IJCAI 2025 MONTREAL

FedCPD: Personalized Federated Learning with Prototype-Enhanced Representation and Memory Distillation

Kaili Jin[1], Li Xu[1], Xiaoding Wang[1*], Sun-Yuan Hsieh[2], Jie Wu[3,4], Limei Lin[1*]

[1]Fujian Normal University

[2]National Cheng Kung University

[3]China Telecom Cloud Computing Research Institute

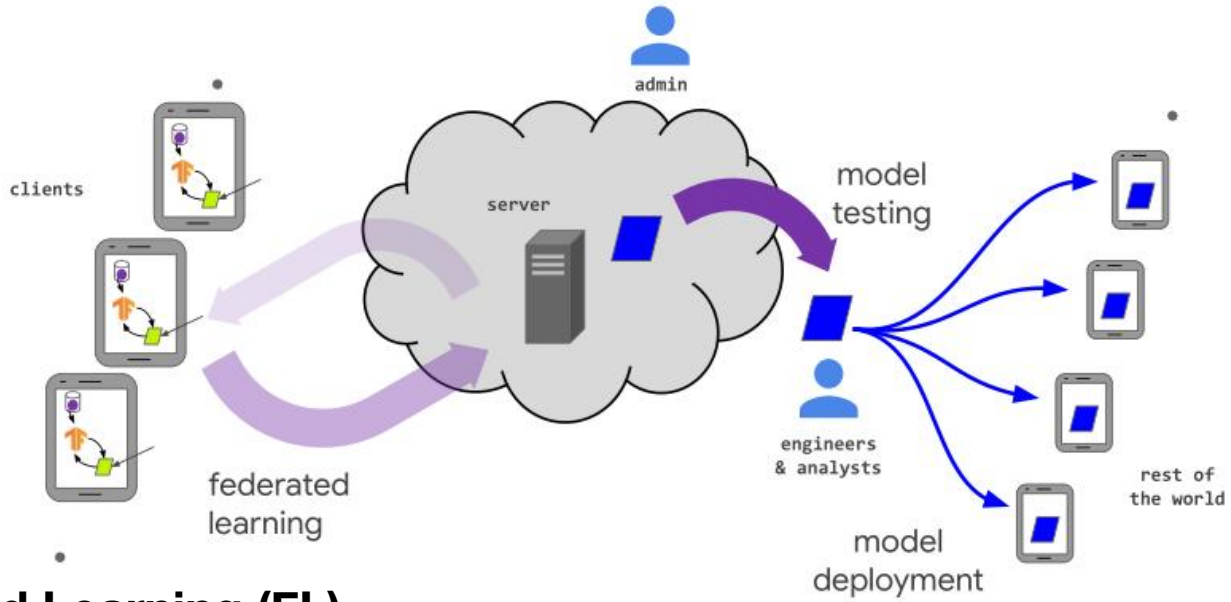
[4]Temple University

福建師範大學 FUJIAN NORMAL UNIVERSITY

國立成功大學 National Cheng Kung University

TEMPLE UNIVERSITY

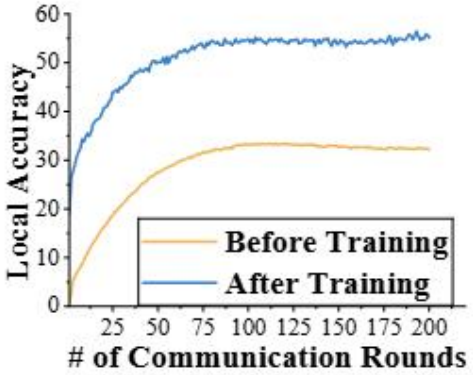
Challenges:



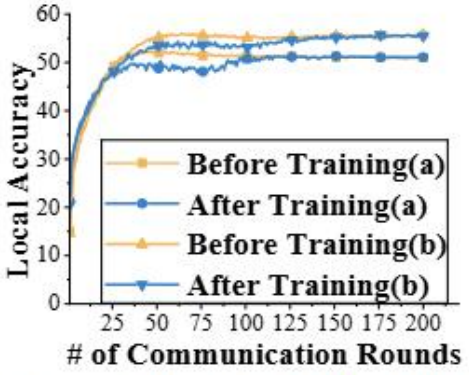
- **Federated Learning (FL)**
  - On-device training; data stays local → privacy
  - **Non-IID** clients → one global model misfits individuals
- **Personalization challenges**
  - **Historical forgetting:** "drop on receive" (global ↓ local accuracy)
  - **Weak generalization:** few/biased local data → overfitting, fuzzy boundaries, poor transfer
- **Limits of prior routes**
  - **Parameter decoupling:** local head only; **extractor stays globally biased** → local nuances lost
  - **Prototype sharing:** helps generalization, **doesn't preserve client history**

Motivation: Historical Information Forgetting

- **Setting:** CIFAR-100, **20** non-IID clients
- **Color code: yellow = received global**, **blue = previous local**
- **Observation:** persistent **"drop on receive"** (global < previous local) → loss of personalization
- **FedRep limitation:** local head kept, but **shared extractor stays global-biased** → lower post-training ceiling
- **Need:** **carry last-round local features forward** to reduce update-induced forgetting



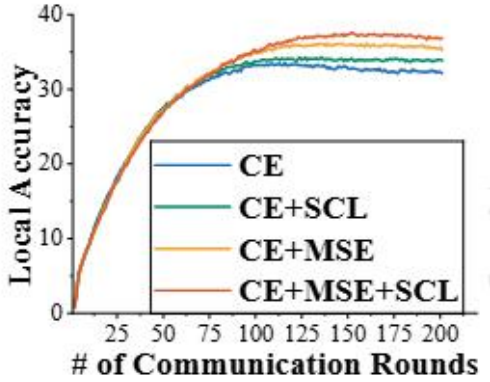
(a) FedAvg



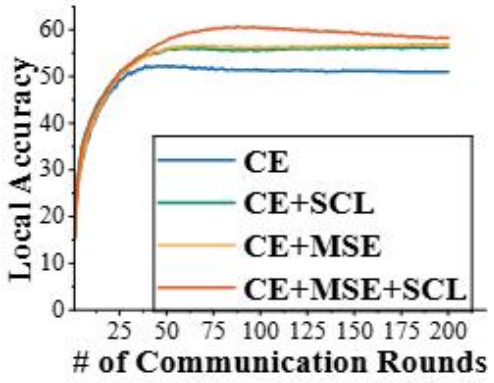
(b) a: FedRep, b: FedRep+FD

Motivation: Limited Generalization Performance

- **Setup:** CE vs **+ alignment** (pull to class center) vs **+ contrast** (push from other classes) vs **both**
- **Fig. 2c (FedAvg):** alignment already yields a clear gain; contrast also helps
- **Fig. 2d (FedRep): alignment + contrast** is **best** (faster, more stable)
- **Need:** inject **class-level consistency + discriminability** to improve **generalization** across clients

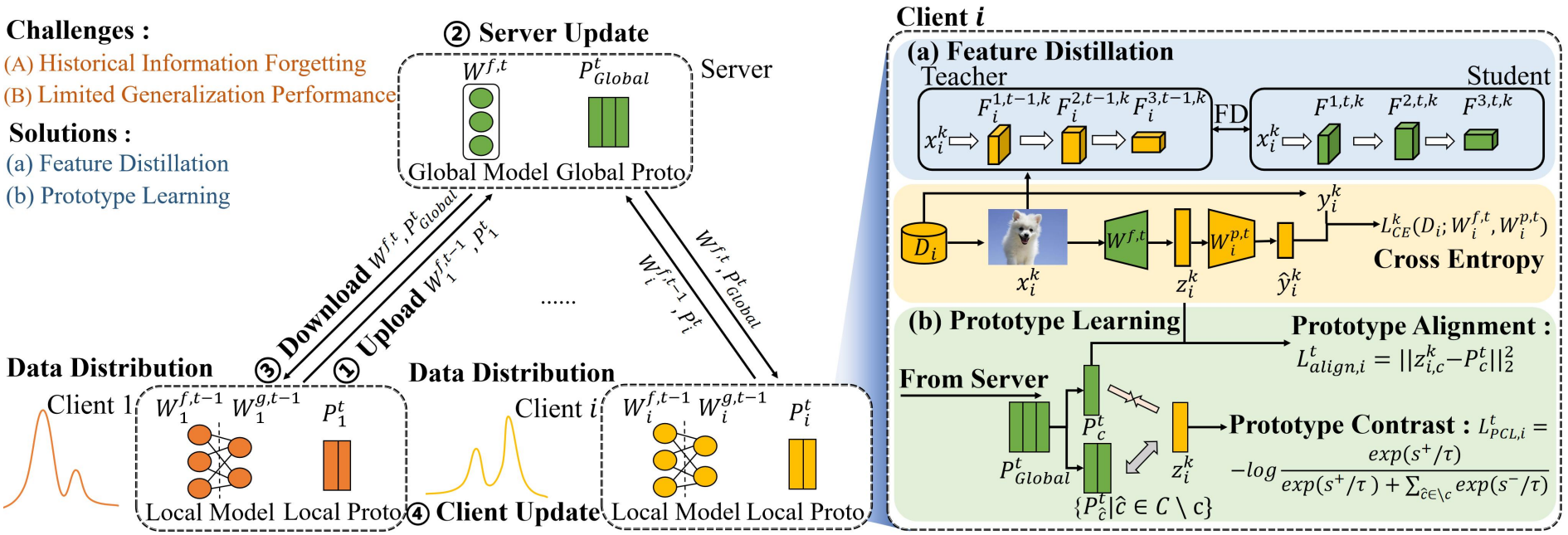


(c) FedAvg



(d) FedRep

Framework Overview



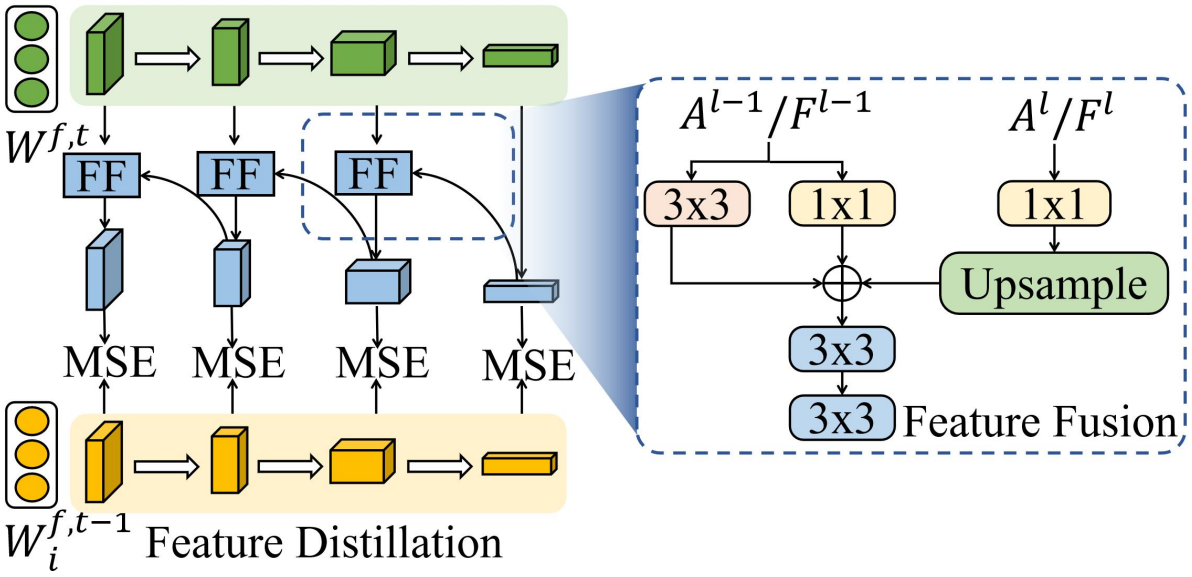
- **Therefore, we tackle historical forgetting and weak generalization through (a) feature distillation, which preserves past knowledge, and (b) prototype learning, which sharpens class separation and improves generalization.**

Feature Distillation

Feature Distillation: Preserve client history, reduce forgetting

Goal: mitigate post-aggregation forgetting and keep client-specific knowledge.

Setup: keep previous local extractor as **Teacher** $W_i^{f,t-1}$; current extractor as **Student** $W_i^{f,t}$.



① **Attention guidance:** apply **CBAM** to each layer's feature map F^l to get attention maps A^l (channel + spatial).

$$A^l = CBAM(F^l)$$

② **Cross-layer fusion (top-down):** three convs — Cv_1 (channel unify 1x1), Cv_3(enhance low-level 3x3), Cv_2 (post-fusion smoothing 2*3x3).

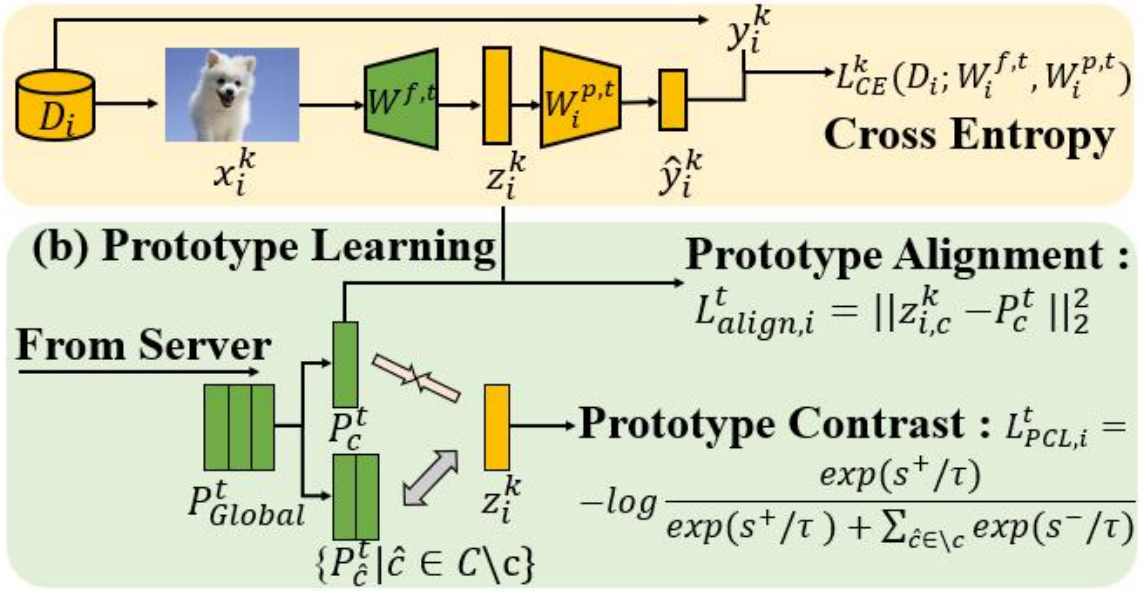
$$A^{l,*} = Cv_2(upsamp\left(Cv_1(A^{l+1})\right) + Cv_1(A^l) + Cv_3(A^l))$$

③ **Alignment loss (MSE):**

$$L_{fd} = ||A_s^{l,*} - A_t^l||_2^2 + ||F_s^{l,*} - F_t^l||_2^2$$

Effect: smoother global→local transition, less "drop on receive," higher personalized accuracy.

Prototype Learning



Objective: enforce **class-level consistency** and **discriminability** across clients.

Embeddings & prototypes: for sample $x_{i,j}$ with label c: $z_{i,j} = f(W_i^{f,t}, x_{i,j})$, global prototype P_c^t

Alignment (pull to class center): $L_{align,i}^t = ||z_{i,j} - P_c^t||_2^2$

Prototype contrast (push from others): with s(· ,·) cosine sim., temp. τ

$$L_{PCL,i}^t = -\log \frac{exp(s(z_{i,j}, P_c^t)/\tau)}{exp(s(z_{i,j}, P_c^t)/\tau) + \sum_{\hat{c} \in \backslash c} exp(s(z_{i,j}, P_{c'}^t)/\tau)}$$

Effect: tighter **intra-class**, larger **inter-class** margins → **faster convergence & stronger generalization**, esp. with sparse per-class data.

Main Results

IJCAI 2025 MONTREAL

Method	Practical heterogeneous ($\beta = 0.1, N = 20$)						Pathological heterogeneous ($N = 20$)					
	FMNIST		CIFAR10		CIFAR100		FMNIST		CIFAR10		CIFAR100	
	Acc.	Std.	Acc.	Std.	Acc.	Std.	Acc.	Std.	Acc.	Std.	Acc.	Std.
Local	97.21	5.57	88.91	11.39	47.77	4.93	93.78	3.02	88.74	5.52	64.13	5.96
FedAvg	85.86	11.87	57.24	12.48	32.35	3.72	87.19	3.25	57.34	11.39	27.10	4.72
FedProx	85.81	11.87	57.17	12.72	32.58	3.80	87.34	3.41	57.36	12.27	27.23	4.80
FedPer	97.64	5.11	90.48	9.45	49.45	4.45	95.27	2.30	90.93	4.82	65.68	4.24
FedRep	97.65	4.42	90.56	9.51	51.01	4.15	95.31	2.04	91.27	4.23	68.27	4.25
FedGH	96.32	11.46	83.06	20.89	49.08	5.06	93.74	3.10	88.80	5.45	65.49	5.71
FedPA	97.20	5.60	90.23	9.89	52.36	4.73	93.76	3.87	90.73	5.23	67.59	5.47
FedProto	97.43	5.77	89.99	10.29	51.63	4.80	93.94	3.34	89.65	5.78	67.41	5.51
FedALA	97.79	4.30	91.02	9.04	55.33	4.03	95.58	2.14	91.57	4.17	67.31	3.66
FedCPD	**97.83**	4.83	**91.78**	**7.87**	**60.23**	3.99	**95.73**	2.19	**91.67**	**3.89**	**72.85**	**2.78**

Table 1: The average test accuracy of the 3 datasets in the real-world environment and the 3 datasets in the pathological heterogeneous environment, as well as the average standard deviation of the accuracy across all clients.

- **Setup:** FMNIST / CIFAR-10 / CIFAR-100; Dirichlet and class-per-client splits; metric: mean client test accuracy; baselines in **Table 1.**
- **Result:** **FedCPD achieves the highest mean accuracy** on all datasets under **both** regimes.
- **Trend:** Larger gains with **more classes / stronger imbalance.**
- **Rationale:** **FD** reduces post-aggregation forgetting; **prototypes** increase intra-class compactness & inter-class separability.

Main Results

- **Heterogeneity control (Table 2):**
  - CIFAR-10: vary Dirichlet **β** to adjust real-world non-IID level
  - CIFAR-100: vary **classes per client** to create pathological non-IID

- **Result:** **FedCPD attains the highest mean test accuracy** under both heterogeneity controls (Table 2) → robust generalization across distributions

- **Fairness (Tables 1-2):** **Lower standard deviation** across clients; FedCPD consistently ranks among the best

- **Interpretation:** Better balance between **global sharing** and **local personalization** → accuracy **and** fairness

Method	CIFAR10		CIFAR100	
	$\beta = 0.5$	$\beta = 1$	cls./clt.=20	cls./clt.=50
	Acc.	Acc.	Acc.	Acc.
Local	88.91	61.23	48.04	31.06
FedAvg	57.24	70.61	30.17	32.05
FedProx	57.17	70.60	30.13	31.62
FedPer	90.48	69.13	51.62	35.59
FedRep	90.56	70.62	53.65	36.77
FedGH	83.06	62.46	48.04	32.16
FedPA	84.90	72.16	52.74	36.19
FedProto	90.40	63.36	50.78	32.88
FedALA	81.15	77.03	54.99	38.27
FedCPD	**92.24**	**79.14**	**62.54**	**48.67**

Table 2: The test accuracy with changes to the β of CIFAR-10 for the real-world heterogeneity evaluation and the label classes for each client (cls./clt.) in CIFAR-100 for the pathological heterogeneity evaluation.

Conclusion:

1.We propose FedCPD, a personalized FL framework that combines attention-guided hierarchical feature distillation and prototype alignment/contrast.

2.We provide theoretical support (convergence upper bound): FD mitigates aggregation-induced forgetting; prototypes enforce intra-class compactness and inter-class separability.

3.Comprehensive experiments on FMNIST/CIFAR-10/100 with Dirichlet and class-per-client splits show state-of-the-art personalized accuracy and robustness — up to +10.40% (generalization) and +4.90% (personalization), with lower client-wise variance.