the inherent complexities of heterogeneous federated learning by enabling flexible model customization and promoting uniform knowledge distribution.
- Experimental analyses confirm Fed2PKD's advantages, highlighting the benefits of embracing model and data diversity.
- Future work will optimize computational efficiency, adapt hyperparameter tuning, extend applicability to diverse data types, and for broader real-world implementations.





# IEEE CLOUD 2024

Thanks for listening!

Presenter: Zaipeng Xie



IEEE



IEEE  
COMPUTER  
SOCIETY



IEEE COMPUTER SOCIETY  
**TCSVC**

*Technical community on Services Computing*