Federated Oriented Learning (FOL): A Practical One-Shot Personalized Federated Learning Framework

Guan Huang and Tao Shu

ICML 2025

July 13, 2025

Presenter: Guan Huang

AUBURN UNIVERSITY
Samuel Ginn College of Engineering

Introduction and Motivation

1. **What is Personalized Federated Learning (PFL)?**
   - Each client trains a local model on its own private data.
   - Shared knowledge is used to adapt a global model into a personalized one for each client.
   - Results in higher accuracy on non-IID local distributions than a single global model.
2. **Why is one-shot PFL Essential?**
   - **Intermittent Connectivity (e.g., LEO satellites, remote IoT):** clients often get only one brief window to exchange models.
   - **Communication Cost:** multiple rounds increase total data transfer and latency—especially costly on low-bandwidth or energy-limited links.
   - **Need for Local Personalization:** with a single exchange—using alignment, pruning, and distillation—clients can still achieve high local accuracy without multiple back-and-forths.

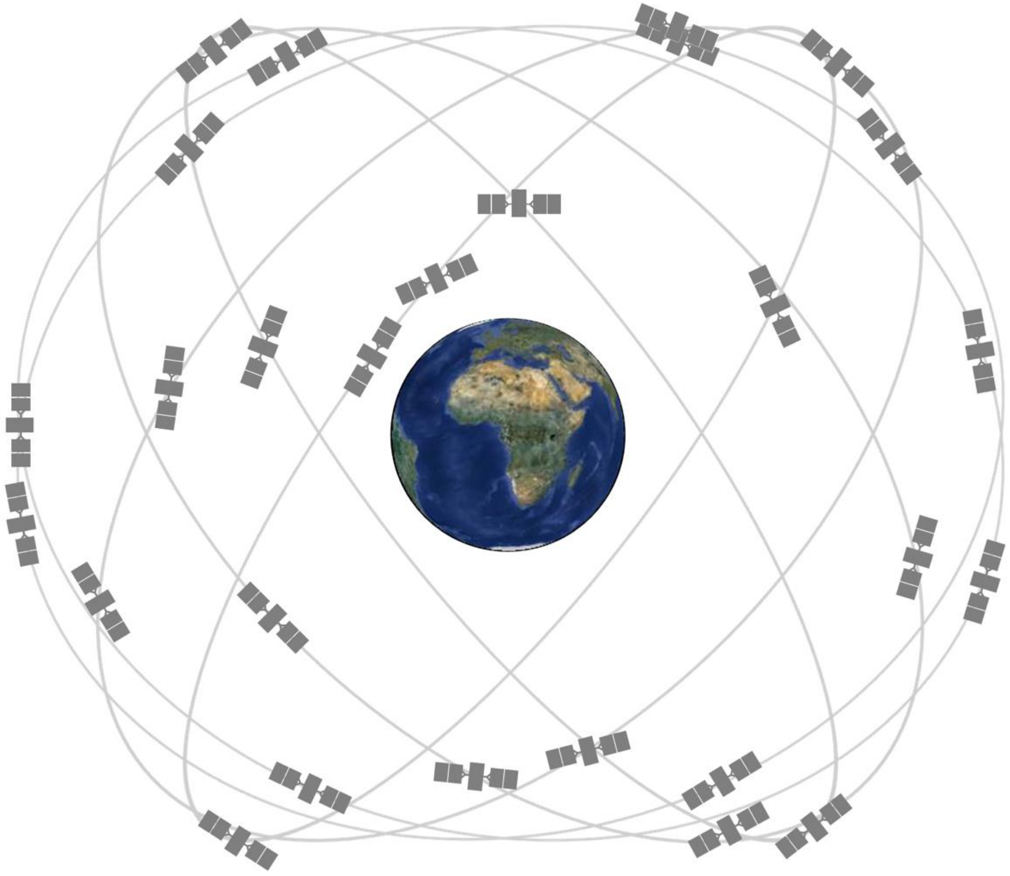

Fig 1. Image Source: Screenshot from https://www.gps.gov/multimedia/images/constellation.jpg. Retrieved [6/4, 2025].

Related Work & Limitations

1. **One-Shot Federated Learning**
   - **Synthetic Data Methods**
     - Server synthesizes proxy data to train a single global model in one communication round.
   - **Ensemble/Distillation Methods**
     - Clients send snapshots once; server ensembles or distills into a single global model.
   - **Limitation**
     - All clients receive the same model — no per-client personalization.

2. **Personalized Federated Learning (PFL)**
   - **Optimization-Based**
     - Clients refine a shared model over multiple rounds using regularization, dynamic aggregation, or second-order updates.
   - **Parameter-Decoupling**
     - Divides the model into a shared backbone and a local head, requiring multiple rounds of updates to obtain a tailored model.
   - **Limitation.**
     - Require repeated communication — impractical for LEO satellites or IoTs with limited contact

FOL Overview

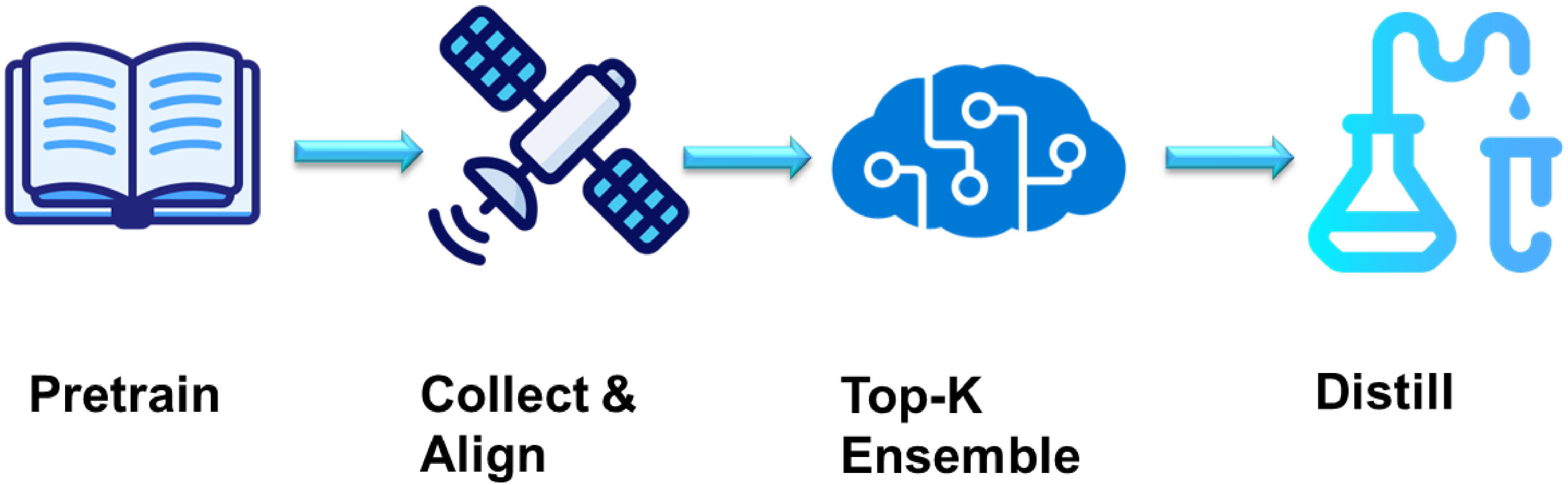

Fig. 2 Architecture Overview

Model Alignment Process (1)

- **Fine-Tuning:** Each received neighbor model $\{\phi_j^{(e)}\}_{j=1}^{Q}$ is fine-tuned on the local training data:

$$\phi'^{(e)}_{j\to k} = \arg\min_{\phi} \frac{1}{|\mathcal{D}^k_{\text{train}}|} \sum_{(x_i,y_i)\in\mathcal{D}^k_{\text{train}}} \ell(f_j(x_i;\phi), y_i),$$

where ϕ is initialized by $\phi \leftarrow \phi_j^{(e)}$.

- **Structured Pruning:** Prune the model using an alignment-aware regularizer:

$$\min_{\substack{\tilde{\phi}^{(e)}_{j\to k},\\ \{\boldsymbol{\alpha}_l\}_{(l,l')\in\mathbb{L}_{\text{shared}}(k,j)},\\ \{\boldsymbol{\alpha}_u\}_{u\in\mathbb{L}_{\text{unshared}}(k,j)}}} \underbrace{\frac{1}{|\mathcal{D}^k_{\text{train}}|} \sum_{(x_i,y_i)\in\mathcal{D}^k_{\text{train}}} \ell(f_j(x_i;\tilde{\phi}^{(e)}_{j\to k}, \{\boldsymbol{\alpha}_l\}, \{\boldsymbol{\alpha}_u\}), y_i)}_{\text{1. Task Loss on Local Data}}$$

$$+ \lambda_p \underbrace{\sum_{(l,l')\in\mathbb{L}_{\text{shared}}(k,j)} \sum_{i=1}^{m_l} \left\| \alpha_{l,i} \mathbf{W}_{l,i}^{(j\to k)} - \mathbf{W}_{l',i}^{k} \right\|_2^2}_{\text{2. Alignment Regularization (Shared Layers Only)}}$$

$$+ \gamma_{\text{shared}} \underbrace{\sum_{l\in\mathbb{L}_{\text{shared}}(k,j)} \left\| \boldsymbol{\alpha}_l \odot \mathbf{W}_l^{(j\to k)} \right\|_{2,1}}_{\text{3. Group-Lasso for Shared Layers}}$$

$$+ \gamma_{\text{unshared}} \underbrace{\sum_{u\in\mathbb{L}_{\text{unshared}}(k,j)} \left\| \boldsymbol{\alpha}_u \odot \mathbf{W}_u^{(j\to k)} \right\|_{2,1}}_{\text{4. Group-Lasso for Unshared Layers}},$$

where α is a gating vector, λ_p and γ are hyperparameters controlling the strength of the alignment regularization and the structured pruning, respectively. $\|\cdot\|_{2,1}$ represents the group-lasso norm. $\odot$ denotes element-wise multiplication.

Model Alignment Process (2)

- **Post-Finetuning:** Refine the pruned model to recover any lost accuracy.

$$\phi_{j \to k}^{(e)} \leftarrow \arg\min_{\phi} \frac{1}{|\mathcal{D}_{\text{train}}^k|} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{train}}^k} \ell(f_j(x_i; \phi), y_i),$$

where ϕ is initialized by $\tilde{\phi}_{j \to k}^{(e)}$.

- **Validation Scoring:** Evaluate each post-tuned neighbor model $\phi_{j \to k}^{(e)}$ and the client's own model $\theta_k^{(e)-}$ on a local validation set $\mathcal{D}_{\text{val}}^k$.

$$\text{score}_k^{(e)}(\theta) = \frac{1}{|\mathcal{D}_{\text{val}}^k|} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{val}}^k} \mathbb{1}(\arg\max f(x_i; \theta) = y_i),$$

where $\mathbb{1}(\cdot)$ is the indicator function.

- **Top-K Selection Prep:** Rank all candidates by validation score (break ties using cosine similarity) for the next ensemble stage.

$$\begin{aligned} \{s_i^{(e)}\}_{i=1}^K \; = \; & \text{TopK}(CB, \{\phi_{j \to k}^{(e)}\}_{j=1}^Q \cup \{\theta_k^{(e)-}\}, \\ & \{\text{score}_k^{(e)}(\phi_{j \to k}^{(e)})\}_{j=1}^Q \cup \{\text{score}_k^{(e)}(\theta_k^{(e)-})\}, K), \end{aligned}$$

Where *CB* specifies that in the event of tied scores, models are further ranked using their cosine similarity to the local model.

Top-K Ensemble → Knowledge Distillation

Top-K Ensemble (Teacher Construction)

- Build a weighted ensemble of the top-K aligned models:

$$A_{\mathbf{w}_k^{(e)}}(x; \{s_i^{(e)}\}_{i=1}^K) = \sum_{i=1}^{K} w_i^{(e)} \cdot f_i(x; s_i^{(e)}),$$

where the optimal weights $\mathbf{w}_k^{(e)}$ is computed by minimizing the following KL-based distillation loss:

$$\mathbf{w}_k^{(e)} = \arg\min_{\mathbf{w}_k^0} \frac{1}{|\mathcal{D}_{\text{train}}^k|} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{train}}^k} \ell(A_{\mathbf{w}_k^0}(x_i; \{s_i^{(e)}\}_{i=1}^K), y_i).$$

Regularization-based Knowledge Distillation

- Distill the weighted ensemble $A_w^{(e)}$ into the client's student model $\theta_k^{(e)+}$ by minimizing following KL-based distillation loss:

$$\mathcal{L}_{\text{KD}}(\theta_k^{(e)+}) = \frac{1}{|\mathcal{D}_{\text{train}}^k|} \sum_{x_i \in \mathcal{D}_{\text{train}}^k} \text{KL}\Big(\text{softmax}\Big(\frac{A_{\mathbf{w}_k^{(e)}}(x_i)}{T}\Big) \parallel \text{softmax}\Big(\frac{f_k(x_i; \theta_k^{(e)+})}{T}\Big)\Big) + \lambda \|\theta_k^{(e)+} - \theta_k^{(e)-}\|^2,$$

where *T > 0* controls the smoothness of the softmax distributions applied to the logits.

Theoretical Analysis

Theorem 1. Risk Discrepancy Bound.

- Let $\theta_k^{(e)}$ be the student model obtained by minimizing the distillation loss $\mathcal{L}_{\mathrm{KD}}(\theta_k^{(e)})$ on $\mathrm{D}^{k}_{\mathrm{train}}$. Then, for a *C*-class problem with L-Lipschitz cross-entropy loss, *T>0*, and softmax outputs in *(α,1−α)*, the empirical risk discrepancy between the student and teacher models is bounded as follows:

$$|\,R_{\mathrm{S}}(\theta_k^{(e)}) - R(A_{\mathbf{w}_k^{(e)}})| \;\leq\; \frac{L\cdot CT}{\alpha(1-\alpha)} \cdot \left(\frac{\mathcal{L}_{\mathrm{KD}}(\theta_k^{(e)})}{2} + \frac{1}{8} \right).$$

Theorem 2. Convergence of Knowledge Distillation.

- Suppose $\{\theta_k^r\}_{r=0}^{R}$ are generated by $\theta_k^{r+1} = \theta_k^r - \eta \nabla \mathcal{L}_{\mathrm{KD},k}(\theta_k^r, \xi_k^r)$, under standard assumptions that the distillation loss $L_{KD,k}$ is *Ls*-smooth and *μ*-strongly convex, and that the variance of the stochastic gradient is bounded by σ^2,then for $r \geq 0$, and any step size *0<η<1/Ls* , the following bound holds:

$$\mathbb{E}[\|\theta_k^r - \theta_k^*\|^2] \;\leq\; \gamma^r \,\|\theta_k^0 - \theta_k^*\|^2 + \sum_{\tau=0}^{r-1} \gamma^\tau \,\beta,$$

Where $\gamma = \left(1 - 2\eta\mu + \frac{L_s^3}{\mu}\eta^2\right),\ \beta = \eta^2\sigma^2$, and θ_k^* is the minimizer of $L_{KD,k}$.

Experimental Setup

- **Datasets**
  - **Satellite imagery:** Wildfire, Hurricane
  - **Vision benchmarks**: SVHN, CIFAR-10, CIFAR-100
- **Setup**
  - **Federated setting:** 70 clients, non-IID partitioning (ψ = 0.1, 0.3, 0.5, 0.7)
  - **One-shot:** Each client communicates with neighbors only once
- **Baselines**
  - Local-only, FedAvg (global, multi-round), DENSE, Co-Boosting (one-shot global), FOL-A (FOL without distillation), FOL (with distillation)
- **Metrics**
  - Client-level accuracy

Quantitative Results

Table 1. Test accuracies (%) on Wildfire and Hurricane ($\psi = 0.7$), reported as mean ± std.

Dataset	Wildfire			Hurricane		
Satellite #	13	28	48	35	32	44
Methods	$\psi = 0.7$					
Local	94.23 ± 1.84	94.12 ± 1.80	90.53 ± 1.57	86.93 ± 1.56	87.34 ± 1.60	89.82 ± 1.82
FOL-A (E=1)	97.19 ± 1.53	97.16 ± 1.24	95.97 ± 1.55	95.34 ± 1.42	96.18 ± 1.02	97.61 ± 1.68
FOL-A (E=2)	97.50 ± 1.12	97.52 ± 1.17	97.33 ± 1.23	96.59 ± 1.76	96.97 ± 1.41	97.87 ± 1.22
FOL-A (E=3)	**97.53 ± 0.76**	**97.70 ± 0.98**	**97.99 ± 0.93**	**96.90 ± 1.09**	**97.47 ± 1.11**	**98.20 ± 1.03**
FOL (E=1)	94.94 ± 1.38	95.21 ± 1.32	91.26 ± 1.62	90.09 ± 1.55	89.87 ± 0.69	91.62 ± 0.58
FOL (E=2)	95.23 ± 1.35	95.57 ± 0.72	91.60 ± 1.29	91.23 ± 1.57	91.77 ± 0.83	95.21 ± 1.49
FOL (E=3)	96.32 ± 0.96	95.75 ± 1.39	91.95 ± 1.31	92.26 ± 1.05	92.41 ± 1.68	95.81 ± 1.88
FOL-AN (E=1)	94.38 ± 1.86	94.86 ± 1.67	91.28 ± 1.82	88.24 ± 1.82	91.14 ± 1.13	92.81 ± 1.10
FOL-AN (E=2)	95.63 ± 1.40	95.04 ± 1.43	93.29 ± 1.51	90.09 ± 0.64	92.47 ± 1.86	94.01 ± 1.70
FOL-AN (E=3)	95.94 ± 0.71	96.45 ± 0.65	95.97 ± 1.43	93.19 ± 1.23	93.04 ± 1.19	96.41 ± 1.26
FOL-N (E=1)	93.44 ± 1.68	94.68 ± 1.79	88.59 ± 2.31	85.76 ± 1.85	89.22 ± 0.93	91.62 ± 1.19
FOL-N (E=2)	94.69 ± 0.53	94.86 ± 0.88	90.60 ± 1.01	89.16 ± 1.31	90.21 ± 1.28	92.22 ± 1.65
FOL-N (E=3)	95.31 ± 1.49	95.21 ± 0.98	91.95 ± 0.97	90.71 ± 0.59	90.51 ± 1.21	94.61 ± 0.73
DENSE	88.75 ± 1.91	87.41 ± 1.63	83.22 ± 1.57	67.49 ± 1.81	69.95 ± 1.70	73.05 ± 1.62
Co-Boosting	90.31 ± 1.26	89.19 ± 1.13	88.02 ± 1.25	72.14 ± 1.52	74.45 ± 1.72	74.04 ± 1.54
FedAvg (E=1)	73.19 ± 1.73	73.94 ± 1.96	68.18 ± 2.02	60.21 ± 1.73	62.03 ± 1.95	66.26 ± 1.62
FedAvg (E=2)	73.13 ± 1.91	72.29 ± 1.74	66.92 ± 1.55	59.44 ± 1.64	64.33 ± 1.33	69.88 ± 1.57
FedAvg (E=3)	74.61 ± 1.54	71.58 ± 1.16	68.48 ± 1.23	63.70 ± 0.71	65.16 ± 1.14	67.82 ± 0.92

Table 2. Test accuracies (%) on Wildfire and Hurricane ($\psi \in \{0.5, 0.3, 0.1\}$), reported as mean ± std.

Dataset	Wildfire			Hurricane		
Satellite #	32	43	48	8	26	44
Methods	$\psi = 0.5$	$\psi = 0.3$	$\psi = 0.1$	$\psi = 0.5$	$\psi = 0.3$	$\psi = 0.1$
Local	79.07 ± 1.71	90.37 ± 1.76	85.50 ± 2.16	86.77 ± 1.90	57.14 ± 2.87	77.78 ± 1.35
FOL-A (E=1)	95.35 ± 1.42	94.07 ± 1.89	90.63 ± 1.92	95.04 ± 1.70	90.48 ± 1.57	88.89 ± 1.92
FOL-A (E=2)	96.52 ± 1.02	94.92 ± 1.25	96.14 ± 1.16	95.34 ± 1.16	91.72 ± 1.26	91.67 ± 1.26
FOL-A (E=3)	**97.67 ± 0.71**	**95.76 ± 0.85**	**96.88 ± 1.01**	**95.87 ± 1.03**	**93.65 ± 1.14**	**94.44 ± 0.87**
FOL (E=1)	90.70 ± 1.75	90.68 ± 1.01	88.46 ± 1.99	89.26 ± 1.25	84.13 ± 1.57	83.33 ± 1.69
FOL (E=2)	91.96 ± 1.09	91.53 ± 1.78	90.63 ± 1.77	90.08 ± 1.74	85.71 ± 1.38	84.43 ± 1.92
FOL (E=3)	93.02 ± 1.22	92.37 ± 1.27	93.75 ± 1.40	90.91 ± 1.38	87.30 ± 1.07	86.11 ± 1.18
FOL-AN (E=1)	90.77 ± 1.38	91.53 ± 1.26	87.51 ± 2.32	91.34 ± 1.70	87.47 ± 2.55	86.73 ± 1.94
FOL-AN (E=2)	93.22 ± 1.85	93.22 ± 1.17	90.63 ± 1.69	92.56 ± 1.18	88.89 ± 1.91	88.67 ± 1.75
FOL-AN (E=3)	95.35 ± 1.25	94.07 ± 1.21	90.94 ± 1.14	93.39 ± 1.37	90.48 ± 1.55	91.39 ± 1.26
FOL-N (E=1)	86.05 ± 1.96	88.14 ± 1.67	85.13 ± 1.92	85.95 ± 1.95	76.19 ± 1.73	80.56 ± 2.11
FOL-N (E=2)	87.35 ± 1.41	89.83 ± 1.76	86.38 ± 2.07	86.74 ± 1.83	80.95 ± 1.94	81.94 ± 1.38
FOL-N (E=3)	90.54 ± 1.51	90.06 ± 1.59	88.47 ± 1.37	87.60 ± 1.49	82.54 ± 1.76	83.37 ± 1.56
DENSE	79.91 ± 1.73	78.63 ± 1.98	52.08 ± 2.03	61.10 ± 1.51	58.73 ± 1.43	46.14 ± 1.81
Co-Boosting	86.05 ± 1.68	85.59 ± 1.65	54.51 ± 1.85	72.29 ± 1.68	52.38 ± 1.85	48.78 ± 1.50
FedAvg (E=1)	53.11 ± 1.82	63.25 ± 1.87	35.33 ± 2.76	66.12 ± 1.50	41.27 ± 1.99	46.14 ± 1.72
FedAvg (E=2)	56.03 ± 2.53	67.52 ± 1.92	45.16 ± 1.97	58.79 ± 1.86	45.16 ± 1.26	42.61 ± 1.86
FedAvg (E=3)	51.07 ± 1.93	66.10 ± 2.05	42.86 ± 1.53	60.33 ± 1.24	44.44 ± 1.76	43.33 ± 1.46

Table 3. Test accuracies (%) on CIFAR-10, CIFAR-100, and SVHN, reported as mean ± std.

Dataset	CIFAR-10	CIFAR-100	SVHN
Satellite #	9	14	21
Methods	$\psi = 0.7$	$\psi = 0.7$	$\psi = 0.5$
Local	60.06 ± 1.97	30.41 ± 2.28	78.97 ± 1.75
FOL-A (E=1)	70.73 ± 2.09	46.32 ± 2.12	85.73 ± 1.63
FOL-A (E=2)	70.93 ± 1.16	47.72 ± 1.53	86.26 ± 1.28
FOL-A (E=3)	**71.02 ± 0.73**	**49.24 ± 1.12**	**88.37 ± 0.92**
FOL (E=1)	65.68 ± 2.14	37.31 ± 2.43	81.09 ± 1.54
FOL (E=2)	66.06 ± 1.05	37.43 ± 1.58	81.62 ± 1.36
FOL (E=3)	66.83 ± 0.82	39.42 ± 1.03	82.85 ± 1.18
FOL-AN (E=1)	61.77 ± 2.36	31.23 ± 2.31	79.62 ± 1.64
FOL-AN (E=2)	62.35 ± 1.31	31.58 ± 1.75	80.92 ± 1.41
FOL-AN (E=3)	63.01 ± 0.71	32.05 ± 1.68	83.15 ± 1.33
FOL-N (E=1)	59.68 ± 2.43	30.64 ± 2.57	80.04 ± 1.74
FOL-N (E=2)	60.44 ± 1.37	30.99 ± 1.74	80.39 ± 1.39
FOL-N (E=3)	61.49 ± 1.43	31.11 ± 2.12	81.15 ± 1.22
DENSE	61.68 ± 2.03	29.59 ± 2.53	69.53 ± 1.57
Co-Boosting	63.11 ± 1.98	33.45 ± 2.12	73.58 ± 1.48
FedAvg (E=1)	47.76 ± 2.63	12.16 ± 3.93	53.08 ± 2.32
FedAvg (E=2)	44.90 ± 2.45	12.75 ± 3.23	58.72 ± 1.79
FedAvg (E=3)	45.57 ± 1.79	12.40 ± 2.89	55.49 ± 1.93

Table 4. Test accuracies (%) for six additional clients on the *Hurricane* dataset with $\psi = 0.7$.

Dataset	**Hurricane**					
Satellite #	41	3	9	22	56	51
Methods	$\psi = 0.7$					
Local	90.45	82.35	88.63	90.67	86.01	91.18
FOL-A (E=1)	94.27	91.18	92.73	93.10	93.87	96.57
FOL-A (E=2)	95.54	94.12	93.64	93.68	95.16	97.06
FOL-A (E=3)	96.18	96.06	94.09	95.40	95.74	97.55
FOL (E=1)	93.11	85.29	90.02	91.95	89.81	93.63
FOL (E=2)	93.63	91.33	91.82	92.53	90.07	94.12
FOL (E=3)	94.27	93.04	92.27	94.25	91.92	95.59
DENSE	70.02	67.35	68.13	71.31	69.57	70.16
Co-Boosting	74.61	69.16	72.51	73.63	75.21	74.47

Summary

- We propose **FOL**, a one-shot personalized federated learning framework tailored for constrained communication environments.
- FOL integrates **model alignment**, **top-K ensemble**, and **regularization-based distillation** to deliver strong personalization in just one exchange.
- We provide **theoretical guarantees** on risk bounds and convergence.
- Experiments across **five diverse datasets and 70 clients** show that FOL outperforms baselines, especially under high data heterogeneity.
- **Future work**: developing more advanced one-shot PFL techniques and integrate stronger privacy guarantees.

Thank you.

Guan Huang
PhD Student, Auburn University
gzh0040@auburn.edu

Tao Shu
Associate Professor, Auburn University
tshu@auburn.edu