# FedPoisonTTP: A Threat Model and Poisoning Attack for Federated Test-Time Personalization

Md Akil Raihan Iftee[1,*], Syed Md. Ahnaf Hasan[1], Amin Ahsan Ali[1], A K M Mahbubur Rahman[1], Sajib Mistry[2], Aneesh Krishna[2]
[1]Center for Computational & Data Sciences (CCDS), Independent University, Bangladesh
[2]Curtin University, Australia

## A. White-Box and Grey-Box Attack Settings

### A.1. White-Box Attacks

In a white-box scenario [7], [8], [9], the adversary possesses full access to the online model parameters $\theta_t$, the benign users' test samples $B_b$, and the gradients used during adaptation. This allows direct optimization of poisoned inputs against the actual online TTA dynamics:

$$\min_{B_a} \mathbb{E}_{(x,y)\in B_b} \left[ L_{\text{atk}}(h(x; \theta_t^*(B_a \cup B_b)), y) \right], \quad (1)$$

where $\theta_t^*$ denotes the updated model after TTA on both poisoned and benign samples. Although analytically convenient, this setting is unrealistic in decentralized or federated deployments where neither benign samples nor online parameters are exposed to adversaries.

### A.2. Grey-Box Attacks

In a grey-box setting [1], [3], [5], adversaries cannot access benign users' samples or observe online model parameters $\theta_t$. Instead, they rely on a surrogate (distilled) model $\hat{\theta}_t$ that approximates the online model's behavior. The attack objective becomes:

$$\begin{aligned} \min_{B_a} \quad & \frac{1}{|B_{ab}|} \sum_{x_i \in B_{ab}} L_{\text{atk}}\left(x_i; \hat{\theta}_t'(B_t)\right), \\ \text{s.t.} \quad & B_t = B_a \cup B_{ab}. \end{aligned} \quad (2)$$

where $B_{ab}$ denotes the adversary's own clean samples. Forwarding $B_t$ jointly can degrade the adaptation process. Here, feature-level distribution consistency is also enforced so that poisoned samples remain statistically aligned with benign ones during adaptation.

## B. Various Attack Objectives

Several attack objectives can be used to generate adversarial samples during test-time adaptation. These objectives guide gradient-based procedures such as PGD to create high-entropy disturbances, low-entropy misclassifications, or feature-consistent perturbations that influence the model's adaptation behavior.

### B.1. Notch High-Entropy (NHE) Objective

NHE [6] constructs a target distribution $Q$ that assigns zero probability to the correct class and distributes uniform probability across all other classes. Minimizing the cross-entropy with respect to this target forces predictions away from the ground-truth and increases output entropy:

$$\begin{aligned} L_{\text{atk}}^{\text{NHE}}(\tilde{x}_i) &= -\sum_k Q_{ik} \log h_k(\tilde{x}_i), \\ \text{where} \quad Q_{ik} &= \begin{cases} 0, & k = y_i, \\ \frac{1}{K-1}, & k \neq y_i. \end{cases} \end{aligned} \quad (3)$$

This design produces high-entropy, label-divergent predictions that strongly influence entropy-based TTA updates.

### B.2. Balanced Low-Entropy (BLE) Objective

BLE [6] addresses the class-collapse problem common in low-entropy attacks. It uses a moving-average confusion matrix $C$ and a label-mapping matrix $M \in \{0,1\}^{K\times K}$ that assigns each class a distinct target class. The objective encourages confident but class-balanced incorrect predictions:

$$L_{\text{atk}}^{\text{BLE}}(\tilde{x}_i) = -\sum_k \mathbf{1}\left(k = \arg\max_{q \neq y_i} M_{y_i,q}\right) \log h_k(\tilde{x}_i). \quad (4)$$

By enforcing balanced misclassification, BLE avoids collapse into a single dominant wrong class and maintains stable target selection across all classes.

### B.3. MaxCE Objective

MaxCE objective [4] maximizes the standard cross-entropy loss with respect to the true label:

$$L^{\text{MaxCE}}(\tilde{x}_i) = -\log h_{y_i}(\tilde{x}_i). \quad (5)$$

*Corresponding author: iftee1807002@gmail.com

This objective drives the model's prediction to reduce confidence in the correct class by pushing up the loss on that class. In practice, a projected gradient method (e.g., PGD) is used to find perturbations $\delta$ within a constrained norm ball that maximize this cross-entropy loss, matching the attack strategy in [4] for adversarial robustness.

### B.4. TePA Entropy-Maximization Objective

TePA [2] generates adversarial samples by maximizing prediction entropy:

$$L^{\text{TePA}}(\tilde{x}_i) = -\sum_k h_k(\tilde{x}_i) \log h_k(\tilde{x}_i). \tag{6}$$

This objective encourages uniform output distributions, which destabilize entropy-based or confidence-based TTA updates by reducing prediction certainty.

### B.5. Distribution Regularization (Feature Consistency)

Feature-level distribution regularization aligns the feature statistics of poisoned data with those of benign data. Gaussian distributions are fitted at each selected feature layer using mean–covariance pairs $(\mu_i^l, \Sigma_i^l)$ for benign samples and $(\tilde{\mu}_i^l, \tilde{\Sigma}_i^l)$ for poisoned samples:

$$L_{\text{reg}} = \frac{1}{L} \sum_{l=1}^{L} \text{KLD}\Big(\mathcal{N}(\mu_i^l, \Sigma_i^l) \parallel \mathcal{N}(\tilde{\mu}_i^l, \tilde{\Sigma}_i^l)\Big). \tag{7}$$

This alignment ensures that poisoned samples remain in-distribution, allowing their influence to transfer to unseen benign samples during adaptation.

### B.6. DIA (Distribution Invading Attack)

In DIA [9], poisoning is formulated as a bilevel optimization problem in which poisoned samples are crafted (outer problem) so that, after test-time adaptation (inner problem) on the mixed batch, the model incurs large loss on the benign samples. The bilevel formulation is:

$$\begin{aligned} \min_{B_a} \quad & \mathbb{E}_{(x,y)\in B_b}\big[L_{\text{atk}}\big(h(x;\theta^*(B_a \cup B_b)), y\big)\big], \\ \text{s.t.} \quad & \theta^*(B_a \cup B_b) = \arg\min_{\theta} L_{\text{TTA}}\big(B_a \cup B_b; \theta\big), \end{aligned} \tag{8}$$

where $B_a$ denotes the poisoned subset and $B_b$ the benign subset in the same test batch. In practice the inner adaptation is approximated and the bilevel problem is converted into a single-level objective that is solved by projected gradient methods; a common approximation replaces $\theta^*$ with the current model parameters (or a surrogate) and optimizes:

$$\min_{B_a} \frac{1}{|B_b|} \sum_{(x,y)\in B_b} L_{\text{atk}}\big(h(x;\theta_{\text{approx}}(B_a)), y\big), \tag{9}$$

with constrained optimization (e.g., PGD) used to enforce perturbation bounds. DIA supports both targeted (flip a chosen sample to a target label) and indiscriminate (degrade overall performance on benign data) variants, and uses projected gradient steps to synthesize the malicious examples.

## References


[1] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017. 1

[2] Tianshuo Cong, Xinlei He, Yun Shen, and Yang Zhang. Test-time poisoning attacks against test-time adaptation models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1306–1324. IEEE, 2024. 2

[3] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018. 1

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2

[5] Binxin Ru, Adam Cobb, Arno Blaas, and Yarin Gal. Bayesopt adversarial attack. In *International conference on learning representations*, 2019. 1

[6] Yongyi Su, Yushu Li, Nanqing Liu, Kui Jia, Xulei Yang, Chuan-Sheng Foo, and Xun Xu. On the adversarial risk of test time adaptation: An investigation into realistic test-time data poisoning. *arXiv preprint arXiv:2410.04682*, 2024. 1

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[8] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 1

[9] Tong Wu, Feiran Jia, Shaokun Zhang, and et al. Uncovering adversarial risks of test-time adaptation. In *Proceedings of Machine Learning Research (PMLR)*, 2023. 1, 2