FedCTTA: A Collaborative Approach to Continual Test-Time Adaptation in Federated Learning

Rakibul Hasan Rajib*, Md Akil Raihan Iftee*, Mir Sazzat Hossain*, A. K. M. Mahbubur Rahman*,
Sajib Mistry†, M Ashraful Amin* and Amin Ahsan Ali*
*Center for Computational & Data Sciences, Independent University, Bangladesh
†Curtin University
rakibul@iub.edu.bd, iftee1807002@gmail.com, {sazzat, akmmrahman}@iub.edu.bd,
Sajib.Mistry@curtin.edu.au, {aminmdashraful, aminali}@iub.edu.bd

***Abstract*—Federated Learning (FL) enables collaborative model training across distributed clients without sharing raw data, making it ideal for privacy-sensitive applications. However, FL models often suffer performance degradation due to distribution shifts between training and deployment. Test-Time Adaptation (TTA) offers a promising solution by allowing models to adapt using only test samples. However, existing TTA methods in FL face challenges such as computational overhead, privacy risks from feature sharing, and scalability concerns due to memory constraints. To address these limitations, we propose Federated Continual Test-Time Adaptation (FedCTTA), a privacy-preserving and computationally efficient framework for federated adaptation. Unlike prior methods that rely on sharing local feature statistics, FedCTTA avoids direct feature exchange by leveraging similarity-aware aggregation based on model output distributions over randomly generated noise samples. This approach ensures adaptive knowledge sharing while preserving data privacy. Furthermore, FedCTTA minimizes the entropy at each client for continual adaptation, enhancing the model's confidence in evolving target distributions. Our method eliminates the need for server-side training during adaptation and maintains a constant memory footprint, making it scalable even as the number of clients or training rounds increases. Extensive experiments show that FedCTTA surpasses existing methods across diverse temporal and spatial heterogeneity scenarios.**

***Index Terms*—Federated learning, Continual Test-Time Adaptation**

I. Introduction

Test-Time Adaptation (TTA) is revolutionizing deep learning by enabling models to adapt dynamically to unseen data distributions during deployment. Traditional machine learning models often degrade in performance when a distribution shift occurs between training and testing data. For instance, an autonomous medical imaging system trained on high-resolution hospital scans may struggle to interpret low-quality scans from rural clinics due to differences in imaging equipment. Existing solutions such as Domain Generalization (DG) [1]–[3] and Domain Adaptation (DA) [4], [5] attempt to mitigate this issue by either training on diverse domains or adapting from a source domain. However, DG requires sufficient domain diversity, and DA depends on access to source data, which may be impractical due to privacy constraints. TTA overcomes these constraints by allowing models to self-adapt using only incoming test samples, eliminating the need for retraining or access to original training data. Recent advancements in TTA [6]–[8] leverage techniques such as entropy minimization, self-supervised learning, and feature alignment, ensuring robust model performance in dynamic and privacy-sensitive environments.

Data privacy concerns, driven by regulations like GDPR [9], challenge traditional machine learning, which relies on centralized data processing. Federated Learning (FL) [10] addresses this by enabling collaborative model training across decentralized clients without sharing raw data, making it ideal for privacy-sensitive domains like healthcare and finance. However, *TTA in FL remains challenging due to heterogeneous and evolving data distributions across clients.* For instance, in federated healthcare, hospitals generate non-IID data due to varying patient demographics, equipment, and practices, causing models to struggle when deployed in settings with unseen data distributions.

Local adaptation, where each client fine-tunes the model using its own test data, fails to leverage broader shifts useful for generalization. *Collaborative adaptation*, where clients share insights without raw data, could improve performance but faces challenges: (1) **privacy risks** from feature or gradient sharing, (2) **model misalignment** due to distinct distribution shifts, and (3) **scalability** issues for resource-constrained clients. Our objective is to *develop privacy-preserving and efficient TTA framework* that enables decentralized models to adapt to evolving conditions while maintaining Federated Learning (FL)'s privacy guarantees. This ensures robust and scalable deployment of FL models in dynamic environments, allowing them to generalize effectively despite distribution shifts across decentralized clients.

Recent approaches have explored TTA in FL to enhance model robustness in decentralized environments. FedICON [11] employs contrastive learning to adapt models to diverse client environments, but its high computational demands make it impractical for resource-limited clients. ATP [7] introduces client-specific adaptation by adjusting module-specific adaptation rates. However, it assumes static test-time distributions and does not explicitly encourage inter-client knowledge sharing, which could enhance robustness by leveraging insights from clients with similar data distributions.

Other methods, such as FedTHE+ [12], improve personalization and adaptivity by ensembling a global generic classifier

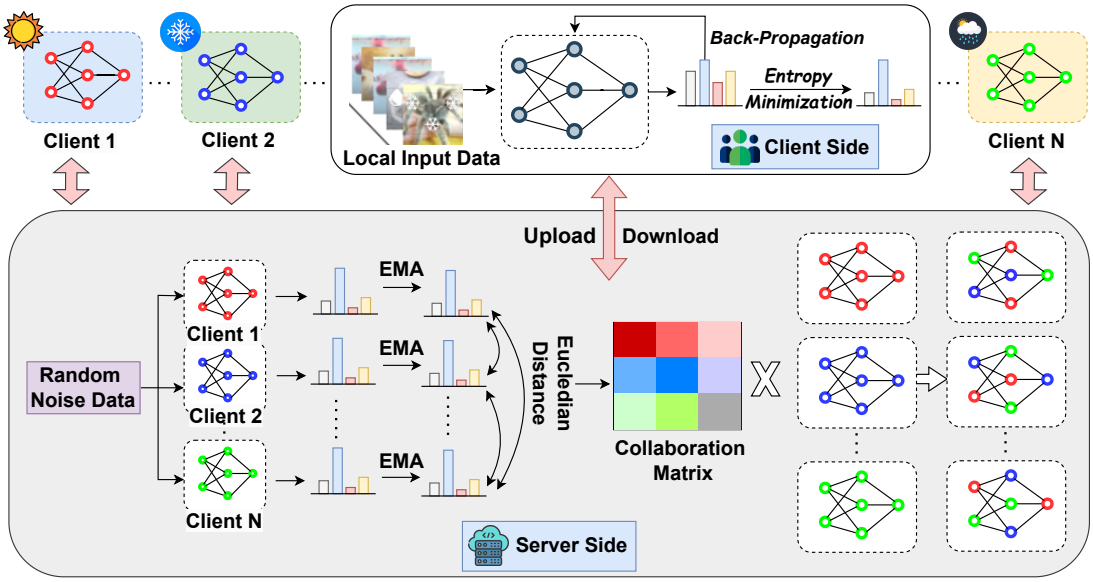


Fig. 1: Illustration of the FedCTTA framework. The server aggregates models received from all the clients based on euclidean distance of outputs probability distribution, then distributes the personalized aggregated models back to the clients for the next round. This process continues iteratively to adapt the models collaboratively across all clients. The figure demonstrates both client-to-server communication (model updates) and server-to-client communication (aggregated model distribution).

and a local personalized classifier in a two-head FL model. However, its performance declines when clients encounter vastly different out-of-distribution data, as combining ensemble classifiers into a more generic global classifier may lead to suboptimal generalization. More recently, FedTSA [13] introduced a collaboration mechanism using temporal-spatial correlations based on local feature means, allowing clients with similar data distributions to improve personalized model aggregation. However, this method introduces privacy concerns since sharing local feature statistics risks sensitive data leakage through reconstruction. Additionally, FedTSA requires server-side learning (6.2 millions parameters) during adaptation, increasing computational complexity. The reliance on storing local feature means in a memory bank also poses scalability challenges, as memory demands grow with the number of clients and training steps.

To address these limitations, we propose **Federated Continual Test-Time Adaptation (FedCTTA)**, a framework designed to enable efficient and privacy-preserving adaptation in federated settings. Unlike existing approaches that rely on sharing local feature means in the server side, FedCTTA avoids direct feature sharing, mitigating privacy risks while facilitating continual adaptation. FedCTTA operates by leveraging entropy minimization or updating batch normalization statistics at each client for local test-time adaptation, ensuring model confidence in evolving target distributions. Instead of relying on stored feature statistics, we incorporate similarity-aware aggregation through functional similarity [14], where the server estimates collaborative relationships between participating clients based on client model outputs over a set of randomly generated noise samples. This allows adaptive knowledge sharing while preserving data privacy. Our approach is computationally efficient, as it eliminates the need for additional training on the server side and reduces the memory footprint by avoiding persistent storage of client-specific embeddings. FedCTTA seamlessly integrates continual adaptation with federated learning, ensuring both domain-aware collaboration and robust model generalization without introducing excessive communication or storage overhead. FedCTTA outperforms existing methods under varying degrees of temporal and spatial heterogeneity. It achieves 66.50% and 63.39% accuracy under the NIID setting, and 67.78% and 64.52% under the IID setting on CIFAR10-C and CIFAR100-C dataset, respectively in TTA-bn method. In contrast, FedTSA achieves 66.19% and 62.93% (NIID), as well as 67.51% and 63.70% (IID). FedCTTA ensures higher accuracy while preserving privacy in decentralized settings.

Our key contributions are as follows:

- We introduce a similarity-aware aggregation technique in federated learning based on functional similarity, where the server calculates collaboration relationships between clients by comparing the outputs of their models over randomly generated noise samples.
- Unlike prior work [12], [13], FedCTTA does not store or share local feature embeddings, ensuring data security and mitigating privacy risks.
- Our method eliminates the need for server-side training during adaptation, significantly reducing computational complexity.
- By avoiding storage of feature means across federated rounds, FedCTTA maintains a constant memory footprint, making it scalable even as the number of clients or

training rounds increases.

II. Related Works

A. Federated Learning

Federated learning is a decentralized approach to training machine learning models while keeping data localized, thereby addressing privacy and security concerns. FedAvg [10] aggregates client models into a global one, while FedAvg+FT further fine-tunes it on local data for personalization. FedProx [15] introduces regularization to handle client data heterogeneity, and FedAVGM [16] incorporates momentum for better aggregation. Li et al. [15] proposed using a globally shared dataset to mitigate performance degradation in non-IID data settings, improving model accuracy by up to 30% on skewed datasets like CIFAR-10. Zhao et al. [17] introduced FedProx, an extension of FedAvg, to handle statistical and system heterogeneity. FedAMP [18] fosters collaboration between clients with similar data, whereas MOON [19] refines local training by leveraging model representation similarity. pFedSD [20] enables clients to distill knowledge from past personalized models, and pFedGraph [21] constructs a collaboration graph based on model similarities. LDAWA [22] improves aggregation by considering angular divergence, while FedTSA [13] utilizes temporal-spatial attention to capture both intra-client and inter-client correlations.

B. Test Time Adaptation

Test-Time Adaptation (TTA) methods enable models to adapt to distribution shifts without access to source data. TENT [23] minimizes entropy by updating BatchNorm parameters, achieving state-of-the-art results on corrupted datasets like ImageNet-C with efficient online updates. DUA [24] extends this by dynamically adjusting BatchNorm statistics using minimal unlabeled test data, improving real-time adaptation in scenarios like autonomous driving. EATA [25] mitigates catastrophic forgetting and noisy updates through entropy-based sample selection and a Fisher regularizer. CoTTA [6] enhances adaptation in non-stationary environments with weight-averaged pseudo-labeling and stochastic restoration of source weights to maintain long-term knowledge. These approaches demonstrate diverse strategies for improving TTA efficiency and robustness across various applications.

C. Federated Test Time Adaptation

Adaptive Test-Time Personalization (ATP) [7] learns module-specific adaptation rates based on client distribution shifts. Clients simulate unsupervised adaptation during training, refining rates to enhance performance on unseen, unlabeled data. FedTHE+ [12] ensembles global and local classifiers for robust test-time personalization and performs unsupervised fine-tuning, improving accuracy across in-domain (ID) and out-of-domain (OOD) distributions. FedICON [11] uses contrastive learning to capture invariant knowledge from inter-client heterogeneity during training and self-supervision for smooth test-time adaptation, tackling intra-client heterogeneity. While leveraging inter-client heterogeneity to address test-time shifts, FedICON requires extensive contrastive learning, which may be computationally intensive for resource-constrained clients. Xu et al. [26] proposed FedCal, a lightweight framework that performs test-time classifier calibration using estimated label priors from global model predictions. FedCal handles label shifts efficiently without extra labeled data, ensuring flexibility for unseen clients.

III. Methodology

A. Problem Definition

Continual test-time adaptation (TTA) addresses the challenge of adapting models to sequentially arriving, non-stationary target domains without access to source data. In federated settings, where clients observe distinct or overlapping domains that evolve over time, this becomes even more challenging.

We consider a federated system with N clients, $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$, each receiving a data stream $\mathcal{D}_t^{(i)}$ over time. The objective is to adapt the client models θ while preserving privacy and preventing catastrophic forgetting, ensuring robust performance despite evolving data distributions.

B. Local Test-time Adaptation: Entropy Minimization

In the client side, we explore two approaches for continual Test-Time Adaptation (TTA): (1) Fine-tuning all model parameters via entropy minimization, and (2) Updating only the batch normalization (BN) layer statistics (mean and variance).

In the first approach, entropy minimization is employed to align the feature distributions with evolving target domains. For an input $x \in \mathcal{D}_t^{(i)}$, the entropy $H(p)$ is defined as:

$$H(p) = -\sum_{k=1}^{K} p_k \log(p_k), \tag{1}$$

where $p = f_{\theta_i}(x)$ is the predicted probability vector, and K is the number of classes. The entropy minimization objective for client C_i is given by:

$$L_{\text{ent}} = \frac{1}{|\mathcal{D}_t^{(i)}|} \sum_{x \in \mathcal{D}_t^{(i)}} H(f_{\theta_i}(x)). \tag{2}$$

The model parameters θ are updated to minimize L_{ent} via gradient descent. Minimizing L_{ent} encourages confident predictions (low uncertainty), facilitating feature alignment with the target domain.

In the second approach, only the running mean μ, variance σ^2, and scale/shift parameters γ and β of the BN layers are updated for each incoming data stream:

$$\mu_i^{\text{new}} = (1-\alpha)\mu_i^{\text{old}} + \alpha \cdot \mathbb{E}_{x \sim \mathcal{D}_t^{(i)}}[x], \tag{3}$$

$$\sigma_i^{2,\text{new}} = (1-\alpha)\sigma_i^{2,\text{old}} + \alpha \cdot \text{Var}(x \sim \mathcal{D}_t^{(i)}), \tag{4}$$

where α is the momentum parameter, $\mathbb{E}[\cdot]$ represents the batch mean, and $\text{Var}[\cdot]$ denotes the batch variance.

By combining BN statistics updates and full parameter updates through entropy minimization, both methods enable

TABLE I: Performance comparison of various federated learning methods with our proposed FedCTTA on CIFAR10-C and CIFAR100-C datasets. We evaluate all methods under two TTA setups: TTA-Grad, where all model parameters are updated during adaptation, and TTA-BN, where only batch normalization layers are updated.

Method	NIID				IID			
	CIFAR10-C		CIFAR100-C		CIFAR10-C		CIFAR100-C	
	TTA-grad	TTA-bn	TTA-grad	TTA-bn	TTA-grad	TTA-bn	TTA-grad	TTA-bn
No-Adapt	58.47±0.19	58.61±0.17	30.22±0.12	30.22±0.12	58.64±0.22	58.55±0.21	30.22±0.12	30.22±0.12
Local	63.82±0.31	64.65±0.29	52.85±0.32	55.99±0.34	63.96±0.33	64.79±0.31	52.94±0.31	56.05±0.34
FedAvg	61.15±0.24	61.45±0.23	51.63±0.17	57.13±0.43	66.12±0.26	67.41±0.27	62.54±0.31	63.96±0.31
FedAvg+FT	63.82±0.27	61.45±0.23	47.83±0.58	57.13±0.43	63.79±0.30	67.41±0.27	61.72±0.59	63.96±0.31
FedProx	61.68±0.22	61.45±0.23	53.00±0.38	57.13±0.43	66.12±0.24	67.41±0.27	62.33±0.67	63.96±0.31
FedAvgM	61.50±0.25	61.37±0.19	52.31±0.46	57.13±0.43	63.60±0.28	67.41±0.27	54.66±0.27	63.96±0.31
MOON	61.58±0.23	61.45±0.23	54.26±0.27	57.13±0.43	66.05±0.25	67.41±0.27	62.40±0.23	63.96±0.31
pFedSD	61.31±0.21	61.45±0.23	53.33±0.37	57.13±0.43	66.14±0.26	67.41±0.27	62.32±0.33	63.96±0.31
pFedGraph	62.38±0.26	64.21±0.25	57.01±0.38	58.73±0.38	66.10±0.29	64.42±0.28	62.48±0.30	58.75±0.63
LDAWA	61.85±0.23	61.45±0.23	53.61±0.33	57.13±0.43	65.92±0.26	67.41±0.27	62.37±0.41	63.96±0.31
FedTSA	63.39±0.27	66.19±0.26	58.03±0.38	62.93±0.29	66.29±0.28	67.51±0.27	62.62±0.36	63.70±0.34
FedCTTA	**66.23±0.28**	**66.50±0.27**	**64.81±0.29**	**63.39±0.28**	**66.64±0.29**	**67.78±0.28**	**64.15±0.28**	**64.52±0.28**

domain-specific adaptation. This allows each client to locally adjust its model to the test domain, improving accuracy, mitigating catastrophic forgetting, and ensuring privacy in federated settings.

C. Similarity-Aware Aggregation

To facilitate efficient collaboration, model aggregation on the server side leverages the functional similarity [14] i.e., similarity of output behavior or probability distributions across clients. Clients exposed to similar domains contribute more to each other's updates while preserving data privacy. Since clients cannot share their data, a set of noise data points, $D_{\text{noise}} = \{z_i\}_{i=1}^{M}$, is randomly generated and used as a reference dataset. In the server, these random noise samples are passed to each client as input to compute similarities in output logits.The similarity between clients is computed based on several methods such as the cosine similarity, cross entropy, negative Euclidean distance between their logits for the random noise data points. Among them, we empirically found that euclidean distance performs the best in this framework, shown in Table IV.

For client i, the logits for a random noise sample z are represented as $f_{\theta_i}(z)$. The similarity between clients i and j is computed using the negative Euclidean distance between their mean logits:

$$D_{ij} = -\|\mu_i - \mu_j\|_2, \tag{5}$$

where μ_i and μ_j are the mean logits over the random noise dataset:

$$\mu_i = \frac{1}{M}\sum_{k=1}^{M} f_{\theta_i}(z_k), \quad \mu_j = \frac{1}{M}\sum_{k=1}^{M} f_{\theta_j}(z_k). \tag{6}$$

A higher (less negative) D_{ij} indicates greater similarity between clients. To derive collaboration weights, the pairwise distances D_{ij} are normalized using the softmax function C_{ij} and the server performs weighted aggregation based on C_{ij}, and each client updates its local model accordingly:

$$\theta_i^{\text{new}} = \sum_{j=1}^{K} \frac{\exp(D_{ij})}{\sum_{k=1}^{K}\exp(D_{ik})}\theta_j, \tag{7}$$

where C_{ij} represents the contribution of client j to client i's aggregation, and K is the total number of clients. θ_j denotes the parameters of client j's model. This approach fosters domain-aware collaboration by prioritizing updates from similar clients, improving adaptation and continuous learning in federated settings.

IV. Experimental Results

A. Implementation Details

1) Datasets and Setup: We evaluate our proposed aggregation method on two standard corruption benchmarks: CIFAR10-C and CIFAR100-C, where a model trained on CIFAR-10 and CIFAR-100, respectively, is adapted to their corrupted versions. These datasets are constructed by applying 15 distinct corruption types at five severity levels to the test and validation images of the original CIFAR datasets. Consistent with prior work, we report the average accuracy across all corruption types and clients at the highest severity level (severity 5). For CIFAR-100 to CIFAR100-C adaptation, we utilize a pretrained ResNeXt-29 model obtained from the Robustbench benchmark [27], while for CIFAR10-C, we employ a pretrained ResNet-8 [28]. To simulate a dynamically evolving test distribution, we progressively alter the corruption type at severity 5 over time. The test data is distributed among 20 clients to emulate a federated learning (FL) setting with decentralized data. Each client processes streaming test data in batches of 10, experiencing sequences of distribution shifts.

Test-Time Adaptation (TTA) methods aim to improve the robustness of a pretrained model when handling unlabeled test data. One approach, TTA-bn, adjusts batch normalization (BN) statistics to match the test distribution, as demonstrated by NORM [29] and DUA [24], without requiring gradient computation. In contrast, TTA-grad methods [6], [23] adapt

TABLE II: Detailed performance comparison under spatial IID and temporal heterogeneity using the TTA-bn method.

Datasets	Time / Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*Jpeg*	*Mean*
	Time	t →															
CIFAR10-C	Source	37.30	38.44	26.08	28.99	33.92	27.44	30.21	34.53	32.89	10.63	36.46	23.53	37.51	41.43	43.70	30.54
	Local	69.76	69.81	63.21	68.69	61.40	63.92	66.58	67.14	67.44	55.35	71.16	39.68	64.93	70.25	71.58	64.72
	FedAvg	72.28	72.46	66.31	71.08	64.22	66.45	69.61	70.03	69.34	58.32	73.60	42.84	68.34	72.92	74.55	67.49
	pFedGraph	68.40	69.45	62.88	68.35	61.48	63.82	65.72	66.70	66.57	55.56	70.37	40.52	65.14	69.28	70.96	64.34
	FedTSA	72.56	72.72	66.06	71.13	64.25	66.35	69.39	70.10	69.66	58.74	73.61	42.61	68.47	72.98	74.69	67.55
	FedCTTA	73.21	73.04	66.27	71.91	64.64	66.32	69.58	70.24	70.38	57.57	74.20	41.93	68.16	73.72	75.00	67.74
CIFAR100-C	Source	14.05	16.64	34.76	41.60	19.35	38.15	43.32	36.48	27.63	20.96	54.91	17.24	35.02	11.45	41.73	30.22
	Local	51.59	53.02	50.26	65.13	50.77	63.23	65.07	58.10	58.17	51.10	66.70	61.25	56.67	59.98	51.57	57.51
	FedAvg	57.33	58.60	56.74	69.07	57.51	68.94	70.92	64.29	64.19	57.44	72.40	67.36	63.70	66.33	57.56	63.50
	pFedGraph	52.05	53.42	50.48	65.60	51.19	63.61	65.49	58.39	58.64	51.39	67.08	61.64	57.14	60.46	52.06	57.91
	FedTSA	57.56	58.75	57.23	69.73	56.27	69.18	71.05	64.33	64.60	56.44	73.10	67.77	63.30	66.58	58.21	63.61
	FedCTTA	56.90	59.58	57.06	72.12	58.47	70.04	71.84	65.32	65.66	58.27	74.34	68.41	64.29	67.32	58.97	64.57

the model using backpropagation with self-supervised losses. Our proposed method focuses on fostering inter-client collaboration to share knowledge across clients and is orthogonal to the TTA strategies employed by individual clients. We evaluate our approach under both TTA-bn and TTA-grad settings for each dataset. For TTA-grad, we adopt entropy minimization as the local adaptation strategy, optimizing with SGD optimizer using a learning rate of 1.0×10^{-5}.

2) Baselines: We compare our proposed method with FedAvg and other regularization-based FL methods, including **FedAvg+FT**, **FedProx**, **FedAvgM**, **MOON**, and **pFedSD**, which we have adapted for test-time adaptation. Additionally, we evaluate our proposed method alongside other personalized federated learning (PFL) and test-time adaptation (TTA) methods. **pFedGraph** [21] constructs a collaboration graph based on model similarities and dataset size to enhance collaboration at the server side. **LDAWA** [22] aggregates model weights by measuring angular divergence between a client's model and the global model and adjusting the aggregation accordingly. **FedTSA** [13] leverages a temporal-spatial attention module to capture both intra-client temporal correlations and inter-client spatial correlations.

3) Spatial and Temporal Heterogeneity: To quantify heterogeneity in our experimental setup, we adopt the notions of temporal heterogeneity and spatial heterogeneity as defined in [13]. These metrics characterize the distribution shifts encountered by clients during continual test-time adaptation.

Spatial Heterogeneity: Spatial heterogeneity at time t, denoted as SH_t, measures the diversity of data distributions among clients:

$$SH_t = \frac{N_{\text{cls}}}{N} \tag{8}$$

where N_{cls} is the number of client clusters with consistent distribution shifts, and N is the total number of clients. Higher SH_t values indicate greater heterogeneity, with $SH_t = 1$ signifying unique distribution shifts for all clients.

Temporal Heterogeity: Temporal heterogeneity for the i-th client, denoted as TH_i, measures the frequency of distribution changes in streaming data:

$$TH_i = \frac{T_{\text{con}}}{T} \tag{9}$$

where T_{con} is the total duration of time slots with consistent distribution shifts, and T is the total duration of all time slots. Higher TH_i values indicate greater heterogeneity, with $TH_i = 1$ signifying a distinct distribution shift in every time slice.

B. Performance Analysis

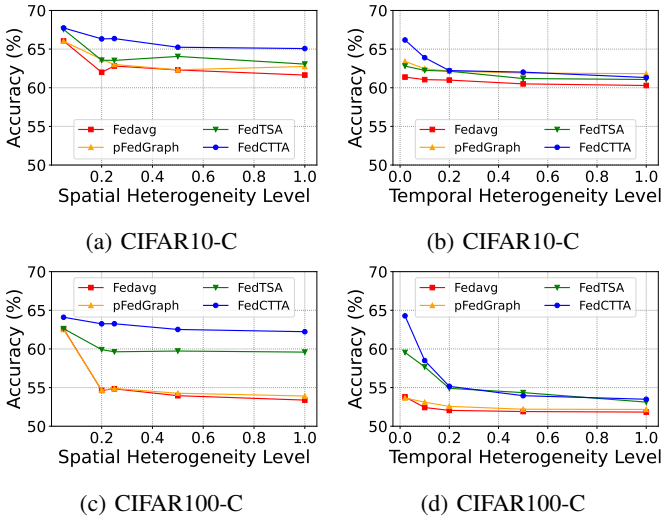
We assessed the performance of our method in both TTA-grad and TTA-bn settings under two distinct scenarios, with data distributed across 20 clients. In the first scenario, we simulated spatial heterogeneity (SH_t = 0.2) with 4 clusters, which we refer to as NIID, while the second scenario involved very low spatial heterogeneity (SH_t = 0.05) with a single cluster, referred to as IID. For both scenarios, temporal heterogeneity (TH_i) was kept constant at 0.02. Since TTA-bn does not require backward optimization, while many state-of-the-art methods, such as FedProx, MOON, pFedSD, and LDAWA, rely on gradient updates, the results from these methods under TTA-bn are consistent with those of FedAvg.

As shown in Table I, the baseline (No-adapt) struggles with corrupted datasets, achieving low accuracy across all settings, underscoring the challenges of distribution shifts in federated learning, particularly under non-IID scenarios. Local adaptation strategies improve performance in IID settings but remain inadequate against severe shifts. While FedAvg and other regularization-based methods such as FedProx, FedAvgM, pFedSD, and MOON perform well in IID settings, their effectiveness declines in NIID scenarios, where personalized federated learning (PFL) methods like FedTSA and pFedGraph achieve better results.

FedTSA and pFedGraph show notable improvements over previous methods, particularly under the NIID setting. They outperform FedAvg and FedProx in both CIFAR10-C and CIFAR100-C, highlighting the importance of inter-client collaboration for improving performance under corrupted conditions. Our method outperforms all other approaches across

TABLE III: The experimental setup and performance in the NIID scenario, where Clients 1–4, Clients 5–7, and Clients 8–10 share similar data distributions throughout the entire lifecycle, with a total of 10 clients.

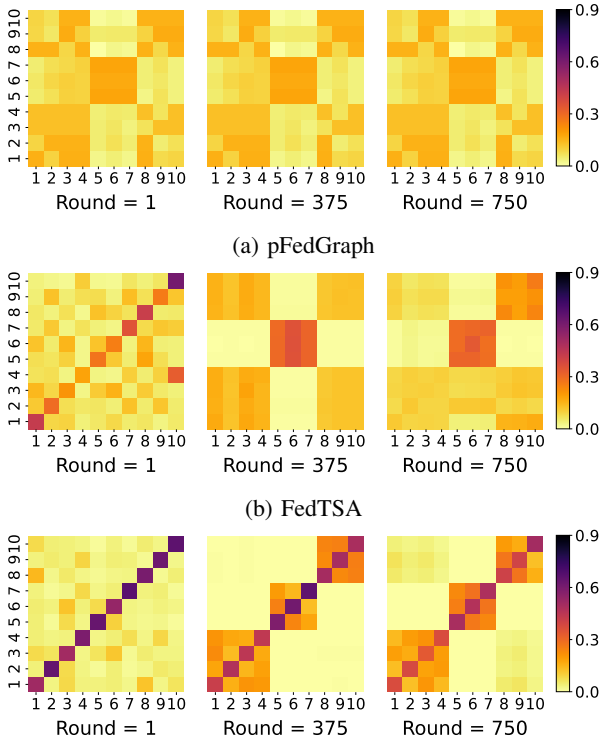
Time	t →															
Client 1-4	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*Jpeg*	*Mean*
Client 5-7	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*Jpeg*	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Mean*
Client 8-10	*Jpeg*	*Pixelate*	*Elastic*	*Contrast*	*Brightness*	*Fog*	*Frost*	*Snow*	*Zoom*	*Motion*	*Glass*	*Defocus*	*Impulse*	*Shot*	*Gaussian*	*Mean*
No-Adapt	71.56	66.60	58.06	46.70	59.88	47.00	60.02	63.48	51.10	48.36	51.22	43.92	64.44	71.44	71.68	58.36
pFedGraph	68.54	66.06	61.72	54.42	66.40	57.04	67.60	54.36	65.54	61.98	69.18	50.86	65.46	67.56	67.66	62.95
FedTSA	68.70	67.30	63.02	56.58	66.48	57.36	68.32	57.26	66.38	62.08	69.18	52.72	66.58	68.20	69.76	63.99
FedCTTA	68.64	67.84	67.20	58.66	67.76	58.56	70.52	58.98	67.18	64.00	70.26	57.02	67.68	70.56	70.22	65.67



(a) CIFAR10-C (b) CIFAR10-C

(c) CIFAR100-C (d) CIFAR100-C

Fig. 2: Comparison of CIFAR-10 and CIFAR-100 under different conditions. Greater spatial heterogeneity, indicating greater distribution differences among clients, leads to performance degradation across all methods, but FedCTTA consistently outperforms. Similarly, increasing temporal heterogeneity, causing frequent distribution shifts, further impacts performance.



(a) pFedGraph

(b) FedTSA

(c) FedCTTA

Fig. 3: The evolution of the collaboration matrix across federated rounds for three methods: pFedGraph, FedTSA, and our proposed FedCTTA. In FedCTTA, clients with similar data distributions naturally form clusters in aggregation weighting.

all settings in both CIFAR10-C and CIFAR100-C, achieving the highest accuracy. This demonstrates the effectiveness of our method in handling both spatial and temporal distribution shifts in federated learning scenarios.

C. Robustness Under Spatial and Temporal Heterogenity

To assess the robustness of our proposed method, we conduct experiments on CIFAR10-C and CIFAR100-C under varying degrees of spatial and temporal heterogeneity. Specifically, when analyzing spatial heterogeneity (SH), we fix temporal heterogeneity (TH_i) at 0.02, and conversely, when varying TH_i, we maintain SH_t at 0.2 to isolate the impact of each factor independently.

As shown in Figure 2a and 2c, our method consistently outperforms the baselines across different levels of spatial heterogeneity. While all methods experience a decline in accuracy as heterogeneity increases, FedAvg suffers the most significant drop, indicating its poor adaptability to spatially non-iid data. Our method demonstrates strong resilience, with only a minor performance decline, highlighting its ability to

effectively handle diverse client distributions.

As illustrated in Figure 2b and 2d, our method also exhibits strong robustness against temporal heterogeneity, outperforming traditional and personalized federated learning baselines in most cases. The core strength of our approach lies in its adaptive aggregation strategy, which leverages temporal similarity between clients to facilitate more effective inter-client collaboration. When temporal heterogeneity is low, distribution shifts occur gradually, allowing our method to retain and utilize historical knowledge more effectively. However, as temporal heterogeneity increases, abrupt shifts in data distribution diminish the relevance of past information. In case of high temporal heterogeneity (TH_i = 1), our method performs comparably to FedTSA.

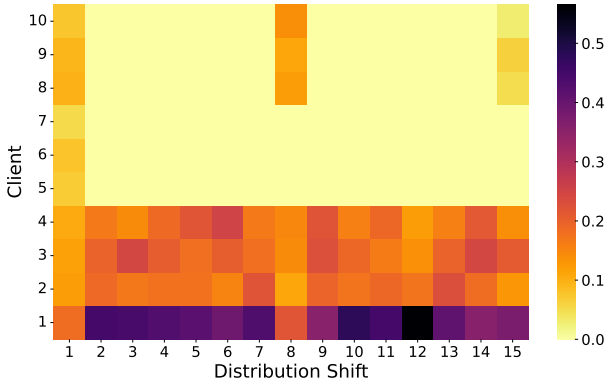
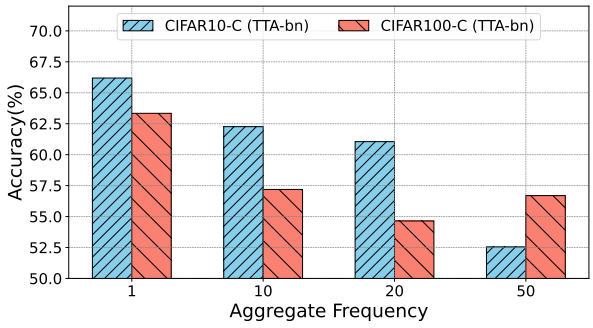


Fig. 4: Time-varying collaboration matrix for Client 1 of Our FedCTTA method in the NIID setting. Throughout all rounds, Clients 1–4 observe the same data distribution. At 8th distribution shift, Clients 8–10 also observe data from the same domain as Clients 1–4, and therefore have higher similarity with Client 1.

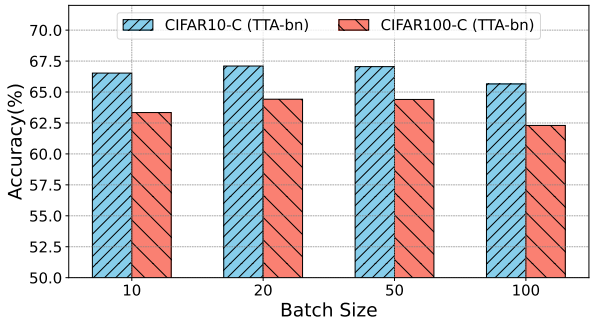
D. Collaboration Relationship Analysis

Figure 3 illustrates the evolution of the collaboration matrix across federated rounds for three methods: pFedGraph, FedTSA, and our proposed FedCTTA. The collaboration matrix quantifies the aggregation weights between clients, where higher values indicate stronger collaboration. This case study evaluates 10 clients on CIFAR10-C, divided into three groups based on the sequence of distribution shifts: Group 1 (Clients 1-4), Group 2 (Clients 5-7), and Group 3 (Clients 8-10). In Figure 3a, pFedGraph exhibits a scattered collaboration pattern across federated rounds, lacking structured inter-client relationships. Figure 3b shows that FedTSA initially relies on self-updates, with limited collaboration emerging over time, but without well-defined client clusters. In contrast, Figure 3c demonstrates that FedCTTA naturally clusters clients with similar data distributions, fostering structured and adaptive collaboration. Figure 4 further illustrates the time-varying collaboration matrix for Client 1 in this setup. Throughout all rounds, Clients 1–4 observe the same data distribution, forming a distinct cluster. At the 8th distribution shift, Clients 8–10 begin observing data from the same domain as Clients 1–4, leading to an increase in similarity with Client 1. These results highlight FedCTTA's effectiveness in leveraging inter-client similarities, where collaboration is determined based on similarity between output logits evaluated on random noise samples. A detailed quantitative comparison under different distribution shifts is provided in Table III.

V. Ablation Study



(a) Effect of aggregation frequency



(b) Effect of batch size

Fig. 5: (a) Gradual performance decline observed as aggregation interval increases across federated rounds. (b) Both very high and very low batch sizes impact performance, affecting generalization and stability. A balanced batch size is ideal.

A. Effect of Aggregation Frequency

We analyze the impact of aggregation frequency on test accuracy for CIFAR10-C and CIFAR100-C using the TTA-bn method. As shown in Figure 5a, increasing the aggregation interval negatively affects performance across federated rounds. A higher aggregation interval (e.g., 50) leads to reduced accuracy, suggesting that frequent model updates and collaboration between clients are crucial for maintaining performance.

B. Effect of Batch Size

Figure 5b illustrates the effect of batch size on accuracy. We observe that both very low (10) and very high (100) batch sizes result in suboptimal performance. A moderate batch size (20 or 50) achieves better results. In the federated setup, each client receives a smaller number of samples per domain, and when the batch size is too large, more frequent distribution shifts occur, leading to reduced performance. Conversely, a

very small batch size can cause unstable updates, impacting accuracy. This trend is consistent across both datasets.

TABLE IV: Comparison of test accuracy using distance measures for output logits and feature embeddings on CIFAR10-C dataset with the TTA-grad method under the NIID setting.

Data	Output Logit (Acc. %)				Feature (Acc. %)	
	Euclid	KL-div	CE	Cosine	Euclid	Cosine
Random Noise	66.19	61.62	61.60	61.62	62.07	61.63
Selected (CIFAR)	65.92	61.65	61.64	61.63	61.80	61.63

C. Ablation on Distance Metric and Auxiliary Dataset

Table IV presents a comparison of test accuracy using different distance measures for output logits and feature spaces, evaluated on random noise and selected CIFAR samples from the CIFAR10-C dataset, with the TTA-grad method under the NIID setting. Our analysis indicates that using random noise data to derive output logits for measuring similarity between client models, along with negative Euclidean distance as the similarity metric to construct the collaboration matrix for personalized model aggregation, achieves the best performance.

VI. Conclusion

We proposed FedCTTA, a privacy-preserving and efficient framework for continual test-time adaptation (CTTA) in federated learning (FL). By leveraging similarity-aware aggregation without sharing feature embedding, FedCTTA ensures adaptive knowledge transfer for different data distribution. It also integrates entropy minimization for confident adaptation to evolving target distributions. Experimental results on CIFAR10-C and CIFAR100-C show that FedCTTA outperforms existing methods in accuracy, robustness, and scalability, even under spatial and temporal heterogeneity. With its low computational overhead and constant memory footprint, FedCTTA is a promising solution for real-world FL applications requiring continual adaptation. Future work could extend FedCTTA to incorporate state of the art CTTA methods and evaluate on more dynamic real-world datasets.

References

[1] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8320–8329.

[2] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[3] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization ," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2018, pp. 1082–10 828.

[4] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[5] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2962–2971.

[6] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7191–7201.

[7] W. Bao, T. Wei, H. Wang, and J. He, "Adaptive test-time personalization for federated learning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[8] R. A. Marsden, M. Döbler, and B. Yang, "Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 2543–2553.

[9] E. Parliament and Council, "Regulation (eu) 2016/679 of the european parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016, official Journal of the European Union, 27 April 2016.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.

[11] Y. Tan, C. Chen, W. Zhuang, X. Dong, L. Lyu, and G. Long, "Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 27 167–27 180.

[12] L. Jiang and T. Lin, "Test-time robust personalization for federated learning," in *International Conference on Learning Representations (ICLR)*, 2023.

[13] J. Zhang, X. Liu, Y. Zhang, G. Zhu, J. Niu, and S. Tang, "Enabling collaborative test-time adaptation in dynamic environment via federated learning," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 4191–4202.

[14] M. Klabunde, T. Schumacher, M. Strohmaier, and F. Lemmerich, "Similarity of neural network models: A survey of functional and representational measures," *arXiv preprint arXiv:2305.06329*, 2023.

[15] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[16] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," 2019. [Online]. Available: https://arxiv.org/abs/1909.06335

[17] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[18] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 9, 2021, pp. 7865–7873.

[19] Q. Li, B. He, and D. Song, "Model-Contrastive Federated Learning ," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2021, pp. 10 708–10 717.

[20] H. Jin, D. Bai, D. Yao, Y. Dai, L. Gu, C. Yu, and L. Sun, "Personalized edge intelligence via federated self-knowledge distillation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 2, pp. 567–580, 2022.

[21] R. Ye, Z. Ni, F. Wu, S. Chen, and Y. Wang, "Personalized federated learning with inferred collaboration graphs," in *International Conference on Machine Learning*. PMLR, 2023, pp. 39 801–39 817.

[22] Y. A. Ur Rehman, Y. Gao, P. P. B. De Gusmao, M. Alibeigi, J. Shen, and N. D. Lane, "L-DAWA: Layer-wise Divergence Aware Weight Aggregation in Federated Self-Supervised Visual Representation Learning," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2023, pp. 16 418–16 427.

[23] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2021.

[24] M. J. Mirza, J. Micorek, H. Possegger, and H. Bischof, "The norm must go on: Dynamic unsupervised domain adaptation by normalization," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 745–14 755.

[25] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International conference on machine learning*. PMLR, 2022, pp. 16 888–16 905.

[26] J. Xu and S.-L. Huang, "A joint training-calibration framework for test-time personalization with label shift in federated learning," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 4370–4374.

[27] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: a standardized adversarial robustness benchmark," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[29] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.