2169: Zero-shot Federated Unlearning via Transforming from Data-Dependent to Personalized Model-Centric

Wenhan Wu[1], Huanghuang Liang[1], Yingling Yuan[2], Jiawei Jiang[1], Kanye Ye Wang[3], Chuang Hu[1,*], Xiaobo Zhou[3] and Dazhao Cheng[1,*] (* Corresponding Authors)

1 School of School Science, Wuhan University, Wuhan, China

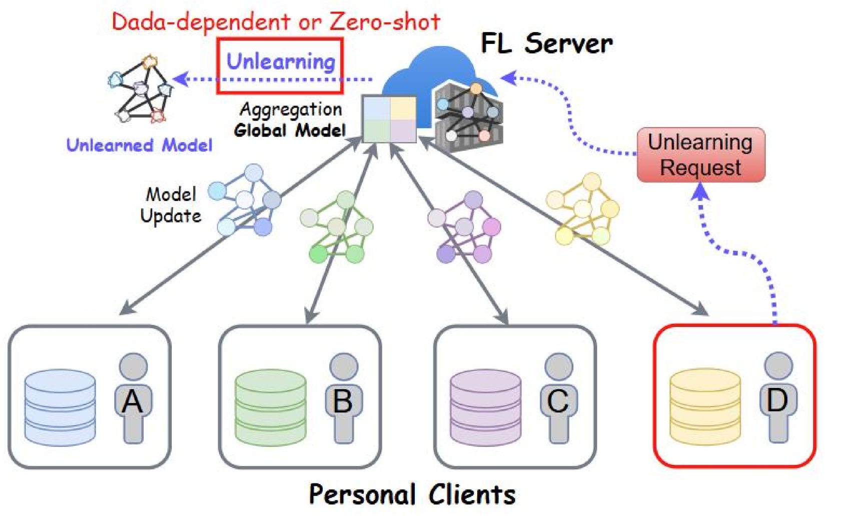
2 Hubei Key Laboratory of Transportation Internet of Things, School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China

3 State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, Macau SAR

Introduction and Motivation

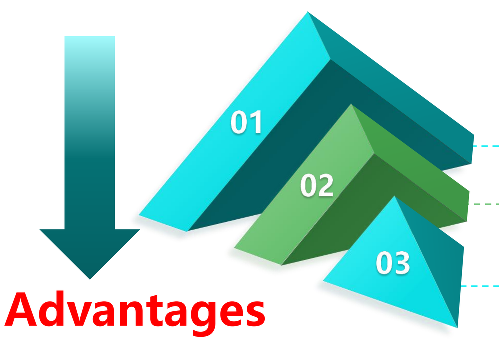
Data-dependent vs Zero-shot Federated Unlearning

Federated Unlearning (FU) addresses the "right to be forgotten" in federated learning by removing specific client data's contribution without retraining from scratch. Existing FUs are *data-dependent*, which assumes that systems can access original training data or stored historical parameter updates during unlearning. However, the assumption cannot always hold in practice, as users usually request the deletion of client data and historical parameter updates due to privacy concerns or storage limitations. Therefore, it is crucial to develop a *zero-shot FU* method without such data access.



Potential Idea: Transforming to Personalized Model-Centric

The main challenge of zero-shot FU is distinguishing and removing the target client's impact without client-specific data. Traditional methods rely on client data distributions or gradient differences, which are unavailable in zero-shot settings. A promising approach is to embed client-specific features directly into the model during training, enabling model-based unlearning. This shifts FU from a data-dependent to a model-centric process, allowing the model's personalized features to identify and erase target client contributions.



- **Differentiation and unlearning through model personalization, enabling zero-shot capability.**
- **Maps samples with the same label to distinct decision spaces via personalization, effectively avoiding catastrophic unlearning.**
- **Enhances learning performance under non-IID data through model personalization.**

Overall Architecture of ZeroFU

We propose the first zero-shot FU framework, ZeroFU. The overall process involves first collaboratively training the FL model, and then during unlearning, the forgotten client C_f sends an unlearning request to the retained clients C_r, followed by executing unlearning to remove its specific contributions.

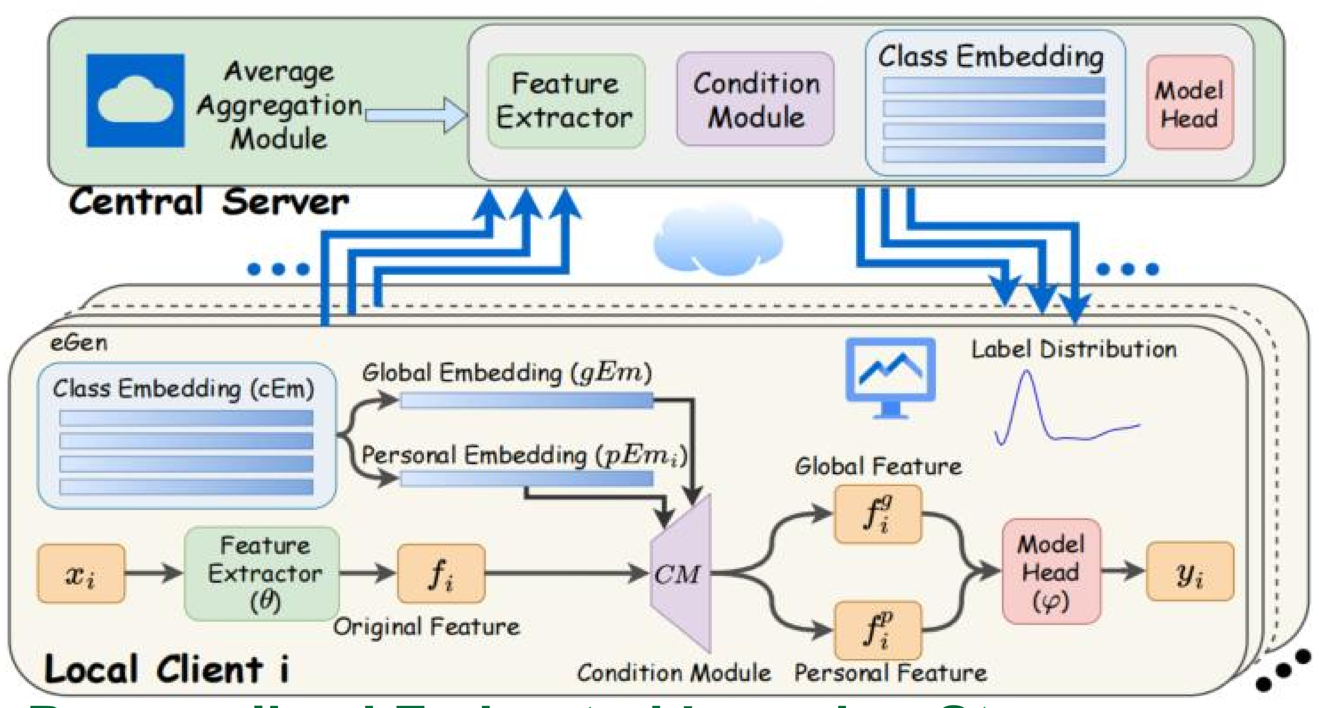
Model Personalization

The backbone is divided into θ and φ. gEm and pEm_i are computed based on the class embedding cEm from eEm and label distribution LD_i of client C_i and then fed into the CM. CM processes the extracting features f_i from θ and generating both global features f_i^g and client-specific features f_i^p, which are then combined to produce the final classification result. θ, φ, CM and eEm share parameters among clients and are trained in an end-to-end manner.

Zero-shot Federated Unlearning

A generator G maximizes the difference $FLoss$ between f_s^p (from student S) and f_f^p (from forgotten client C_f). Knowledge distillation (KD) minimizes the similarity measures $KLoss$ (output similarity) and $ALoss$ (intermediate layer similarity) between the teacher (retained client) and student. $FLoss$ is also minimized to effectively mask the personalized information of the forgotten client.

Personalized Learning Framework with Condition Module



Objective: Training personalized U – *class* classification models for each client by embedding client-specific features.

Personalized Federated Learning Steps:

1. Feature Extractor: θ extracts features from input x_i: $f_i = \theta(x_i; \omega_\theta)$.

2. Class Embedding Generation:

Global Embedding: $gEm = \frac{1}{U}\sum_{y=0}^{U-1} cEm_y$.

Personal Embedding: $pEm_i = \sum_{u=0}^{U-1} cEm_u * LD_i$, where $LD_i = \mathbb{E}_{D_i}\mathbb{I}(y_i, y)$.

3. Feature Transformation with Conditional Module:

Generates global (f_i^g) and personalized (f_i^p) features:

$$f_i^g = ReLU(CM_B(gEm;\ \omega^{CM}) + (CM_W(gEm;\ \omega^{CM}) + \mathbf{1}) \odot f_i),$$
$$f_i^p = ReLU(CM_B(pEm_i;\ \omega^{CM}) + (CM_W(pEm_i;\ \omega^{CM}) + \mathbf{1}) \odot f_i).$$

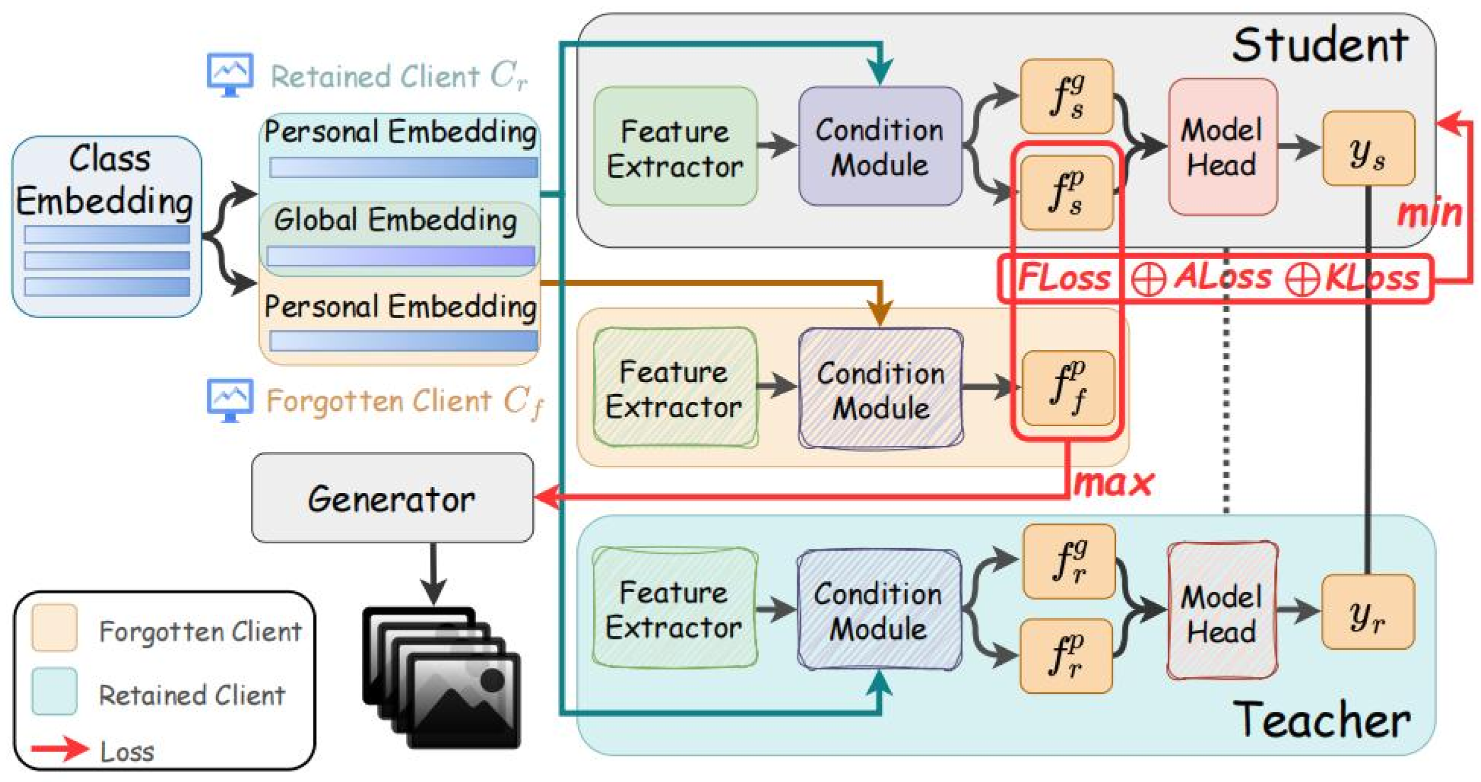
4. Model Head Output: $\hat{y}_i = \varphi([f_i^p;\ f_i^g];\ \omega^{\varphi})$.

5. Local Loss SGD: $\mathcal{L}_i(\omega_i) = \mathcal{L}_i^{CE} + \mathcal{L}_i^{EM} + \lambda_1|\omega^{cGen}|_2^2 + \lambda_2|\omega^{CM}|_2^2$.

- Cross-entropy loss: $\mathcal{L}_i^{CE} = CrossEntropyLoss(\hat{y}_i, y_i)$.
- Class embedding guidance loss: $\mathcal{L}_i^{EM} = -\log\left(\frac{\exp(\cos\_sim(f_i^g, cEm_{y_i}))}{\sum_{u=1}^{U}\exp(\cos\_sim(f_i^g, cEm_u))}\right)$, encouraging feature vectors to align with their own class embeddings while staying distant from others, preserving class specificity.

6. Aggregation: Client C_i shares updated parameters ω_i^t={θ, φ, $eGen$, CM} with the server and aggregate global parameters: $\omega^t \leftarrow \frac{1}{\sum_{i\in O^t}|D_i|}\sum_{i\in O^t}|D_i|\,\omega_i^t$.

Zero-shot Unlearning Framework with Knowledge Distillation



Objective: Removing C_f's influence on the C_r's under a zero-shot setting by erasing personalized feature information.

Zero-shot Client-level Unlearning Steps:

1. Data Generation: Use a generator $G(z;\ \omega_G)$ to synthesize pseudo-samples x from random noise $z \sim \mathcal{N}(0,1)$. x feed into teacher $R(x, \omega_r)$ on C_r, student S, and forgetting model $F(x, \omega_f)$ on C_f.

2. Parameter Freezing: $R(x, \omega_r)$ and $F(x, \omega_f)$ guide $S(x, \omega_s)$, which shares the freezing part of teacher's structure: $\omega_s^{CM} = \omega_r^{CM}$, $\omega_s^{cGEn} = \omega_r^{cGen}$.

3. Zero-shot Knowledge Distillation Loss:

- Forgetting Personalized Features: Minimize feature cosine similarity between S and F: $FLoss = 1 - \cos\_sim(f_s^p, f_f^p)$, f_s^p and f_f^p are generated from personalized embeddings $[gEm, pEm_r]$ and $[gEm, pEm_f]$, respectively.
- Maintaining Knowledge on C_r: Align outputs of R and S using Kullback-Leibler divergence: $KLoss = \tau^2\sum_{i=0}^{U-1}\sigma\left(\frac{\hat{y}_r}{\tau}\right)_i \log\frac{\left(\frac{\hat{y}_r}{\tau}\right)_i}{\log\left(\frac{\hat{y}_r}{\tau}\right)_i}$.
- Preserve attention patterns between R and S with attention loss:

$$ALoss = \frac{1}{|\mathcal{N}_L|}\sum_{l\in\mathcal{N}_L}\left\|\frac{f(A_l^{(r)})}{\|f(A_l^{(r)})\|_2} - \frac{f(A_l^{(s)})}{\|f(A_l^{(s)})\|_2}\right\|_2.$$

4. Student and Generator Co-Optimization and Training Loop:

- Each KD round, S minimizes total loss using SGD with T_k distillation steps: $\min_{\omega_s^\theta, \omega_s^\varphi} F_{stu}$, where $F_{stu} = FLoss + \beta\,KLoss + \gamma\,ALoss$.
- Each unlearning round, G maximizes $FLoss$ adversarially to strengthen the forgetting signal: $\min_{\omega^G} FLoss$.

Evaluation Results

Performance on Retained and Forgotten Data

DataSet	ζ	C_r	C_f	Origin D_r	Origin D_f	Retrained D_r	Retrained D_f	ZeroFU D_r	ZeroFU D_f	FedMM D_r	FedMM D_f	FedGKT D_r	FedGKT D_f	FedBadT D_r	FedBadT D_f
MNIST	0.01	0	1	96.40	99.22	96.30	0.24	92.63	0.00	73.11	0.00	71.14	0.00	88.75	0.00
		8	9	98.54	99.91	98.81	0.01	97.05	0.01	89.11	0.00	89.11	0.00	89.11	1.13
	0.1	4	5	99.10	99.26	99.03	57.18	96.03	38.95	20.98	0.00	60.70	0.12	82.66	27.53
		3	6	96.02	97.29	94.41	83.79	91.04	83.54	44.89	65.19	54.60	0.00	85.00	73.23
SVHN	0.01	0	1	96.02	95.13	96.02	0.00	96.02	0.00	96.02	0.00	96.02	0.00	96.02	4.96
		8	9	95.99	99.99	95.84	0.00	95.99	0.00	95.99	0.00	93.44	0.00	95.99	8.95
	0.1	4	5	98.13	70.87	98.13	60.26	98.13	46.60	98.13	36.98	98.13	39.98	98.13	38.10
		3	6	88.23	59.93	86.08	54.60	88.26	45.41	88.16	0.00	88.16	0.00	88.16	8.40
FMNIST	0.01	0	1	99.45	99.06	99.45	6.27	99.45	8.69	100.00	12.70	71.95	0.00	100.00	96.28
		8	9	99.98	99.44	99.50	12.96	99.98	12.13	99.99	0.00	99.98	0.00	99.98	84.92
	0.1	4	5	83.12	85.91	88.13	13.35	79.78	16.11	63.12	16.11	59.41	16.11	59.40	83.48
		3	6	76.74	98.11	88.41	13.88	84.81	17.31	49.61	0.00	21.04	0.00	73.84	95.41
CIFAR10	0.01	0	1	98.78	100.00	98.78	0.00	98.78	0.00	98.78	0.00	98.78	0.00	98.78	100.00
		8	9	80.36	99.96	77.88	0.00	78.76	0.00	21.28	0.00	75.68	0.00	80.13	99.96
	0.1	4	5	86.08	80.27	79.77	28.22	78.10	29.01	71.01	7.07	48.70	2.30	71.11	12.11
		3	6	87.19	91.70	80.15	3.76	76.67	2.01	47.48	0.00	41.20	9.58	65.30	92.21

- **Compared with zero-shot machine unlearning methods applied to federated scenarios.**

Dataset	Model	FedEraser D_r	FedEraser D_f	FedRecovery D_r	FedRecovery D_f	Knot D_r	Knot D_f	ZeroFU D_r	ZeroFU D_f
MNIST	retrain	99.01	98.44	99.01	98.44	99.15	98.48	97.78	96.53
	unlearn	96.23	95.84	92.08	91.89	97.66	97.12	94.05	94.77
SVHN	retrain	94.07	89.07	94.07	89.07	97.89	89.36	93.81	89.01
	unlearn	87.55	87.50	81.24	78.90	94.97	93.18	90.45	89.03
FMNIST	retrain	93.79	91.09	93.79	91.09	93.97	90.99	92.24	92.75
	unlearn	90.36	90.09	86.42	83.78	86.27	85.90	91.46	90.88
CIFAR10	retrain	88.73	68.23	88.73	68.23	90.65	87.75	86.70	85.38
	unlearn	85.08	63.57	75.34	63.21	85.33	85.28	85.43	83.78

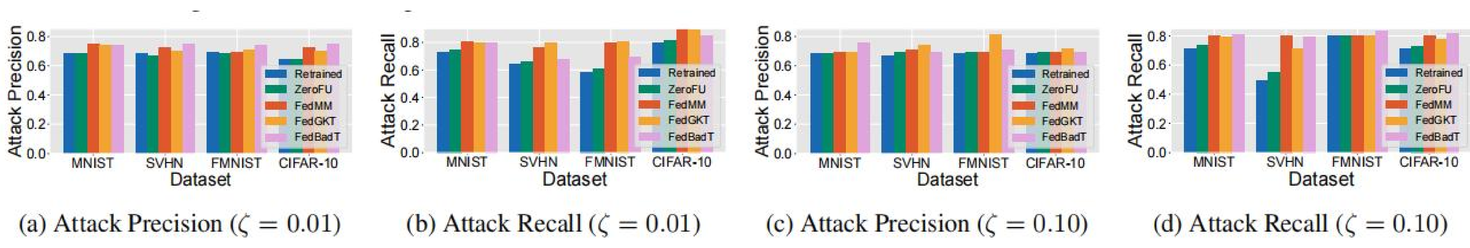
(a) IID Scenario

Dataset	Model	FedEraser D_r	FedEraser D_f	FedRecovery D_r	FedRecovery D_f	Knot D_r	Knot D_f	ZeroFU D_r	ZeroFU D_f
MNIST	retrain	98.72	87.79	98.72	87.79	97.37	88.72	94.41	83.79
	unlearn	77.16	65.89	90.78	82.36	94.34	85.46	91.04	83.54
SVHN	retrain	84.26	59.26	84.26	59.26	81.95	79.05	86.08	54.60
	unlearn	57.81	42.94	70.12	40.25	80.31	77.63	88.26	45.41
FMNIST	retrain	58.46	22.49	58.46	22.49	58.22	12.55	88.41	13.88
	unlearn	42.32	15.70	49.56	10.98	48.09	38.21	84.81	17.31
CIFAR10	retrain	52.79	39.58	52.79	39.58	71.31	33.89	79.77	28.22
	unlearn	44.10	34.54	40.78	32.08	53.41	33.91	78.10	29.01

(b) non-IID Scenario ($\zeta = 0.10$)

- **Compared with non-zero-shot federated unlearning methods.**

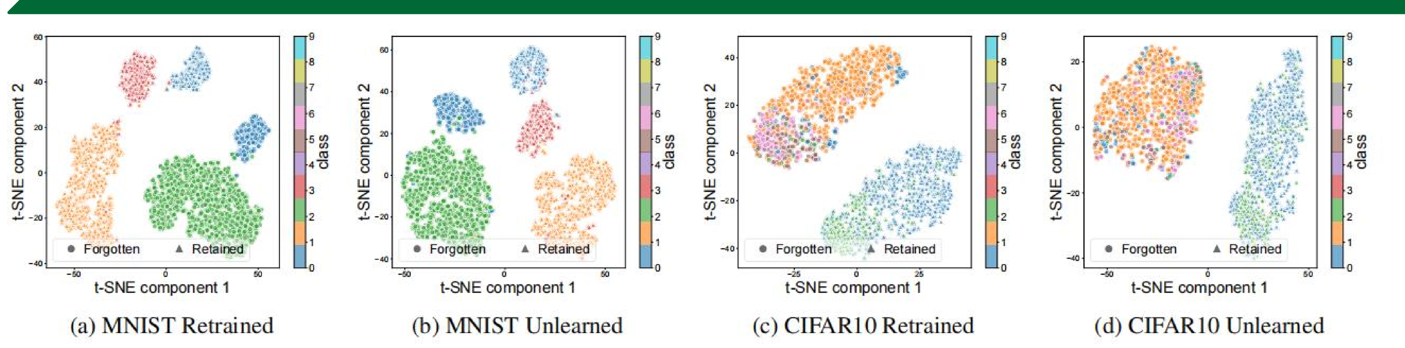
Privacy Leakage of Forgotten Data via MIA



(a) Attack Precision ($\zeta = 0.01$) (b) Attack Recall ($\zeta = 0.01$) (c) Attack Precision ($\zeta = 0.10$) (d) Attack Recall ($\zeta = 0.10$)

- **Membership Inference Attack on forgotten data D_f.**

Visualization of the Personalized Feature



(a) MNIST Retrained (b) MNIST Unlearned (c) CIFAR10 Retrained (d) CIFAR10 Unlearned

t-SNE shows catastrophic forgetting on data with the same label is avoided.

Contact Us Paper Link