This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Stable Federated Continual Test-Time Adaptation in Wild World

Liwen Wang[1], Xingbo Dong[1⋆], Iman Yi Liao[2], Zhe Jin[1⋆]
[1]Anhui Provincial Key Laboratory of Secure Artificial Intelligence,
Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging,
School of Artificial Intelligence, Anhui University, [2]University of Nottingham (Malaysia Campus)
liwenwang@stu.ahu.edu.cn {xingbo.dong, jinzhe}@ahu.edu.cn Iman.Liao@nottingham.edu.my

Abstract

*Federated Learning (FL) enables collaborative model training while preserving privacy, but faces challenges with client data heterogeneity and domain shifts during deployment. Although Personalized Federated Learning (PFL) mitigates heterogeneity, it typically requires labelled data from target clients, which is an impractical assumption. Test-Time Adaptation (TTA) offers label-free adaptation, yet its direct use in a continual federated setting risks destabilizing the global model and causing catastrophic forgetting. To address this, we consider the **Fed**erated **C**ontinual **T**est-**T**ime **A**daptation (FedCTTA) setting, where unlabeled clients arrive sequentially, requiring online adaptation and continuous global model updates. We propose BPFedCTTA, a framework that employs **B**ayesian **P**rior-guided **A**daptation (BPA) for stable local adaptation via Maximum a Posteriori estimation, and **U**ncertainty-**G**ated **S**ingle-client **A**ggregation (UGSA) to selectively integrate updates based on client uncertainty. This approach balances adaptation with knowledge retention, thereby mitigating forgetting. Extensive experiments on cross-domain classification and segmentation show BPFedCTTA outperforms existing FL, PFL, and TTA methods in sequential adaptation and global model improvement. The source code is available in https://github.com/LiwenWang919/BPFedCTTA.*

1. Introduction

Federated Learning (FL) [34] provides a foundational framework for collaborative model training across distributed data distributions while preserving data privacy. However, in practical deployments after training, FL systems often encounter test data from distributions that differ from the training data, arising from dynamic and non-stationary environments. This necessitates the development of models capable of continually adapting to new distributions after deployment.

⋆ Corresponding authors.

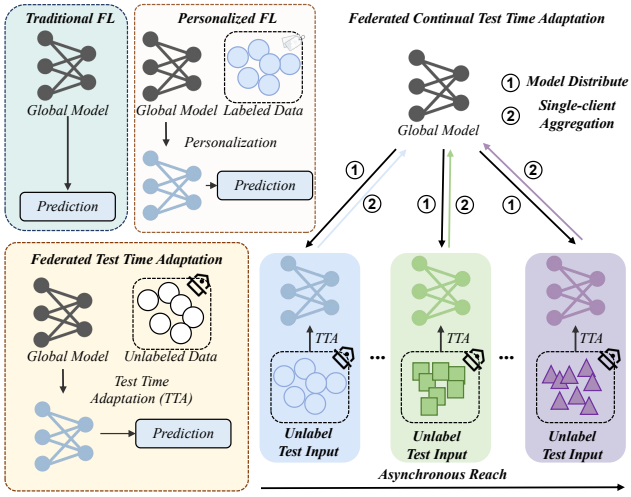


Figure 1. Illustration of the Federated Continual Test-Time Adaptation (FedCTTA) setting.

Federated Domain Generalization (FedDG) [28, 31, 56] can improve model performance on new data distributions, but since models are not fine-tuned on these new distributions, the gains are limited. While Personalized Federated Learning (PFL) [22, 45] offers adaptation mechanisms, it is largely unsuitable for the post-training scenario as it typically requires access to labels from the new data distributions for effective personalization. Although Test-Time Adaptation (TTA) methods [2, 3, 17, 38, 48] has shown effectiveness in single-source label-free adaptation scenarios, it remains fundamentally mismatched with the federated context: the FL global model represents a mixture of knowledge aggregated from multiple heterogeneous sources, whereas standard TTA operates on models trained in centralized settings with single-domain assumptions. Direct application of TTA in FL violates this structural premise, as unsupervised updates on heterogeneous client data can disrupt the carefully balanced global model, leading to performance degradation.

Recent work [40] has pioneered continual test-time adaptation (CTTA) in federated learning. However, their setting

is limited to predefined client clusters with synchronous updates, which does not reflect the sequential and asynchronous nature of real-world deployments. Moreover, their approach primarily addresses spatial heterogeneity with known shifts and does not account for the extreme temporal heterogeneity arising from clients that arrive sequentially with completely unrelated distributions. In contrast, our work is designed to handle such unforeseen and extreme spatio-temporal shifts while maintaining strict data privacy and enabling safe global model evolution. Additionally, their method lacks mechanisms to constrain local adaptation stability or quantify update reliability in continual settings, leaving the global model vulnerable to corruption from noisy client updates over time.

Under this setting, direct use of existing TTA or PFL techniques is inadequate. We identify two distinctive challenges:

- **(C1) Unstable label-free local adaptation.** The FL global model is a generalized average residing in a flat loss basin to serve multiple clients [4, 34, 51]. This structure, while good for generalization, makes it highly sensitive to noisy updates [38]. A naive TTA update, driven by a single client's (potentially noisy) unlabeled data, lacks a proper regularizer and can easily push the model parameters out of this stable basin, leading to rapid overfitting to the new client and catastrophic local model drift.
- **(C2) Unsafe continual global evolution.** In this sequential, label-free setting, the server has no ground truth to verify the quality or reliability of an adapted client's update. A single client's update, if overfitted or based on a major domain shift, can be destructive. Naively aggregating this unvetted update (*e.g.*, via FedAvg) can poison the global model and lead to catastrophic forgetting of the knowledge acquired from previous clients. Critically, standard continual learning strategies that use data replay are infeasible due to FL's core privacy constraints.

We provide a theoretical justification of both challenges in the Appendix. To address these coupled challenges, we introduce **B**ayesian **P**rior-guided **Fed**erated **C**ontinual **T**est-**T**ime **A**daptation (BPFedCTTA), a unified probabilistic framework that reformulates the entire continual adaptation process through a Bayesian lens. In this Bayesian view, the global model acts as an evolving prior, and each client's local adaptation is reformulated as approximate posterior inference guided by its unlabeled data.

Specifically, to tackle **C1** (local instability), we propose the **B**ayesian **P**rior-guided **A**daptation (BPA) module. BPA treats the global model parameters as the mean of a Gaussian prior and performs Maximum a Posteriori (MAP) estimation for the local update. This Bayesian formulation provides a principled, data-driven regularization, anchoring the local model to the high-probability region of the global prior and preventing the model from drifting or overfitting to noisy unlabeled data. This probabilistic approach provides a natural solution for **C2** (unsafe aggregation). By treating the local adaptation as posterior inference, we can rigorously quantify the predictive uncertainty of the adapted model (*e.g.*, via output entropy), which serves as a direct measure of the update's reliability. We leverage this to design the **U**ncertainty-**G**ated **S**ingle-client **A**ggregation (UGSA) mechanism. UGSA uses this client-quantified uncertainty to gate the aggregation process. Reliable, low-uncertainty updates are integrated, while high-uncertainty (unreliable) updates are attenuated. This selection process, which approximates sequential Bayesian filtering, ensures the global model evolves safely, integrating new knowledge without being corrupted by unreliable updates, thus mitigating catastrophic forgetting.

Our key contributions are summarized as follows:

1. We define and formalize the realistic and challenging **Fed**erated **C**ontinual **T**est-**T**ime **A**daptation (FedCTTA) setting, which differs from existing approaches that operate under constrained scenarios [40] and highlights the core coupled challenges of unstable local adaptation (C1) and unsafe global aggregation (C2) in sequential, unlabeled, and heterogeneous federated environments.
2. We propose BPFedCTTA, a novel Bayesian framework that unifies adaptation and aggregation to solve these challenges. Our innovations include: (a) Bayesian Prior-guided Adaptation (BPA), which reformulates local TTA as MAP estimation to ensure stable, regularized adaptation, and (b) Uncertainty-Gated Single-client Aggregation (UGSA), which introduces a principled uncertainty-based filter to enable safe, continual global model evolution from unreliable updates.
3. Extensive experiments on cross-domain classification and segmentation show that BPFedCTTA achieves superior adaptability on new clients while maintaining global model stability, significantly outperforming existing FL, PFL, and TTA baselines.

2. Related Work

2.1. Personalized Federated Learning

Personalized Federated Learning (PFL) addresses data heterogeneity across clients through various personalization strategies. Common approaches include local fine-tuning, regularization methods like FedProx [26], meta-learning frameworks such as Per-FedAvg [11], and parameter decoupling techniques. Some methods personalize feature extractors while sharing classifiers [29], while others share feature extractors but personalize classifiers [1, 7]. Additional strategies employ prototype alignment [46] or bi-level optimization as in pFedMe [45] and Ditto [27]. However, most PFL methods assume labeled client data for personalization—an impractical requirement in real-world deployments where new clients lack annotations. Furthermore, these approaches focus solely on the training phase and cannot handle test-time domain shifts or sequential client arrival, making them

unsuitable for the Federated Continual Test-Time Adaptation (FedCTTA) setting.

2.2. Federated Domain Generalization

Domain Generalization (DG) aims to learn a model from multiple source domains that can generalize well to *unseen* target domains [25]. This goal is aligned with ours in addressing domain shift under data privacy constraints. Typical DG methods focus on the training phase, falling into three main categories: style augmentation via feature-level transformations [16], learning domain-invariant representations [5, 10], and other regularization techniques. However, centralized training across domains is often infeasible due to data privacy constraints [44]. Federated Domain Generalization (FedDG) addresses this by training local models on distributed source domains and aggregating them to enhance global generalization [28, 31]. Despite recent progress, FedDG remains underexplored. Methods like EL-CFS share spectral information to improve generalization but may still risk privacy [31], while others, such as adaptive weighting schemes, face practical limitations [56]. However, Static FedDG models, limited by their source domains, cannot adapt post-deployment. Our FedCTTA setting, however, involves unlabeled clients with novel distributions. This requires continual, online adaptation, motivating our work to integrate TTA with federated learning.

2.3. Test-Time Adaptation

Test-Time Adaptation (TTA) adapts a pre-trained model to an unlabeled, unseen test distribution at inference time. Prevailing strategies often rely on output-level objectives, such as entropy minimization [20, 38, 47, 53] or self-training with pseudo-labels [13, 33, 55]. While computationally efficient, these approaches are prone to error accumulation as incorrect pseudo-labels are propagated, particularly under significant domain shifts. To improve robustness, other methods employ feature-level consistency regularization [18, 32, 49, 57] or contrastive learning [6, 54] to enhance feature discrimination. Nonetheless, many of these techniques still depend on model predictions (e.g., from a teacher model) that can be unreliable for out-of-distribution (OOD) samples [9]. A critical challenge emerges when applying TTA in the context of Federated Learning (FL), related to the geometry of the global model's loss landscape. Federated models aggregated via averaging mechanisms like FedAvg tend to converge to flat minima, a beneficial artifact of aggregating heterogeneous client objectives [34, 35]. While flat minima are proven to enhance generalization in centralized settings [15, 19], their interaction with TTA in federated contexts is largely unexplored. Specifically, standard unsupervised TTA updates [38, 47], while effective for single-domain adaptation, can destabilize the carefully calibrated global model. These updates, driven by a local test distribution, risk deviating from the flat basin, leading to overfitting on the current client's data and compromising the model's robust generalization capabilities [2, 17]. This highlights a critical need for TTA methods that are explicitly regularized to preserve the global model's flatness, ensuring robust adaptation without sacrificing cross-domain performance.

3. Methodology

In this section, we present our proposed framework, Bayesian Prior-guided Federated Continual Test-Time Adaptation (BPFedCTTA Fig. 2), designed for the challenging Federated Continual Test-Time Adaptation (FedCTTA) setting (Sec. 3.1). Inspired by Bayesian learning principles, BPFedCTTA adopts a two-level structure: clients perform local adaptation by treating the global model as a prior and approximating Bayesian inference (Sec. 3.2), while the server facilitates continual global model evolution through an uncertainty-aware aggregation mechanism (Sec. 3.3). This approach aims to achieve robust personalization on sequential unlabeled clients while mitigating catastrophic forgetting.

3.1. Problem Setting (FedCTTA)

As shown in Fig. 1, we consider an FL system initially trained on N source clients $\{\mathcal{C}_s\}_{s=1}^{N}$ with labeled data $\mathcal{D}_s = \{(\mathrm{X}_i^s, \mathrm{Y}_i^s)\}_{i=1}^{n_s} \sim \mathrm{P}_s$ for each client $\mathcal{C}_s$, which produces a pre-trained global model $\theta_G^{(0)}$ for deployment.

Definition 1 FedCTTA. *During depolyment, a sequence of K new target clients $\{\mathrm{C}_k\}_{k=1}^{K}$ arrive over time $t_1, t_2, \ldots, t_K$. Each client C_k has only an unlabeled data stream $\mathrm{X}_k = \{\mathrm{X}_i^k\}_{i=1}^{n_k} \sim \mathrm{P}_k$, where the client distributions P_k (and P_s) are heterogeneous. At time t_k, client C_k joins, downloads the current global model $\theta_G^{(k-1)}$ (where $\theta_G^{(k)}$ denotes the global model state after considering clients up to time t_k), performs local test-time adaptation using X_k to obtain a personalized model estimate θ'_k, and transmits relevant information back to the server. The server asynchronously receives these updates and is tasked with continually refining the global model $\theta_G^{(k-1)} \to \theta_G^{(k)}$.*

Lemma 1 Objective function of FedCTTA. *The objective function can be formulated as:*

$$\min_{\{\theta'_k, \theta_G^{(k)}\}_{k=1}^{K}} \sum_{k=1}^{K} \left[\underbrace{\tilde{\mathcal{R}}_k(\theta'_k)}_{\textit{Obj. 1: Local Stability}} + \lambda \cdot \underbrace{\left(\sum_{s=1}^{N} \mathcal{R}_s(\theta_G^{(k)}) \right)}_{\textit{Obj. 2: Global Evolution}} \right] \tag{1}$$

where $\tilde{\mathcal{R}}_k(\theta) = \mathbb{E}_{x \sim \mathrm{P}_k}[\ell(\theta; x)]$ denotes the expected risk of model θ on the distribution of client C_k, with $\ell(\cdot)$ being the task-specific loss function. Similarly, $\mathcal{R}_s(\theta)$ represents the risk on source client $\mathcal{C}_s$. The hyperparameter $\lambda > 0$ controls the trade-off between personalization and global

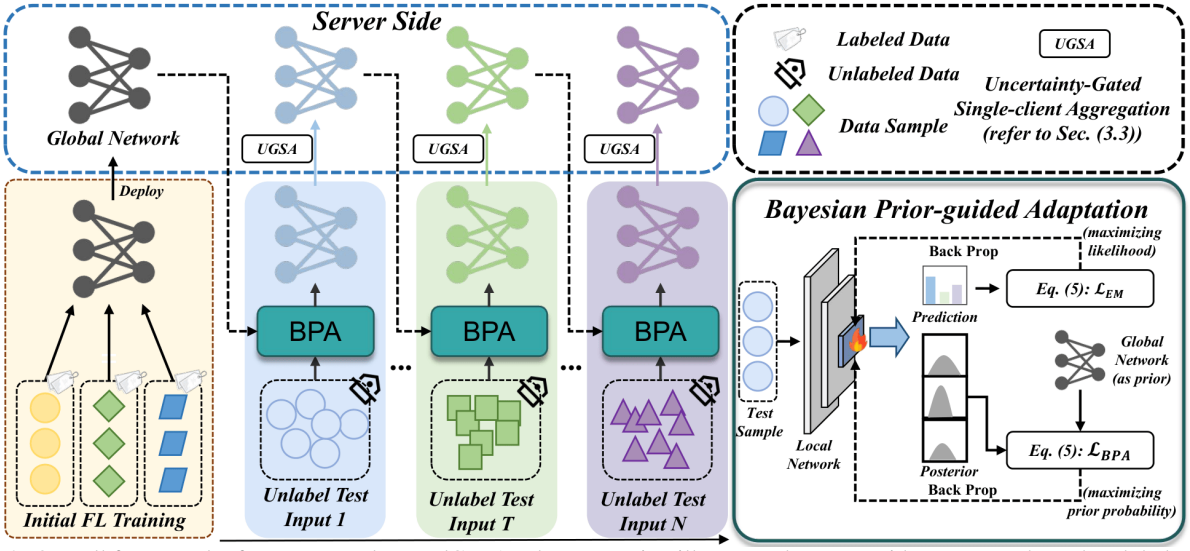


Figure 2. Overall framework of our proposed BPFedCTTA. The top section illustrates the server-side process, where the global model is continually updated via Uncertainty-Gated Single-client Aggregation (UGSA) based on client-reported uncertainty. The bottom section depicts the client-side Bayesian Prior-guided Adaptation (BPA) module, which leverages the global model as a prior to perform stable local test-time adaptation via Maximum a Posteriori estimation, using only unlabeled test data.

stability. Note that for target clients C_k, only unlabeled data is available during adaptation, making the estimation of $\tilde{\mathcal{R}}_k$ challenging in practice.

In the FL setting, our work considers a realistic and challenging scenario where unlabeled clients arrive sequentially and asynchronously, exhibiting extreme spatio-temporal heterogeneity. Specifically, each client originates from a distinct and unforeseen data distribution, with no temporal overlap or inter-client relationship. We enforce strict privacy by prohibiting any feature or logit sharing, and focus on continual global model evolution to maintain a universally adaptable model. By contrast, [40] adopts synchronous cluster-based clients, focuses on spatial heterogeneity with predefinable shifts, relaxes privacy via sharing noise-sample logits for similarity calculation, and prioritizes client personalization over global model improvement.

3.2. Bayesian Prior-guided Adaptation

Motivation. Adapting a complex model θ using only unlabeled data X_k is inherently ill-posed and prone to instability or overfitting, especially when X_k is limited or noisy. Standard TTA methods often lack sufficient regularization tailored for FL models residing in flat loss basins [34]. Inspired by Bayesian Inference where a prior regularizes learning from limited data, we propose BPA, which leverages the current global model $\theta_G^{(k-1)}$ as prior to guide local adaptation.

Formalism. We treat the local adaptation on client k as approximating the posterior distribution $p(\theta|\mathrm{X}_k, \theta_G^{(k-1)})$. We define a Gaussian prior centered at the current global model, representing the belief that the optimal parameters should remain close to the well-generalized global solution:

$$p(\theta|\theta_G^{(k-1)}) = \mathcal{N}(\theta|\theta_G^{(k-1)}, \Sigma_0), \quad \Sigma_0 = \sigma_0^2 I. \tag{2}$$

The likelihood of the unlabeled data X_k given parameters θ is modeled implicitly via a TTA loss function $\mathcal{L}$, which encourages desirable prediction properties such as high confidence and local consistency:

$$p(\mathrm{X}_k|\theta) \propto \exp(-\mathcal{L}(\theta; \mathrm{X}_k)). \tag{3}$$

Instead of computing the full posterior, BPA seeks the Maximum a Posteriori (MAP) estimate θ'_k, corresponding to the mode of the posterior distribution:

$$\begin{aligned} \theta'_k &= \arg\max_\theta p(\theta|\mathrm{X}_k, \theta_G^{(k-1)}) \\ &\propto \arg\max_\theta \left[\log p(\mathrm{X}_k|\theta) + \log p(\theta|\theta_G^{(k-1)})\right] \\ &= \arg\min_\theta \left(\mathcal{L}(\theta; \mathrm{X}_k) - \log \mathcal{N}(\theta|\theta_G^{(k-1)}, \Sigma_0)\right). \end{aligned} \tag{4}$$

Expanding the log-prior yields the following optimization objective:

$$\begin{aligned} \theta'_k &= \arg\min_\theta \Big(\mathcal{L}(\theta; \mathrm{X}_k) \ + \tfrac{1}{2}(\theta - \theta_G^{(k-1)})^\top \Sigma_0^{-1} (\theta - \theta_G^{(k-1)})\Big) \\ &= \arg\min_\theta \left(\mathcal{L}_{\mathrm{EM}}(\theta) + \mathcal{L}_{\mathrm{BPA}}(\theta)\right), \end{aligned} \tag{5}$$

where $\mathcal{L}_{\mathrm{BPA}}$ is the Bayesian adaptation objective and $\mathcal{L}_{\mathrm{EM}}$ is entropy minimization loss.

To enhance robustness under domain shift, we further let Σ_0 adaptively scale with model uncertainty estimated from

prediction entropy:

$$\Sigma_0^{-1} = \mathrm{diag}\left(\frac{1}{\sigma_0^2} \cdot \mathbb{E}_{x \in \mathrm{X}_k}[\mathcal{H}(p(y|x;\theta_G^{(k-1)}))]\right), \quad (6)$$

where $\mathcal{H}$ denotes Shannon entropy. This dynamic precision effectively tightens the prior when the model is confident and relaxes it under high uncertainty, yielding a more principled and data-dependent Bayesian regularization.

The client obtains θ_k' by performing several gradient descent steps on $\mathcal{L}_{\text{BPA}}$ and $\mathcal{L}_{\text{EM}}$. The Bayesian-derived quadratic regularizer anchors the adaptation within the high-probability region of the global model while the adaptive covariance in Eq. 6 modulates learning strength according to uncertainty, thereby preventing drift and ensuring stable test-time adaptation on unlabeled, heterogeneous data streams.

3.3. Uncertainty-Gated Single-client Aggregation

Motivation. After clients adapt locally via BPA, the server receives their updated parameters. To prevent performance degradation from domain shift or unreliable updates, UGSA integrates these contributions using an uncertainty-weighted average for stable global model evolution.

Formalism. We view the aggregation step as a Bayesian belief update on the global model parameters. Let the global model at round $(k-1)$ follow a Gaussian distribution:

$$p(\theta_G^{(k-1)}) = \mathcal{N}(\theta_G^{(k-1)}, \Sigma_G^{(k-1)}), \quad (7)$$

and the adapted client posterior be approximated as

$$q_k(\theta) = \mathcal{N}(\theta_k', \Sigma_k), \quad (8)$$

where Σ_k encodes the adaptation uncertainty on client k. Following the principle of Bayesian fusion, the new global distribution $p(\theta_G^{(k)})$ is derived by minimizing a weighted sum of Kullback–Leibler divergences, balancing fidelity to the client's adapted posterior against consistency with the previous global prior:

$$p(\theta_G^{(k)}) = \arg\min_{p \in \mathcal{P}} \left\{ w_k KL(q_k \| p) + (1 - w_k) KL(p_G^{(k-1)} \| p) \right\}, \quad (9)$$

where $w_k \in [0,1]$ is the reliability gate derived from client uncertainty. This formulation is mathematically equivalent to a Bayesian belief update with tempered likelihoods. Specifically, minimizing the objective in Eq. (9) yields a posterior that is a geometric mean (in density space) of the prior and the client contribution:

$$p(\theta_G^{(k)}) \propto p(\theta_G^{(k-1)})^{1-w_k}\, q_k(\theta)^{w_k}, \quad (10)$$

where w_k denotes the reliability gate derived from client uncertainty. Under Gaussian assumptions, the closed-form update for mean and covariance becomes:

$$\begin{aligned} (\Sigma_G^{(k)})^{-1} &= (\Sigma_G^{(k-1)})^{-1} + w_k \Sigma_k^{-1}, \\ \theta_G^{(k)} &= \Sigma_G^{(k)} \left[(\Sigma_G^{(k-1)})^{-1} \theta_G^{(k-1)} + w_k \Sigma_k^{-1} \theta_k' \right]. \end{aligned} \quad (11)$$

Uncertainty gating. We quantify the uncertainty of client k's update using the predictive entropy of the adapted model:

$$\sigma_k = \mathbb{E}_{x \in \mathrm{X}_k}[\mathcal{H}(p(y|x;\theta_k'))], \quad (12)$$

and define the gating weight as

$$w_k = \frac{\exp(-\beta \sigma_k)}{\sum_j \exp(-\beta \sigma_j)}, \quad (13)$$

where β controls the sharpness of uncertainty attenuation. This exponential weighting ensures that clients with higher confidence (low entropy) contribute more strongly to the global update, while uncertain clients are down-weighted to prevent noisy drift.

Interpretation. Eq. 11 can be viewed as a federated Bayesian filtering process, where the global parameters evolve recursively:

$$p(\theta_G^{(k)} | \mathrm{X}_{1:k}) \propto q_k(\theta | \mathrm{X}_k)^{w_k}\, p(\theta_G^{(k-1)} | \mathrm{X}_{1:k-1}). \quad (14)$$

This formulation reveals UGSA as a precision-weighted evidence accumulation mechanism that continually refines the global model under unlabeled and heterogeneous conditions. The adaptive covariance update in Eq. 11 propagates uncertainty from each client to the server, allowing the global model to maintain calibrated confidence across rounds.

In practice, we maintain a diagonal approximation of Σ_G for computational efficiency and normalize w_k across participating clients. The update can be equivalently expressed as:

$$\theta_G^{(k)} = \theta_G^{(k-1)} + \gamma\, w_k\, (\theta_k' - \theta_G^{(k-1)}), \quad (15)$$

where γ is the server-side learning rate.

4. Experiments

4.1. Datasets

Image Classification Benchmark

CIFAR10-C Benchmark extends CIFAR10 [21], comprising 32×32 color images from 10 classes. It includes 15 different corruptions, each at five severity levels, applied to the test images of CIFAR-10, resulting in a total of 10,000 images. **CIFAR100-C Benchmark** extends CIFAR-100 [21], containing 32×32 color images from 100 classes. It includes 15 different corruptions, each at five severity levels, applied to the test images of CIFAR-100.

Image Segmentation Benchmark

CitySpace to ACDC Benchmark Cityscapes Dataset [8] provides images and annotations for 30 categories, which are grouped into 8 classes. ACDC [42] comprises 4006 images, which are evenly distributed across four common adverse visual conditions: fog, night, rain, and snow.

Retinal Fundus Segmentation Benchmark comprises five public datasets from different medical centers, used for the

Table 1. Image Classification results (Accuracy in %) on the CIFAR10-C and CIFAR100-C. All results employ ResNet-18 [14] as backbone for training on the source dataset (CIFAR10 and CIFAR100) with a round of 50 and a client of 10. Bold text indicates the best.

Datasets	Method	Time/Client →															Mean
		Gauss.	Shot	Imp.	Def.	Glass	Mot.	Zoom	Snow	Fro.	Fog	Bri.	Con.	Ela.	Pix.	JPEG	
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	70.28	70.46	64.31	69.08	62.22	64.45	67.61	68.03	67.34	56.32	71.60	40.84	66.34	70.92	72.55	65.49
	FedProx	70.29	70.13	63.40	69.75	60.27	63.99	66.83	68.04	67.17	54.30	71.49	41.02	64.30	68.73	71.11	64.72
	FedBN	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
	FedGA	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
	Tent	70.15	70.23	63.85	69.42	62.18	64.56	67.24	67.89	68.12	56.08	71.89	40.35	65.67	71.03	72.31	65.41
	CoTTA	70.89	70.95	64.72	70.18	63.05	65.28	68.13	68.74	68.85	56.92	72.46	41.18	66.52	71.78	73.04	66.28
	BeCoTTA	71.85	71.96	65.89	71.24	64.32	66.38	69.27	69.88	69.91	58.05	73.42	42.47	67.86	72.64	74.18	67.36
	TCA	71.42	71.58	65.36	70.75	63.78	65.94	68.79	69.36	69.42	57.63	73.05	41.89	67.28	72.31	73.67	66.92
	FedTHE+	72.42	72.58	66.49	71.83	64.95	67.08	69.92	70.61	70.48	58.79	74.05	43.12	68.57	73.24	74.86	67.93
	ATP	72.13	72.27	66.18	71.52	64.67	66.72	69.58	70.25	70.13	58.41	73.78	42.76	68.19	72.89	74.52	67.65
	FedCTTA	73.26	73.73	66.50	71.95	65.37	66.82	69.59	70.47	69.71	57.58	74.79	44.28	**68.99**	72.79	**75.60**	68.10
	TTA-FedDG	72.47	72.61	66.58	71.94	65.03	**67.22**	70.06	70.74	**70.62**	**58.96**	74.12	43.28	68.72	**73.38**	74.99	68.09
	BPFedCTTA	**74.38**	**74.53**	**67.69**	**73.51**	**66.14**	66.95	**70.28**	**71.24**	69.52	56.14	**75.21**	**44.56**	68.04	73.35	75.12	**68.44**
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	53.33	54.60	52.74	65.07	53.51	64.94	66.92	60.29	60.19	53.44	68.40	63.36	59.70	62.33	53.56	59.50
	FedProx	52.38	53.90	50.83	64.92	51.93	62.95	65.03	58.15	58.30	51.77	67.51	61.04	57.48	61.05	51.62	57.92
	FedBN	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
	FedGA	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
	Tent	52.84	54.26	51.59	66.38	52.08	64.52	66.24	59.35	59.42	52.35	68.15	62.48	58.12	61.63	53.21	59.02
	CoTTA	53.67	55.13	52.74	67.25	53.16	65.48	67.19	60.42	60.58	53.42	69.28	63.57	59.34	62.78	54.38	60.18
	BeCoTTA	55.36	56.84	55.03	68.98	55.29	67.37	69.08	62.53	62.87	55.53	71.53	65.72	61.75	65.05	56.69	62.47
	TCA	54.52	55.98	53.89	68.12	54.23	66.43	68.14	61.48	61.73	54.48	70.41	64.65	60.55	63.92	55.54	61.33
	FedTHE+	56.94	58.43	57.18	70.58	57.28	69.13	70.85	64.51	65.03	57.51	73.65	67.74	64.03	67.19	58.87	64.61
	ATP	56.21	57.69	56.16	69.83	56.34	68.30	70.01	63.57	64.00	56.57	72.64	66.78	62.94	66.17	57.83	63.59
	FedCTTA	57.18	58.92	57.73	72.35	57.73	69.64	70.39	66.00	65.82	58.19	74.22	68.03	63.96	67.75	58.82	64.45
	TTA-FedDG	**57.62**	59.12	**58.15**	71.28	58.17	69.91	71.64	65.40	66.01	58.40	74.61	68.65	65.07	68.16	59.86	65.58
	BPFedCTTA	57.28	**60.17**	57.97	**73.02**	**59.45**	**70.85**	**72.92**	**66.26**	**67.06**	**59.37**	**75.89**	**69.20**	**65.42**	**68.83**	**60.01**	**67.58**

joint segmentation of the optic disc (OD) and optic cup (OC). The datasets are Site A (RIM-ONE [12]), Site B (REFUGE [39]), Site C (ORIGA [58]), Site D (REFUGE-Validation [39]), Site E (REFUGE-Test [39]), and Site F (Drishti-GS [43]), containing 159, 400, 650, 400, 400, and 101 images, respectively.

MRI Prostate Segmentation Benchmark is for prostate segmentation and is gathered from 6 different medical centers. The data consists of T2-weighted MRI scans from three publicly datasets: NCI-ISB113, I2CVB, and PROMISE12. The six clinical centers (clients) are denoted as RUNMC, BMC, I2CVB, UCL, BIDMC, and HK [24, 30, 37].

4.2. Experimental Setup

Experimental Protocol. As represented in Fig. 1, our evaluation protocol is divided into two distinct phases to simulate a real-world deployment:

- **Phase 1: Initial Federated Training.** We first train an initial global model, θ_G, using the standard FedAvg algorithm. This phase involves only the designated source clients and their labeled data.
- **Phase 2: Continual Test-Time Adaptation.** After the initial training, the new clients are introduced sequentially. For each client, the current global model θ_G performs BPA on its unlabeled data stream. After adaptation, knowledge is fused back into θ_G using the proposed aggregation mechanism. This updated global model is then deployed to the next client, simulating a continually evolving system.

Unlike [40], which operates under synchronized cluster-based settings with predefined shifts, our evaluation protocol reflects extreme spatio-temporal heterogeneity by simulating sequential client arrival from unseen and diverse domains. This setting is more aligned with real-world deployments where clients join asynchronously and exhibit completely unrelated data distributions.

Comparison Methods. To comprehensively compare relevant approaches, we selected several methods as baselines for comparison: 1) standard federated learning methods (FedAvg [34]), 2) personalized federated learning methods (FedProx [26], FedBN [52], FedGA [56]), 3) standard test-time adaptation methods (Tent [47]) and continual test-time adaptation methods (CoTTA [48], BeCoTTA [23], TCA [36]), and 4) methods combining TTA with federated learning (ATP [2], FedTHE+ [17], FedCTTA [40], TTA-FedDG [28]). We implemented the aforementioned methods on the brand-new FedCTTA setting and reported the performance.

Implementation Details. We use a consistent setup across all benchmarks for fair comparison. We employ ResNet-18 [14] for CIFAR10-C and CIFAR100-C (severity 5), Segformer-B5 [50] for Cityscapes-to-ACDC, and U-Net [41] for the Retinal Fundus and Prostate MRI benchmarks.To simulate data heterogeneity for CIFAR and Cityscapes, we partition data using a Dirichlet distribution ($\alpha = 0.5$). For the medical datasets (Retinal Fundus and Prostate MRI), we

Table 2. Semantic segmentation results (mIoU in %) on the Cityscapes-to-ACDC benchmark. The model is continually adapted to the four adverse conditions without model reset. All results are evaluated based on the Segformer-B5 [50] architecture. Bold text indicates the best.

Method	Sequence 1 →				Mean	Sequence 2 →				Mean	Sequence 3 →				Mean	Sequence 4 →				Mean
	Fog	Night	Rain	Snow		Snow	Rain	Night	Fog		Night	Fog	Snow	Rain		Rain	Snow	Fog	Night	
Source	69.10	40.30	59.70	57.80	56.73	57.80	59.70	40.30	69.10	56.73	40.30	69.10	57.80	59.70	56.73	59.70	57.80	69.10	40.30	56.73
Local	63.45	38.92	55.83	54.27	53.12	54.05	55.64	38.76	63.28	52.93	38.59	63.12	53.84	55.46	52.75	55.27	53.62	62.95	38.43	52.57
FedAvg	64.78	39.85	56.94	55.63	54.30	55.34	56.68	39.62	64.52	54.04	39.38	64.25	55.04	56.41	53.77	56.15	54.75	63.98	39.15	53.51
FedProx	65.40	40.52	56.80	56.16	55.07	56.39	57.20	40.82	65.10	55.25	40.78	65.43	56.01	56.88	55.37	56.62	53.99	65.69	41.72	54.80
FedBN	65.32	40.47	57.62	56.38	54.95	56.21	57.45	40.32	65.18	54.79	40.16	65.03	56.03	57.28	54.63	57.11	55.86	64.87	40.01	54.46
FedGA	66.15	41.23	58.37	57.14	55.72	56.89	58.13	41.05	65.94	55.50	40.86	65.72	56.63	57.89	55.28	57.65	56.38	65.50	40.68	55.05
Tent	67.84	42.15	59.83	58.42	57.06	56.84	58.12	40.78	66.23	55.49	39.54	64.86	55.48	56.73	54.15	55.18	53.96	63.45	38.27	52.72
CoTTA	69.52	43.28	61.45	60.13	58.60	59.67	60.83	42.65	68.94	58.02	42.08	68.42	59.28	60.35	57.53	59.82	58.74	67.85	41.53	56.99
BeCoTTA	70.82	44.63	62.78	61.45	59.92	60.38	61.49	43.27	69.16	58.58	42.16	67.85	59.42	60.54	57.49	59.43	58.15	66.38	41.28	56.31
TCA	71.25	45.18	63.24	61.97	60.41	61.52	62.86	44.73	70.84	59.99	44.25	70.36	61.03	62.41	59.51	61.87	60.41	69.78	43.72	58.95
FedTHE+	72.14	46.32	64.38	63.25	61.52	62.97	64.12	45.94	71.93	61.24	45.55	71.71	62.68	63.85	60.95	63.57	62.38	71.48	45.15	60.65
ATP	71.68	45.73	63.79	62.58	60.95	62.24	63.48	45.36	71.42	60.63	44.98	71.15	61.89	63.16	60.30	62.83	61.53	70.87	44.59	59.96
FedCTTA	72.24	46.47	64.65	63.29	61.66	64.38	65.43	47.97	73.35	62.78	48.57	73.29	65.38	66.59	63.46	66.45	66.51	73.90	46.26	63.28
TTA-FedDG	72.59	46.87	64.93	63.89	62.07	63.68	64.73	46.52	72.38	61.83	46.16	72.16	63.46	64.52	61.58	64.30	63.23	71.93	45.79	61.31
BPFedCTTA	**73.25**	**47.68**	**65.84**	**64.73**	**62.88**	**65.82**	**66.95**	**48.93**	**74.12**	**63.96**	**49.75**	**74.86**	**66.48**	**67.63**	**64.68**	**68.15**	**67.03**	**75.24**	**50.32**	**65.19**

Table 3. Performance comparison (Dice Score, %) on the Retinal Fundus and Prostate MRI segmentation benchmark. Results are shown for individual target domains(using Sites A B C as source domains for continually testing on Site D E F, and vice-versa). All results are evaluated based on the U-Net [41] architecture. Bold text indicates the best.

Method	Retinal Fundus Segmentation								MRI Prostate Segmentation							
	Sequence 1 (A B C → D E F)				Sequence 2 (D E F → A B C)				Sequence 3 (A B C → D E F)				Sequence 4 (D E F → A B C)			
	Time/Client →			Mean	Time/Client →			Mean	Time/Client →			Mean	Time/Client →			Mean
	Site D	Site E	Site F		Site A	Site B	Site C		Site D	Site E	Site F		Site A	Site B	Site C	
Source	69.73	70.28	69.62	69.88	69.35	70.07	68.94	69.45	69.86	70.41	69.75	70.01	69.42	70.14	68.98	69.51
Local	70.32	70.78	70.15	70.42	69.83	70.47	69.38	69.89	70.27	70.73	70.08	70.36	69.76	70.41	69.25	69.81
FedAvg	71.05	70.20	71.50	70.92	73.04	55.22	75.05	67.77	68.03	72.51	69.97	70.17	65.11	73.02	75.78	71.30
FedProx	71.20	70.85	72.01	71.35	75.01	60.10	75.49	70.20	70.08	73.04	69.79	70.97	68.23	73.14	75.53	72.30
FedBN	72.50	71.90	72.13	72.18	74.78	68.19	**76.12**	73.03	72.54	72.01	67.03	70.53	68.01	72.98	**75.82**	72.27
FedGA	72.10	71.55	73.00	72.22	74.20	66.94	74.92	72.02	66.26	69.93	68.61	68.27	70.84	72.86	74.91	72.87
Tent	71.45	71.32	71.03	71.27	70.94	70.81	70.55	70.77	71.38	71.25	71.01	71.21	70.87	70.69	70.42	70.66
CoTTA	72.83	72.51	72.04	72.46	72.32	72.15	71.83	72.10	72.76	72.59	72.17	72.51	72.25	72.07	71.81	72.04
BeCoTTA	73.27	73.54	73.13	73.31	72.76	73.09	72.62	72.82	73.21	73.47	73.06	73.25	72.69	73.12	72.55	72.79
TCA	73.72	74.01	73.82	73.85	73.21	73.54	73.07	73.27	73.67	73.95	73.56	73.73	73.14	73.67	73.00	73.27
FedTHE+	74.63	75.27	74.59	74.83	74.12	74.95	74.00	74.36	74.58	75.22	74.40	74.73	71.05	74.88	73.93	73.29
ATP	74.16	74.48	74.11	74.25	73.65	74.08	73.52	73.75	74.12	74.43	74.05	74.20	73.58	74.12	73.45	73.72
FedCTTA	74.73	75.82	74.72	75.09	75.02	74.37	73.36	74.25	75.74	74.66	75.61	75.34	74.74	75.04	74.98	74.92
TTA-FedDG	**75.08**	75.75	75.13	75.32	74.58	75.42	74.48	74.83	75.04	**75.91**	75.07	75.34	74.51	75.34	74.41	74.75
BPFedCTTA	74.90	**76.16**	**75.50**	**75.52**	**75.50**	**75.67**	75.03	**75.11**	**76.01**	74.50	**75.90**	**75.47**	**74.90**	**75.60**	74.68	**75.06**

use the natural domain split, where each medical center is a client, preserving real-world domain shifts.For our BPFedCTTA, the client-side BPA uses an Adam optimizer with a learning rate of $\eta = 1e-4$. The server-side UGSA learning rate is $\gamma = 0.5$. All experiments were implemented in PyTorch and run on 2 NVIDIA 3090 GPUs.

Evaluation Metrics. We evaluate classification performance using standard accuracy. For segmentation tasks, we use Mean Intersection over Union (mIoU, %) for natural images and the Dice score (%) for medical images.

4.3. Experimental Results

Results on Image Classification Benchmarks. As shown in Table 1, our BPFedCTTA achieves state-of-the-art average accuracy on both CIFAR10-C (68.44%) and CIFAR100-C (67.58%), outperforming all baselines. FedAvg provides a strong baseline (65.49%, 59.50%) but lacks specialized adaptation. Personalized FL methods such as FedBN (64.34%, 57.91%) perform poorly as they require labeled data. Standard TTA methods (Tent: 65.41%, CoTTA: 66.28%) show significant performance degradation in sequential settings (e.g., CoTTA drops from 58.60% to 56.99% in Table 2), highlighting their instability, while continual TTA methods (BeCoTTA: 67.36%, TCA: 66.92%) narrow the gap but still trail our approach. Notably, BPFedCTTA surpasses federated TTA competitors FedTHE+ (67.93%) and TTA-FedDG (68.09%) on CIFAR10-C, demonstrating robustness.

Results on Natural Image Semantic Segmentation. Table 2 shows BPFedCTTA achieves the highest average mIoU on Cityscapes-to-ACDC, with progressive improvement across sequences (Seq1: 62.88%, Seq4: 65.19%), indicating effective continual learning. In contrast, the source model remains static (56.73%), while FedAvg (54.30%→53.51%) and Local fine-tuning (53.12%→52.57%) degrade over time. Tent collapses from 57.06% to 52.72%, and CoTTA drops from 58.60% to 56.99%, highlighting their instability. The progressive improvement in mIoU across sequences (from 62.88% to 65.19%) confirms that UGSA successfully integrates reliable knowledge from sequential clients.

Results on Medical Image Segmentation. On medical benchmarks (Table 3), BPFedCTTA delivers robust performance, achieving top-2 mean Dice scores in 3 of 4 sequences.

Table 4. Comprehensive ablation study of BPFedCTTA components across all benchmarks. Bold text indicates the best.

Method	CIFAR		Cityscapes to ACDC				Retinal Fundus		Prostate MRI	
	CIFAR10-C	CIFAR100-C	Seq1	Seq2	Seq3	Seq4	Seq1	Seq2	Seq3	Seq4
FedAvg	65.49	59.50	54.30	54.04	53.77	53.51	70.92	67.77	70.17	71.30
+ Tent	64.72	57.51	57.06	55.49	54.15	52.72	71.27	70.77	70.66	70.46
+ BPA	67.15	62.89	59.82	59.24	58.71	58.30	73.31	72.82	72.79	72.46
+ Tent + UGSA	66.83	61.45	58.53	58.98	58.46	58.12	72.46	72.10	72.04	71.85
BPFedCTTA (Ours)	**67.74**	**64.57**	**62.88**	**63.96**	**64.68**	**65.19**	**75.52**	**75.11**	**75.47**	**75.06**

Table 5. Comparison of uncertainty estimation methods for UGSA across multiple benchmarks. We report accuracy (%) on CIFAR10-C and CIFAR100-C datasets, computational time per client (seconds), and forgetting measure ($\Delta\%$, lower is better).

Uncertainty Method	CIFAR Accuracy		Time/Client (s)	Forgetting ($\Delta\%$)
	CIFAR10-C	CIFAR100-C		
Predictive Entropy	68.44	64.57	**4.492**	4.21
Max Probability	68.10	63.95	6.112	4.58
Energy Score	68.62	64.89	8.834	3.95
Prediction Consistency	68.79	65.12	19.457	3.82
MC Dropout ($T = 10$)	**68.75**	**64.98**	64.920	**3.45**
Ensemble ($M = 3$)	68.65	64.85	20.125	3.54

For Retinal Fundus, it leads Sequence 1 (75.52%) over TTA-FedDG (75.32%) and FedTHE+ (74.83%), and remains highly competitive in Sequence 2 (75.11%). On Prostate MRI, it achieves the highest score in Sequence 4 (75.06%). The consistent performance across sequences (75.52%, 75.11%, 75.47%, 75.06%) demonstrates greater stability compared to baselines such as FedGA (72.22%→68.27%). These results validate our framework's generalizability in privacy-sensitive medical domains.

4.4. Further Empirical Analysis

Effectiveness of BPA and UGSA. Ablation results in Table 4 confirm the necessity of both components. FedAvg performs poorly. While Tent improves classification, it fails catastrophically in sequential scenarios (dropping from 57.06% to 52.72%). BPA alone substantially boosts performance over Tent (*e.g.*, +2.43% on CIFAR10-C, +5.38% on CIFAR100-C) and improves stability. The variant + Tent + UGSA shows that aggregation alone mitigates forgetting but is less effective. The full BPFedCTTA achieves the best performance, with progressive improvement on Cityscapes-to-ACDC (62.88% → 65.19%), confirming BPA and UGSA are complementary: BPA ensures stable local adaptation, while UGSA enables safe global consolidation.

Effectiveness of uncertainty estimation methods. As shown in Tab. 5, Entropy provides the best speed-accuracy trade-off, while Energy Score offers improved accuracy and reduced forgetting with minimal computational overhead. Multi-view methods (Prediction Consistency, MC Dropout, Ensemble) achieve the best accuracy and stability at significantly higher computational cost.

Effectiveness over heterogeneity of source data. As shown in Fig. 3, BPFedCTTA maintains superior performance across all Non-IID levels, achieving 64.5% and 62.5% under extreme heterogeneity ($\alpha = 0.05$), 1.4% and 1.8% higher than the strongest baseline. Performance degradation from

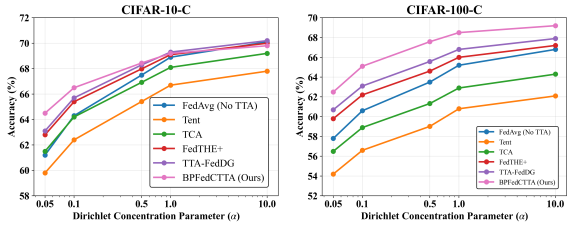


Figure 3. Performance comparison of different methods under varying degrees of data heterogeneity (Non-IID).

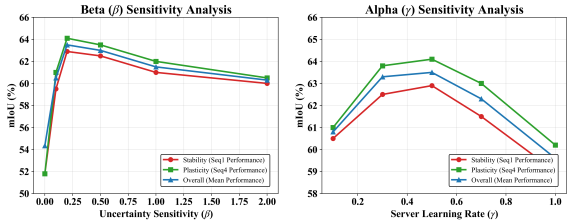


Figure 4. Hyperparameters (γ and β) sensitivity analysis of the UGSA mechanism on the CitySpace to ACDC benchmark.

$\alpha = 10.0$ to $\alpha = 0.05$ is only 5.3% and 6.7%, lower than FedAvg's 9.0% and 9.3%, demonstrating robustness in heterogeneous federated scenarios.

Effectiveness of γ and β in UGSA module. Sensitivity analysis (Fig. 4) shows β is crucial: at $\beta = 0.0$ (UGSA off), Stability collapses to 51.8%. Stability peaks at $\beta = 0.2$ (62.9%), with overall performance also optimal (63.5% mIoU), balancing Stability and Plasticity. With $\beta = 0.2$, $\gamma = 0.5$ maximizes overall performance (63.5% mIoU), while lower γ (0.1) suppresses Plasticity (61.0%) and higher γ (1.0) reduces Stability (59.0%).

5. Conclusion

This paper tackles the challenging Federated Continual Test-Time Adaptation (FedCTTA) scenario, where unlabeled clients with heterogeneous data arrive sequentially. We identify two core challenges: unstable local adaptation and destructive global aggregation. To address these, we propose BPFedCTTA, a Bayesian framework that unifies local adaptation and global evolution via two key components: (1) Bayesian Prior-guided Adaptation (BPA), which stabilizes local updates using the global model as a prior, and (2) Uncertainty-Gated Single-client Aggregation (UGSA), which selectively integrates reliable client updates to mitigate catastrophic forgetting. Extensive experiments demonstrate that BPFedCTTA achieves superior and more stable performance, outperforming existing methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62376003), Anhui Provincial Natural Science Foundation, China (Grant No. 2308085MF200), and the Anhui Provincial Key Research and Development Program under Grant NO.202304a05020047.

References

[1] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. 2

[2] Wenxuan Bao, Tianxin Wei, Haohan Wang, and Jingrui He. Adaptive test-time personalization for federated learning. *Advances in neural information processing systems*, 36:77882–77914, 2023. 1, 3, 6

[3] Wenxuan Bao, Ruxi Deng, Ruizhong Qiu, Tianxin Wei, Hanghang Tong, and Jingrui He. Latte: Collaborative test-time adaptation of vision-language models in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 133–143, 2025. 1

[4] Debora Caldarola, Pietro Cagnasso, Barbara Caputo, and Marco Ciccone. Beyond local sharpness: Communication-efficient global sharpness-aware minimization for federated learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25187–25197, 2025. 2

[5] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019. 3

[6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 3

[7] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022. 2

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[9] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023. 3

[10] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *European conference on computer vision*, pages 200–216. Springer, 2020. 3

[11] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020. 2

[12] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. RIM-ONE: An open retinal image database for optic nerve evaluation. In *Computers in Biology and Medicine*, pages 1–6. IEEE, 2011. 6

[13] Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 3

[16] Philip TG Jackson, Amir Atapour Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: data augmentation via style randomization. In *CVPR workshops*, pages 10–11, 2019. 3

[17] Liangze Jiang and Tao Lin. Test-time robust personalization for federated learning. *arXiv preprint arXiv:2205.10920*, 2022. 1, 3, 6

[18] Sanghun Jung, Jungsoo Lee, Nanhee Kim, Amirreza Shaban, Byron Boots, and Jaegul Choo. Cafa: Class-aware feature alignment for test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19060–19071, 2023. 3

[19] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 3

[20] Hiroaki Kingetsu, Kenichi Kobayashi, Yoshihiro Okawa, Yasuto Yokota, and Katsuhito Nakazawa. Multi-step test-time adaptation with entropy minimization and pseudo-labeling. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4153–4157. IEEE, 2022. 3

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[22] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4)*, pages 794–797. IEEE, 2020. 1

[23] Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. Becotta: Input-dependent online blending of experts for continual test-time adaptation. In *International Conference on Machine Learning*, pages 27072–27093. PMLR, 2024. 6

[24] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 6

[25] Haidong Li, Jiongcheng Li, Xiaoming Guan, Binghao Liang, Yuting Lai, and Xinglong Luo. Research on overfitting of deep learning. In *2019 15th international conference on computational intelligence and security (CIS)*, pages 78–81. IEEE, 2019. 3

[26] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 2, 6

[27] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021. 2

[28] Haoyuan Liang, Xinyu Zhang, Shilei Cao, Guowen Li, and Juepeng Zheng. Tta-feddg: Leveraging test-time adaptation to address federated domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18658–18666, 2025. 1, 3, 6

[29] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020. 2

[30] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 6

[31] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1023, 2021. 1, 3

[32] Xingguo Lv, Xingbo Dong, Liwen Wang, Jiewen Yang, Lei Zhao, Bin Pu, Zhe Jin, and Xuejun Li. Test-time domain generalization via universe learning: A multi-graph matching approach for medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15621–15631, 2025. 3

[33] Jing Ma. Improved self-training for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23701–23710, 2024. 3

[34] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 3, 4, 6

[35] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013. 3

[36] Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, and Tao Zhou. Maintaining consistent inter-class topology in continual test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15319–15328, 2025. 6

[37] Bloch Nicholas, Madabhushi Anant, Huisman Henkjan, Freymann John, Kirby Justin, et al. Nci-proc. ieee-isbi conf. 2013 challenge: Automated segmentation of prostate structures. *The Cancer Imaging Archive*, 5, 2015. 6

[38] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3

[39] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020. 6

[40] Rakibul Hasan Rajib, Md Akil Raihan Iftee, Mir Sazzat Hossain, AKM Rahman, Sajib Mistry, M Ashraful Amin, and Amin Ahsan Ali. Fedctta: A collaborative approach to continual test-time adaptation in federated learning. *arXiv preprint arXiv:2505.13643*, 2025. 1, 2, 4, 6

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6, 7

[42] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10765–10775, 2021. 5

[43] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-GS: Retinal image dataset for optic nerve head (onh) segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 53–56. IEEE, 2014. 6

[44] Yuwei Sun, Ng Chong, and Hideya Ochiai. Feature distribution matching for federated domain generalization. In *Asian conference on machine learning*, pages 942–957. PMLR, 2023. 3

[45] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020. 1, 2

[46] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8432–8440, 2022. 2

[47] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2020. 3, 6

[48] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1, 6

[49] Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20050–20060, 2023. 3

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers.

Advances in neural information processing systems, 34:12077–12090, 2021. 6, 7

[51] Yicheng Xu, Wubin Ma, Chaofan Dai, Yahui Wu, and Haohao Zhou. Generalized federated learning via gradient norm-aware minimization and control variables. *Mathematics*, 12 (17):2644, 2024. 2

[52] Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized federated learning with inferred collaboration graphs. In *International conference on machine learning*, pages 39801–39817. PMLR, 2023. 6

[53] Yanyu Ye, Wei Wei, Lei Zhang, Chen Ding, and Yanning Zhang. Domain consistency learning for continual test-time adaptation in image semantic segmentation. *Pattern Recognition*, 165:111585, 2025. 3

[54] Linjing You, Jiabao Lu, and Xiayuan Huang. Probabilistic contrastive test-time adaptation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3

[55] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023. 3

[56] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3954–3963, 2023. 1, 3, 6

[57] Siqi Zhang, Lu Zhang, and Zhiyong Liu. Test-time adaptation for object detection via dynamic dual teaching. *Image and Vision Computing*, page 105740, 2025. 3

[58] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. In *Engineering in Medicine and Biology Society*, pages 3065–3068. IEEE, 2010. 6