# Persistent Test-time Adaptation in Recurring Testing Scenarios

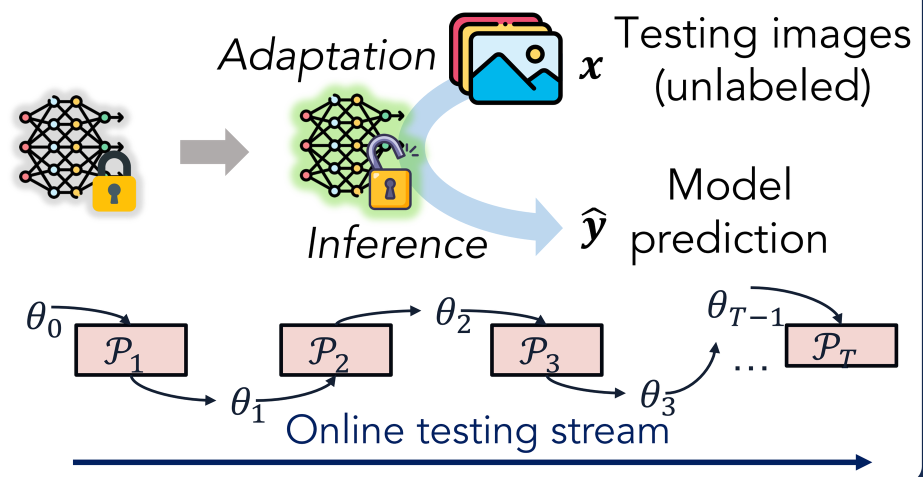
Trung-Hieu Hoang [1], Duc Minh Vo [2], Minh N. Do [1]

[1] Department of Electrical & Computer Engineering, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, USA [2] The University of Tokyo, Japan

## INTRODUCTION

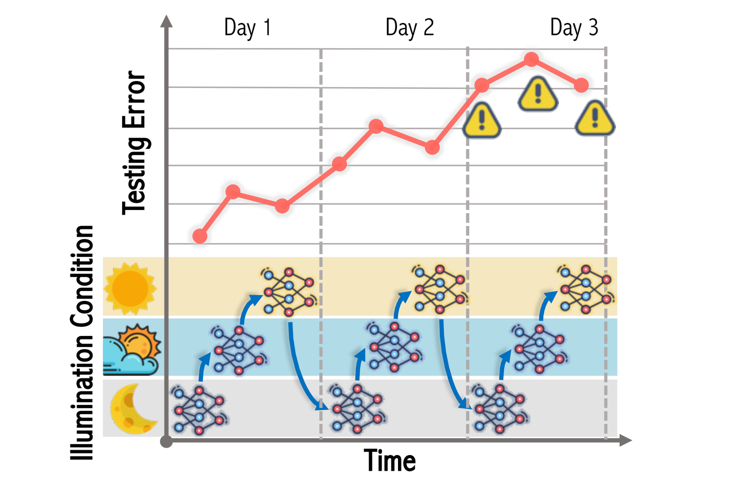
Continual Test-time Adaptation (TTA) operates on an ML classifier $f_t: \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta_t \in \Theta$ gradually changing over time. The model explores an online stream of testing data $X_t \sim P_t$ for adapting itself $f_{t-1} \to f_t$ (self-supervised learning) before predicting $\hat{Y}_t = f_t(X_t)$.

### Continual Test-time Adaptation Scheme



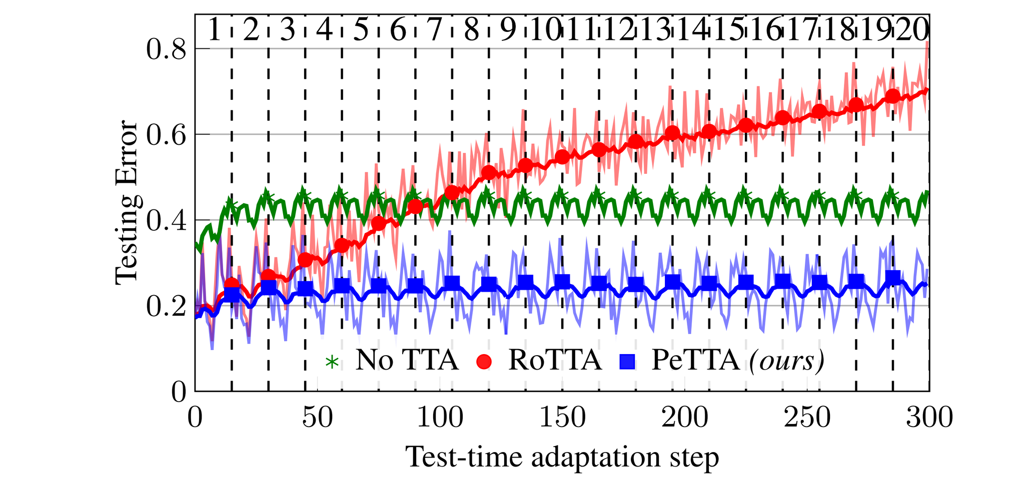

Does the model adaptability persist after a long time adapting to multiple data shifts?

### Hypothetical Setting




- In practice, testing environments may *change recurringly*.
- Preserving adaptability when visiting the same testing condition is *not guaranteed*.

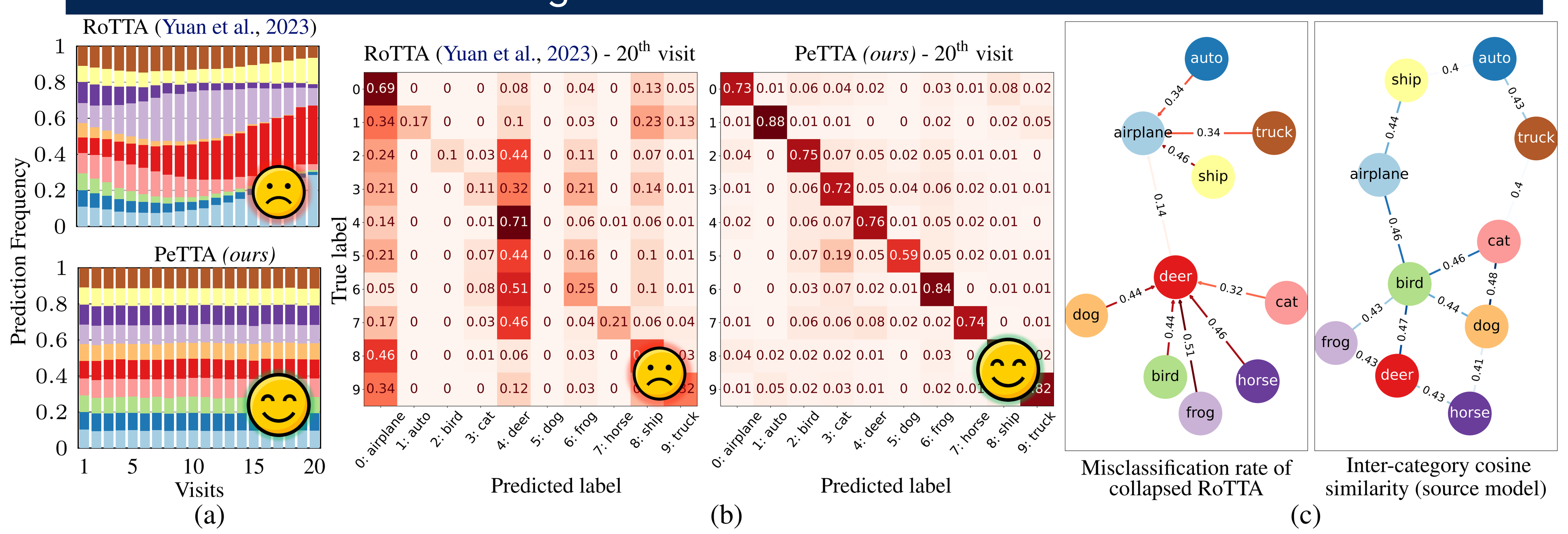
### Empirical Experiment on CIFAR-10-C




- Testing error of RoTTA (Yuan, 2023), a baseline TTA algorithm raises - *performance degradation.*
- Quickly exceeding the error of the source model (without TTA, accepting domain shift as-it-is).
- PeTTA *(ours)* demonstrates its stability.

Recurring Test−time Adaptation: $\mathcal{P}_1 \to \mathcal{P}_2 \to \cdots \to \mathcal{P}_D \to \cdots \to \mathcal{P}_1 \to \mathcal{P}_2 \to \cdots \to \mathcal{P}_D$

### TTA Under Recurring TTA on CIFAR-10 → CIFAR-10-C Task for 20 Visits




(a) Histogram of model PeTTA achieves a persisting performance while RoTTA degrades.
(b) Confusion matrix at the last visit (c) Force-directed graph showing (left) the most prone to misclassification; (right) similar categories tend to be easily collapsed.

## $\epsilon$-PERTURBED GAUSSIAN MIXTURE MODEL CLASSIFIER ($\epsilon$-GMMC)

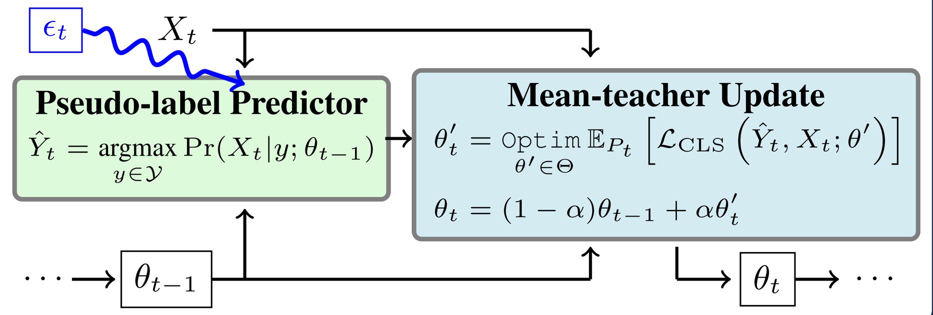
*$\epsilon$-GMMC - a simple yet representative **failure case** of TTA for **theoretical analysis***

**Setting:** A simplified continual TTA process

- Let $p_{y,t} = \Pr(Y_t = y)$; $\hat{p}_{y,t} = \Pr(\hat{Y}_t = y)$.
- Binary classification $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \{0,1\}$.
- Underlying distribution follows a mixture of 2 Gaussian: $P_t(x,y) = p_{y,t}\mathcal{N}(x; \mu_y, \sigma_y^2)$.

**Main Task:** predicting $X_t$ was sampled from cluster 0 or 1 (negative or positive).

### $\epsilon$-GMMC




**ε−GMMC** performs 2 main steps:

- *Predicting pseudo-labels ($\hat{Y}_t$).*
- *Updating with mean teacher model.*

**Key Idea:** The *predictor is* **perturbed** for retaining a **false negative rate (FNR)** of $\varepsilon_t = \Pr\{Y_t = 1 | \hat{Y}_t = 0\}$ to simulate undesirable effects of the testing stream in TTA, making model prone to collapse.

### A Mathematical Definition of Model Collapse

**Definition 1 (Model Collapse).** *A model is said to be collapsed from step $\tau \in \mathcal{T}, \tau < \infty$ if there exists a non-empty subset of categories $\tilde{\mathcal{Y}} \subset \mathcal{Y}$ such that $\Pr\{Y_t \in \tilde{\mathcal{Y}}\} > 0$ but the marginal $\Pr\{\hat{Y}_t \in \tilde{\mathcal{Y}}\}$ converges to zero in probability:*

$$\lim_{t \to \tau} \Pr\{\hat{Y}_t \in \tilde{\mathcal{Y}}\} = 0.$$

*Factors contributing to the model collapse:*

*(i)* ***Data-dependent factors***: the prior data distribution ($p_0$), the nature difference between two categories ($|\mu_0 - \mu_1|$) from the dataset.

*(ii)* ***Algorithm-dependent factors***: update rate ($\alpha$), the false negative rate at each step ($\varepsilon_t$).

**Assumption 1 (Static Data Stream).** *The marginal distribution of the true label follows the same Bernoulli distribution $\mathrm{Ber}(p_0)$: $p_{0,t} = p_0$, $(p_{1,t} = p_1 = 1 - p_0), \forall t \in \mathcal{T}$.*

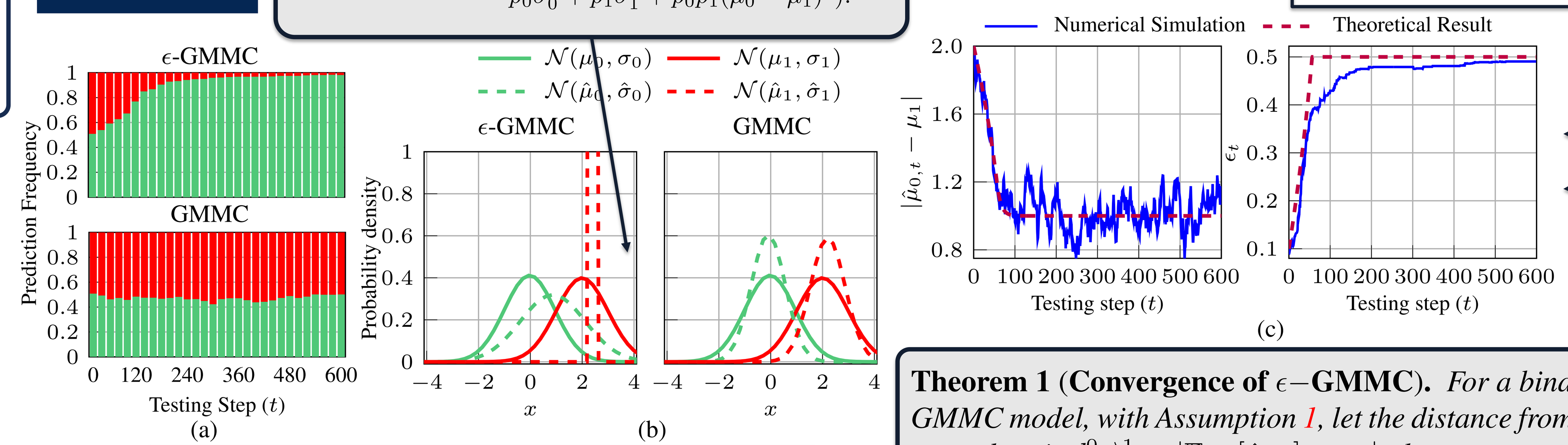
**Lemma 2 ($\epsilon$-GMMC After Collapsing).** *For a binary $\epsilon$-GMMC model, with Assumption 1, if $\lim_{t \to \tau} \hat{p}_{1,t} = 0$ (collapsing), the cluster 0 in GMMC converges in distribution to a single-cluster GMMC with parameters:*

$$\mathcal{N}(\hat{\mu}_{0,t}, \hat{\sigma}_{0,t}^2) \xrightarrow{d.} \mathcal{N}(p_0\mu_0 + p_1\mu_1, p_0\sigma_0^2 + p_1\sigma_1^2 + p_0p_1(\mu_0 - \mu_1)^2).$$

**Corollary 1 (A Condition for $\epsilon-$GMMC Collapse).** *With fixed $p_0, \alpha, \mu_0, \mu_1$, $\epsilon-$GMMC is collapsed if there exists a sequence of $\{\epsilon_t\}_{\tau - \Delta_\tau}^{\tau}$ $(\tau \geq \Delta_\tau > 0)$ such that:*

$$p_1 \geq \epsilon_t > 1 - \frac{d_{t-1}^{0\to1}}{|\mu_0 - \mu_1|}, \quad t \in [\tau - \Delta_\tau, \tau].$$

### ε−GMMC Simulation Result




**Lemma 1 (Increasing FNR).** *Under Assumption 1, a binary $\epsilon$-GMMC would collapsed (Def. 1) with $\lim_{t \to \tau} \hat{p}_{1,t} = 0$ (or $\lim_{t \to \tau} \hat{p}_{0,t} = 1$, equivalently) if and only if $\lim_{t \to \tau} \epsilon_t = p_1$.*

**Theorem 1 (Convergence of $\epsilon-$GMMC).** *For a binary $\epsilon$-GMMC model, with Assumption 1, let the distance from $\hat{\mu}_{0,t}$ toward $\mu_1$ is $d_t^{0\to1} = |\mathbb{E}_{P_t}[\hat{\mu}_{0,t}] - \mu_1|$, then:*

$$d_t^{0\to1} - d_{t-1}^{0\to1} \leq \alpha \cdot p_0 \cdot \left(|\mu_0 - \mu_1| - \frac{d_{t-1}^{0\to1}}{1 - \epsilon_t}\right).$$

(a) Histogram of model predictions. (b) The probability density function of the two clusters after convergence (*dashed line*) versus the true data distribution. (c) Distance toward $\mu_1$ and false-negative rate ($\varepsilon_t$)coincides with the theoretical analysis.

## PERSISTENT TEST-TIME ADAPTATION (PeTTA)

**Key Idea:** *Striking a balance between **adaptation and preventing model collapse***

With $\phi_{\theta_t}$ is the deep feature extractor of $f_t$, let $\boldsymbol{z} = \phi_{\theta_t}(\boldsymbol{x})$. Keeping track of a collection of the running mean of feature vector $\boldsymbol{z}$: $\{\hat{\mu}_t^y\}_{y \in \mathcal{Y}}$ in which $\hat{\boldsymbol{\mu}}_t^y$ is exponential moving average updated with vector $\boldsymbol{z}$ if $f_t(\boldsymbol{x}) = y$.

### Persistent TTA

**(1) Sensing the divergence from $\boldsymbol{\theta}_0$**

$$\gamma_t^y = 1 - \exp\left(-(\hat{\boldsymbol{\mu}}_t^y - \boldsymbol{\mu}_0^y)^T (\boldsymbol{\Sigma}_0^y)^{-1} (\hat{\boldsymbol{\mu}}_t^y - \boldsymbol{\mu}_0^y)\right)$$

$\boldsymbol{\mu}_0^t$, $\boldsymbol{\Sigma}_0^t$ are pre-computed on the source distribution

**(2) Adaptive Learning Rate $\boldsymbol{\alpha}_t$ and Regularization $\boldsymbol{\lambda}_t$**

$$\bar{\gamma}_t = \frac{1}{|\hat{\mathcal{Y}}_t|} \sum_{y \in \hat{\mathcal{Y}}_t} \gamma_t^y, \quad \hat{\mathcal{Y}}_t = \left\{\hat{Y}_t^{(i)} | i = 1, \cdots, N_t\right\}$$

$$\lambda_t = \bar{\gamma}_t \cdot \lambda_0, \qquad \alpha_t = (1 - \bar{\gamma}_t) \cdot \alpha_0,$$

**PeTTA**

$$\theta'_t = \operatorname{Optim}_{\theta' \in \Theta} \mathbb{E}_{P_t}\left[\mathcal{L}_{\mathrm{CLS}}\left(\hat{Y}_t, X_t; \theta'\right) + \mathcal{L}_{\mathrm{AL}}\left(X_t; \theta'\right)\right] + \lambda_t \mathcal{R}(\theta')$$

$$\theta_t = (1 - \alpha_t)\theta_{t-1} + \alpha_t\theta'_t.$$

**(3) Anchor Loss**

$$\mathcal{L}_{\mathrm{AL}}(X_t; \theta) = -\sum_{y \in \mathcal{Y}} \Pr(y | X_t; \theta_0) \log \Pr(y | X_t; \theta)$$

## EXPERIMENTAL RESULTS

Average classification error on the task *ImageNet* → *ImageNet-C* for 20 recurring TTA visits.

| Method | *Recurring TTA visit →* 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | 82.0 | | | | | | | | | | | | | | | | | | | | 82.0 |
| LAME (Boudiaf et al., 2022) | 80.9 | | | | | | | | | | | | | | | | | | | | 80.9 |
| CoTTA (Wang et al., 2022) | 98.6 | 99.1 | 99.4 | 99.4 | 99.5 | 99.5 | 99.5 | 99.5 | 99.6 | 99.7 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.6 | 99.7 | 99.7 | 99.5 |
| EATA (Niu et al., 2022) | 60.4 | 59.3 | 65.4 | 72.6 | 79.1 | 84.2 | 88.7 | 92.7 | 95.2 | 96.9 | 97.7 | 98.1 | 98.4 | 98.6 | 98.7 | 98.8 | 98.8 | 98.9 | 98.9 | 99.0 | 89.0 |
| RMT (Döbler et al., 2022) | 72.3 | 71.0 | 69.9 | 69.1 | 68.8 | 68.5 | 68.4 | 68.3 | 70.0 | 70.2 | 70.1 | 70.2 | 72.8 | 76.8 | 75.6 | 75.1 | 75.1 | 75.2 | 74.8 | 74.7 | 71.8 |
| MECTA (Hong et al., 2023) | 77.2 | 82.8 | 86.1 | 87.9 | 88.9 | 89.4 | 89.8 | 89.9 | 90.0 | 90.4 | 90.6 | 90.7 | 90.7 | 90.8 | 90.8 | 90.9 | 90.8 | 90.8 | 90.7 | 90.8 | 89.0 |
| RoTTA (Yuan et al., 2023) | 68.3 | 62.1 | 61.8 | 64.5 | 68.4 | 75.4 | 82.7 | 95.1 | 95.8 | 96.6 | 97.1 | 97.9 | 98.3 | 98.7 | 99.0 | 99.1 | 99.3 | 99.4 | 99.5 | 99.6 | 87.9 |
| RDumb (Press et al., 2023) | 72.2 | 73.0 | 73.2 | 72.8 | 72.2 | 72.8 | 73.3 | 72.7 | 71.9 | 73.0 | 73.2 | 73.1 | 72.0 | 72.7 | 73.3 | 73.1 | 72.1 | 72.6 | 73.3 | 73.1 | 72.8 |
| ROID (Marsden et al., 2024) | 62.7 | 62.3 | 62.3 | 62.3 | 62.5 | 62.3 | 62.4 | 62.4 | 62.3 | 62.6 | 62.5 | 62.3 | 62.5 | 62.4 | 62.5 | 62.4 | 62.4 | 62.5 | 62.4 | 62.5 | 62.4 |
| TRIBE (Su et al., 2024) | **63.6** | 64.0 | 64.9 | 67.8 | 69.6 | 71.7 | 73.5 | 75.5 | 77.4 | 79.8 | 85.0 | 96.5 | 99.4 | 99.8 | 99.9 | 99.8 | 99.8 | 99.9 | 99.9 | 99.9 | 84.4 |
| PeTTA *(ours)*(*) | 65.3 | **61.7** | **59.8** | **59.1** | **59.4** | **59.6** | **59.8** | **59.3** | **59.4** | **60.0** | **60.3** | **61.0** | **60.7** | **60.4** | **60.6** | **60.7** | **60.8** | **60.7** | **60.4** | **60.2** | **60.5** |

*Does model reset help?* A comparison with a *reset-based approach* at different frequencies.

| Reset Every | *Recurring TTA visit →* 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $T = 1000$ | 72.2 | 73.0 | 73.2 | 72.8 | 72.2 | 72.8 | 73.3 | 72.7 | 71.9 | 73.0 | 73.2 | 73.1 | 72.0 | 72.7 | 73.3 | 73.1 | 72.1 | 72.6 | 73.3 | 73.1 | 72.8 |
| $T = 5000$ | 70.2 | 70.8 | 71.6 | 72.1 | 72.4 | 72.6 | 72.9 | 73.1 | 73.2 | 73.6 | 73.7 | 73.9 | 74.0 | 74.0 | 74.3 | 74.1 | 74.1 | 73.8 | 73.5 | 71.9 | 73.0 |
| $T = 75000$ | 67.0 | 67.1 | 67.2 | 67.5 | 67.5 | 67.6 | 67.8 | 67.6 | 67.6 | 67.6 | 67.5 | 67.7 | 67.6 | 67.9 | 68.1 | 67.9 | 67.4 | 67.5 | 67.7 | 67.5 | 67.6 |
| PeTTA *(ours)*(*) | **65.3** | **61.7** | **59.8** | **59.1** | **59.4** | **59.6** | **59.8** | **59.3** | **59.4** | **60.0** | **60.3** | **61.0** | **60.7** | **60.4** | **60.6** | **60.7** | **60.8** | **60.7** | **60.4** | **60.2** | **60.5** |

## CONTRIBUTIONS

✓ A new testing scenario – **recurring TTA**.
✓ Theoretical analysis on performance degradation of TTA on **$\epsilon$ −perturbed Gaussian Mixture Model Classifier**.
✓ A new baseline – **persistent TTA (PeTTA)**.

AUTHOR | PAPER | PROJECT