# Persistent Test-time Adaptation in Recurring Testing Scenarios

Trung-Hieu Hoang [1], Duc Minh Vo [2], Minh N. Do [1,3]

[1] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, USA

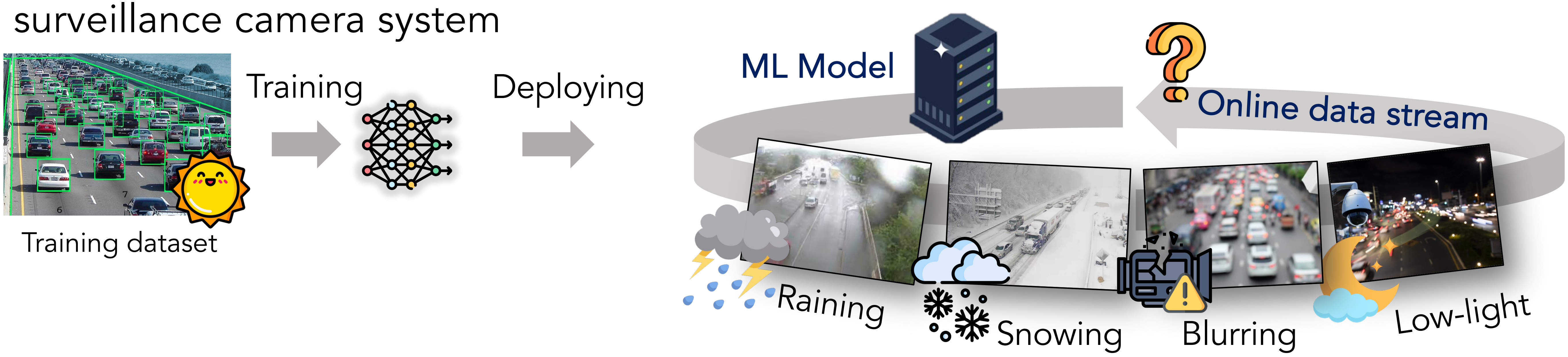
[2] The University of Tokyo, Japan

[3] VinUni-Illinois Smart Health Center, VinUniversity

**Best paper award** – Community Track, 1st Workshop on Test-Time Adaptation: Model, Adapt Thyself! (MAT), Conference on Computer Vision and Pattern Recognition (CVPR) Workshop 2024.

# Context: Motivation for Test-time Adaptation (TTA)

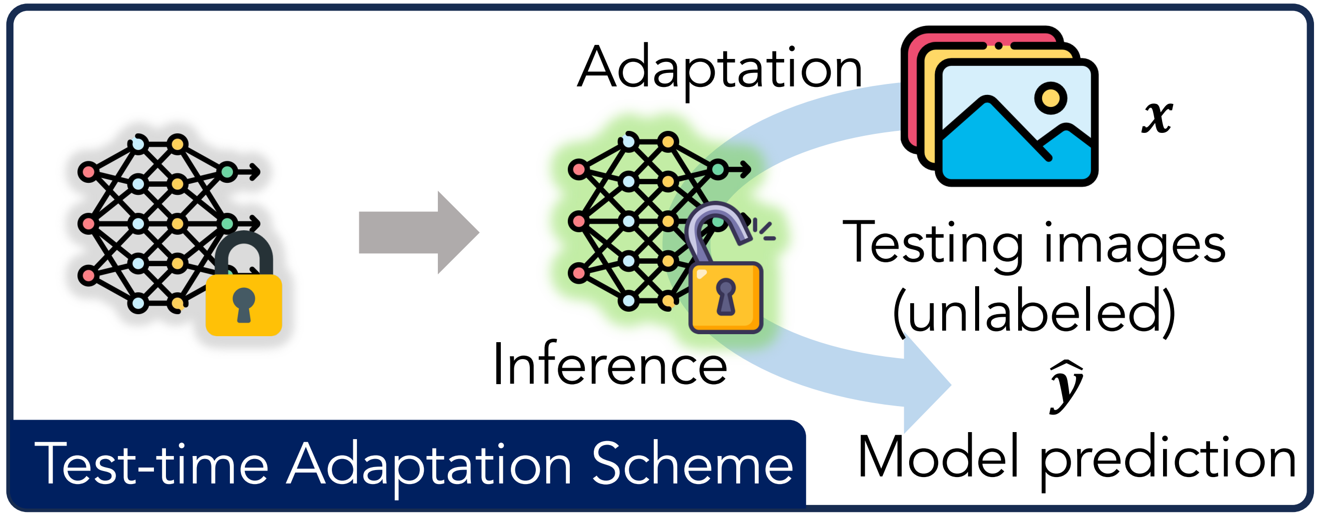
Let's look at a real-world scenario: deploying a machine learning (ML) model for traffic surveillance camera system




Unforeseen circumstances can introduce *domain-shift* and *severely reduce* ML model's performance at test-time

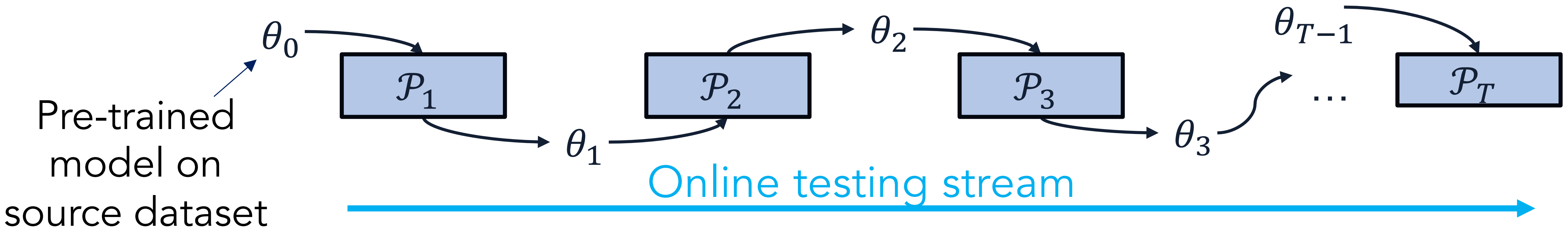
*How can we fix it?*

- Making the ML model *learnable* at *test time*
- Utilizing *unlabeled* data at test time for adaptation
- TTA has been showing many "good" results!



# Background: Test-time Adaptation for ML Model Deployment

**Test-time Adaptation (TTA):** TTA operates on an ML classifier $f_t\colon \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta_t \in \Theta$ *gradually changing* over time.




**Online testing stream:** The model explores an online stream of testing data $X_t \sim \mathcal{P}_t$ for adapting itself $f_{t-1} \to f_t$ (self-supervised learning) before predicting $\hat{Y}_t = f_t(X_t)$.

**What could go wrong?**

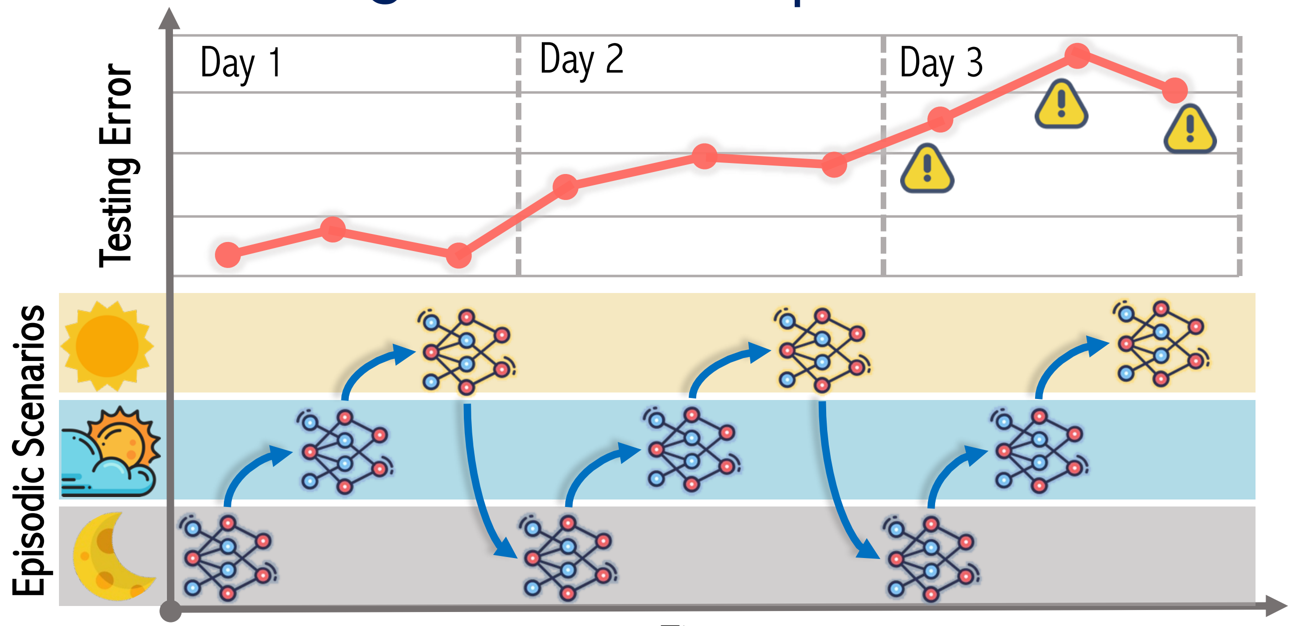
Does the model performance/adaptability persist after a long time adapting to multiple environments?

*Unfortunately, can not be guaranteed … we call it "TTA model collapsing"*

# Benchmark: Recurring Test-time Adaptation

## Hypothetical Setting

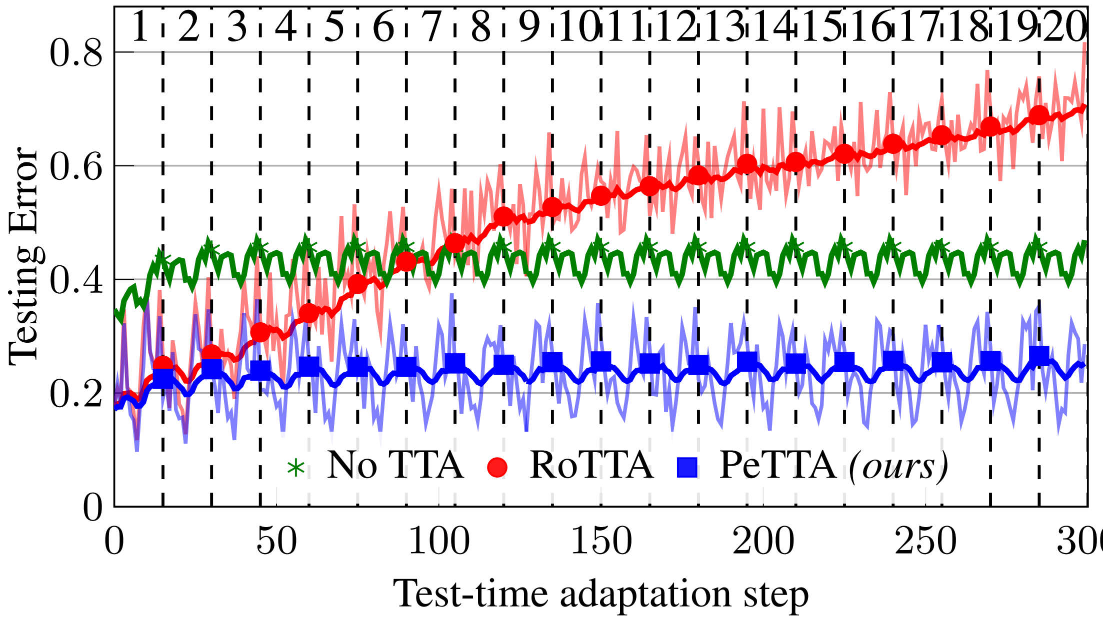
### Recurring Test-time Adaptation (TTA)




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

## Empirical Experiment

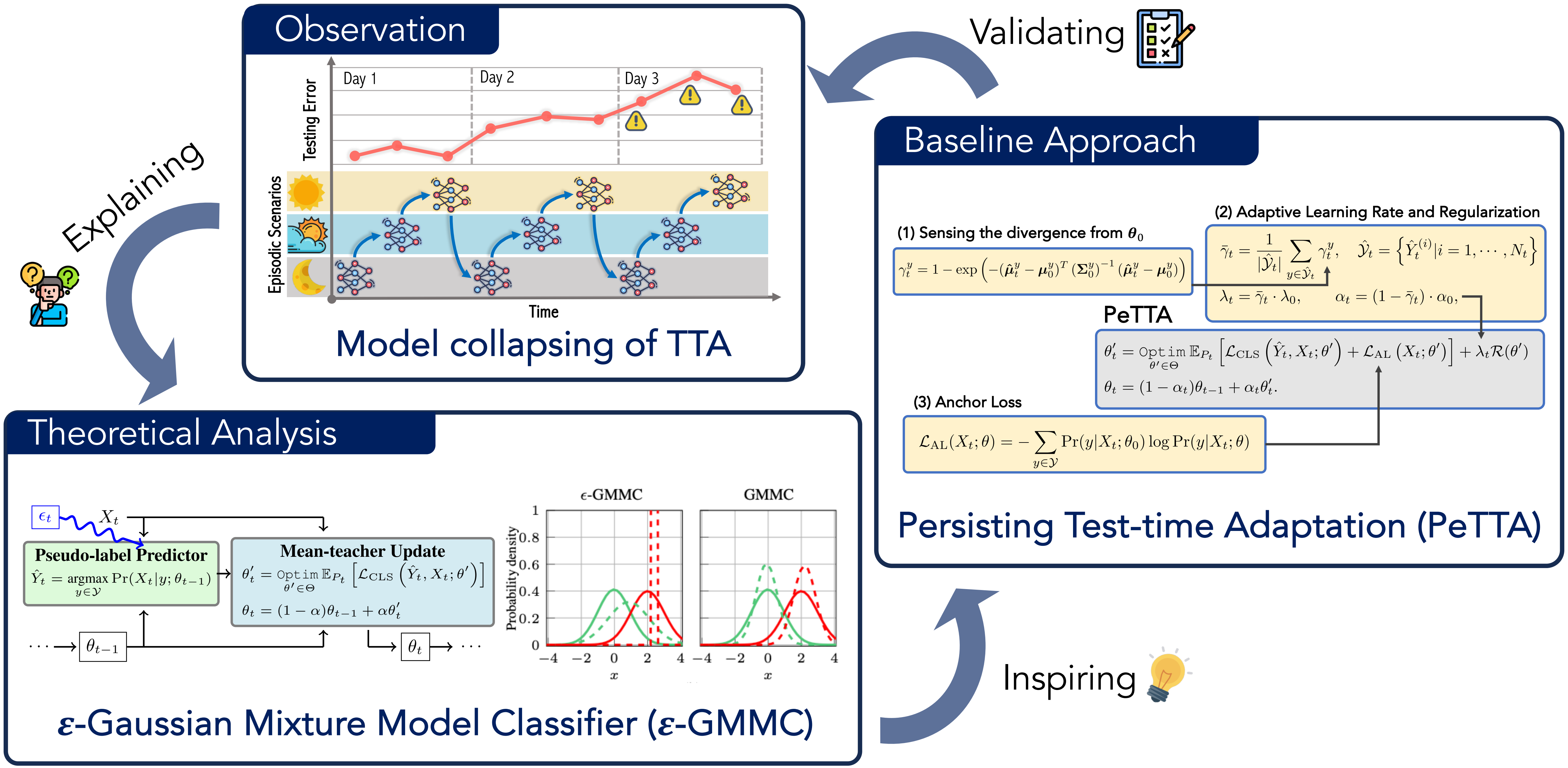
### Recurring TTA on CIFAR-10-C (corrupted)




- 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:** $\mathcal{P}_1 \rightarrow \mathcal{P}_2 \rightarrow \cdots \rightarrow \mathcal{P}_D \rightarrow \cdots \rightarrow \mathcal{P}_1 \rightarrow \mathcal{P}_2 \rightarrow \cdots \rightarrow \mathcal{P}_D$

# Overview: Persistent Test-time Adaptation (PeTTA)

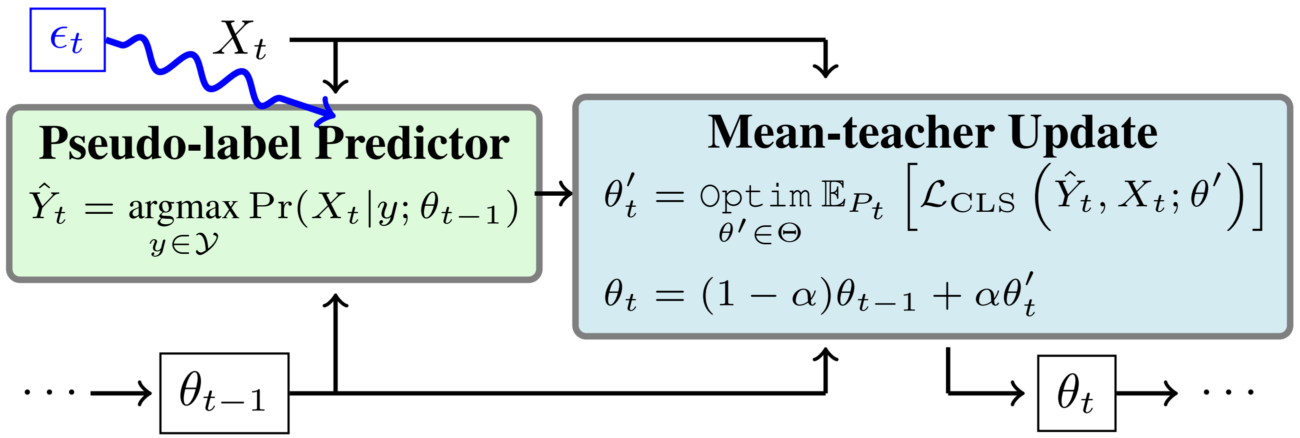


# Theoretical Analysis: $\varepsilon$-Gaussian Mixture Model Classifier (**$\varepsilon$-GMMC**)

**Goal:** Simulating *a simple yet representative* failure case of TTA for theoretical analysis

## Theoretical $\varepsilon$-Gaussian Mixture Model Classifier (**$\varepsilon$-GMMC**)

- **Data stream:** $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \{0,1\}$ and the underlying joint distribution $P_t(x, y) = p_{y,t}\mathcal{N}(x; \mu_y, \sigma_y^2)$ with $p_{y,t} = \Pr(Y_t = y)$ - true label and $\hat{p}_{y,t} = \Pr(\hat{Y}_t = y)$ - predicted label
- **Task:** predicting $X_t$ was sampled from cluster 0 or 1 (negative or positive)
- **Procedure:** *pseudo-label* $\hat{Y}_t$ prediction and a *mean-teacher* update




*A formal 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:*

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

- **"Noisy" pseudo-label predictor:** 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 TTA testing stream
- How this simple TTA model will be collapsed?

# Key Analysis Results

We then obtained the following theoretical results:

- **Why collapsing?**

*Increasing* the false-negative rate leads to model collapse

**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 $\lim_{t\to\tau} \epsilon_t = p_1$.*

- **After collapsing?**

Converging to a *single-cluster* model (instead of 2)

**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:*

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

- **How?** Conditions and factors that contribute to the model collapse

**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:*

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

**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 $\{\epsilon_t\}_{\tau-\Delta_\tau}^{\tau}$ $(\tau \ge \Delta_\tau > 0)$ such that:*

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

*Factors contributing to the model collapse:*

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

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

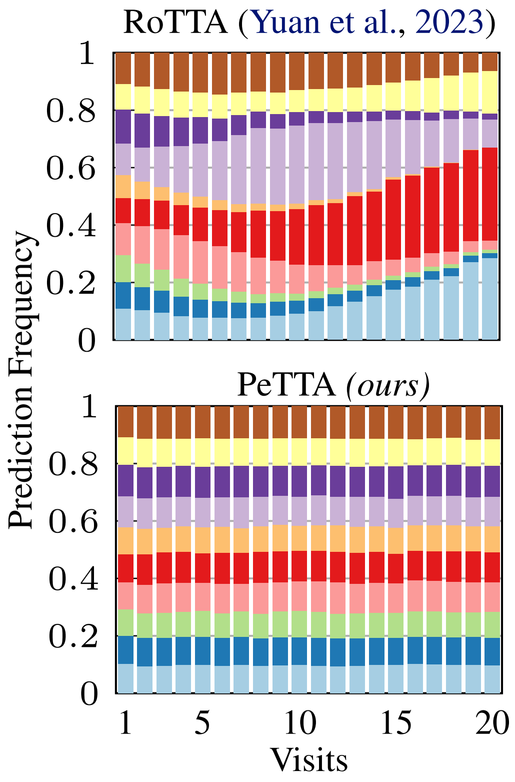
Under the static data stream assumption

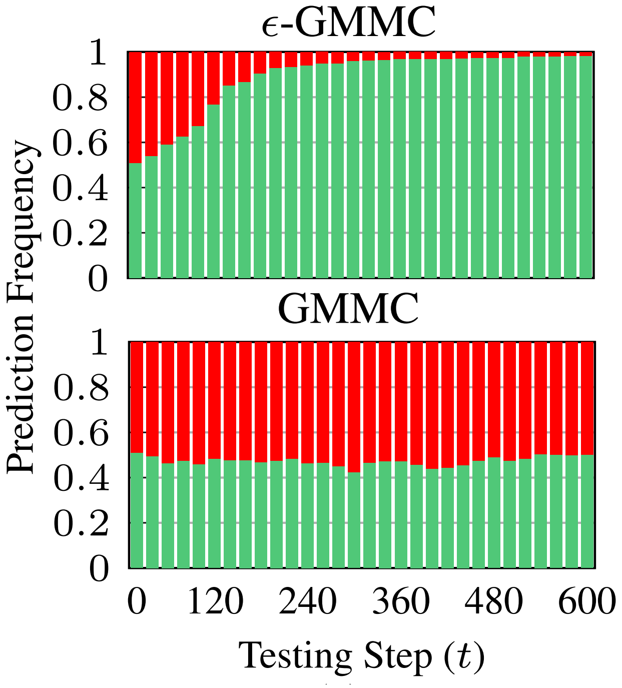
**Assumption 1 (Static Data Stream).** *The marginal distribution of the true label follows the same Bernoulli distribution $\mathrm{Ber}(p_0)$: $p_{0,t} = p_0$, $(p_{1,t} = p_1 = 1 - p_0)$, $\forall t \in \mathcal{T}$.*

# Simulation Results: Collapsing Behavior of ε-GMMC

We perform a numerical simulation to *empirically validate* the theoretical analysis

## Histogram of Predictions



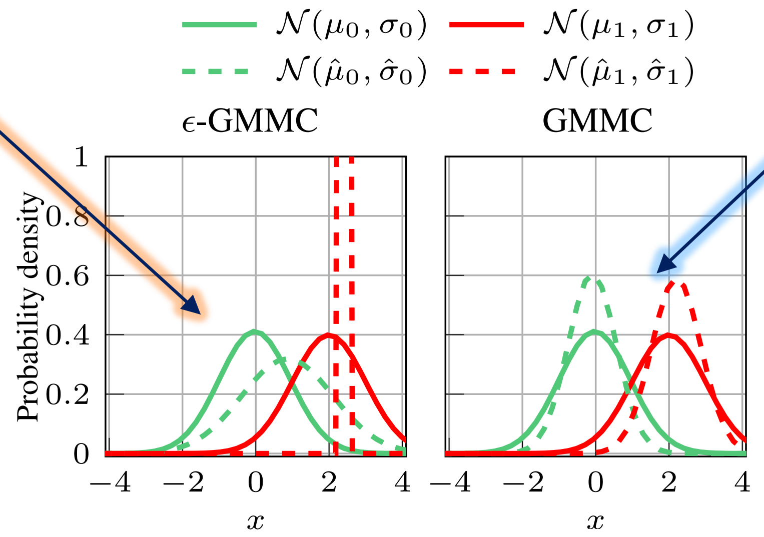

On real dataset[1]

Simulation

$\epsilon$ -GMMC simulates a *similar* collapsing pattern observed on RoTTA/CIFAR-10-C

## ε-GMMC After Collapsing

ε-GMMC collapsed into a single cluster model (predicting all zeros)

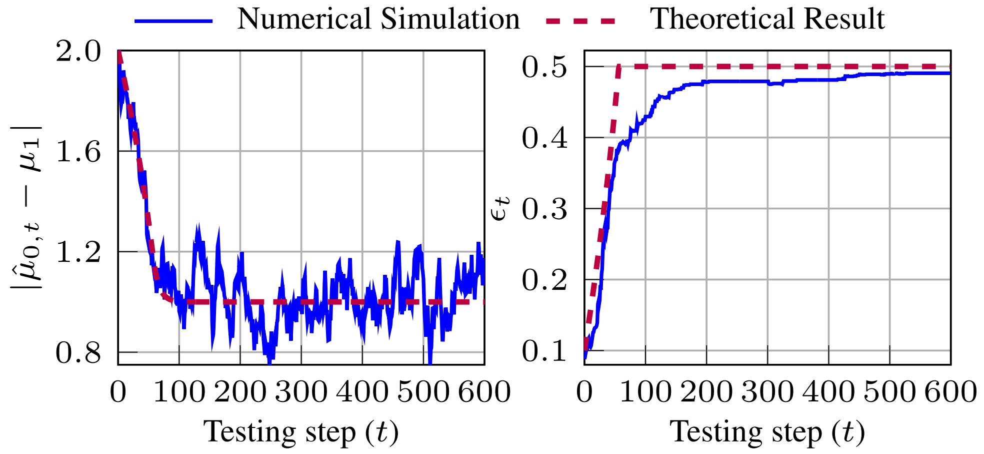



Collapsing

None collapsing

GMMC model *without* the noisy pseudo labels converges to the true distributions

## Collapsing Trajectory of ε-GMMC




Numerical simulation on a two-Gaussian models *aligns* with the theoretical analysis result

[1] Each column on these plots shows the histogram of model prediction (class labels are color-coded). CIFAR-10-C has an equal number of images for 10 classes. Hence, predictions from an ideal model should follow a uniform distribution.

# Proposed Approach: Persisting Test-time Adaptation (PeTTA)

- **Notation:** 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 the value of vector $\boldsymbol{z}$ if $f_t(\boldsymbol{x}) = y$
- **Key Idea:** Sensing the divergence of $\phi_{\theta_t}$ from $\phi_{\theta_0}$, and adjust the adaptation objective correspondingly
- With $\boldsymbol{\mu}_0^t$, $\boldsymbol{\Sigma}_0^t$ are pre-computed on the *source dataset*, we can:

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

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

(2) Adaptive Learning Rate and Regularization

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

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

PeTTA

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

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

Adaptation loss (↑ performance)

Regularization (↑ collapse prevention)

(3) Anchor Loss

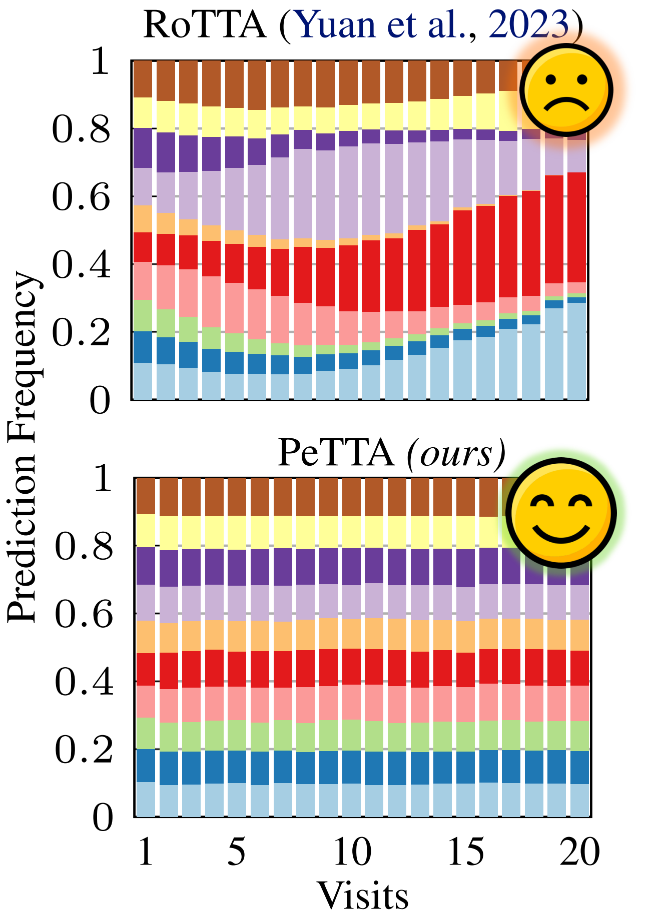
$$\mathcal{L}_{\text{AL}}(X_t; \theta) = -\sum_{y \in \mathcal{Y}} \Pr(y | X_t; \theta_0) \log \Pr(y | X_t; \theta)$$

PeTTA is an "elaborated" version of the regular mean-teacher update model

# Qualitative Results of PeTTA on CIFAR10-C

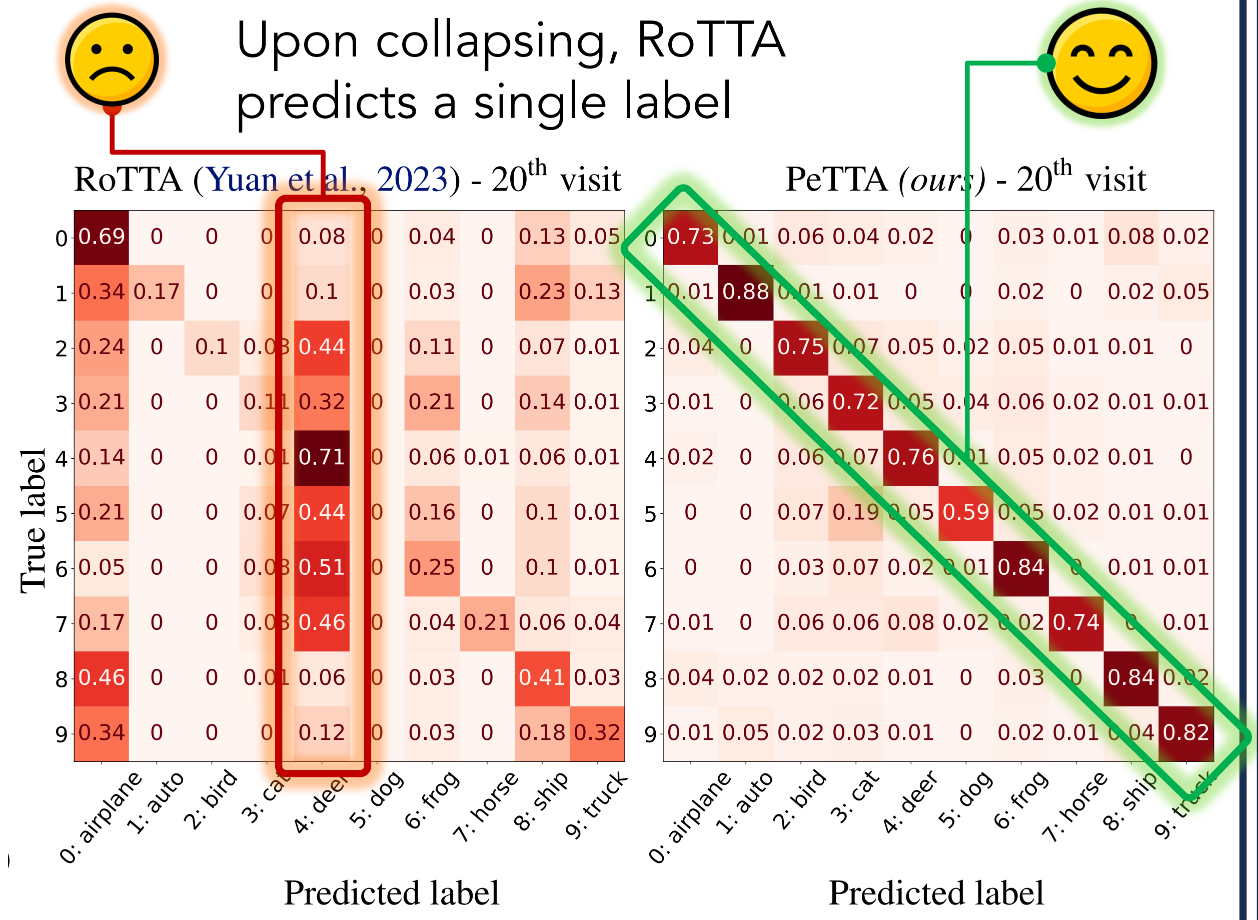
We qualitatively compare the performance of PeTTA (Persistent Test-time Adaptation) and RoTTA (Robust Test-time Adaptation [Yuan, 2023]) and analyze the model collapse on CIFAR10-C dataset

## Histogram of Predictions[1]




Class labels are color-coded, on CIFAR-10-C we have 10 classes

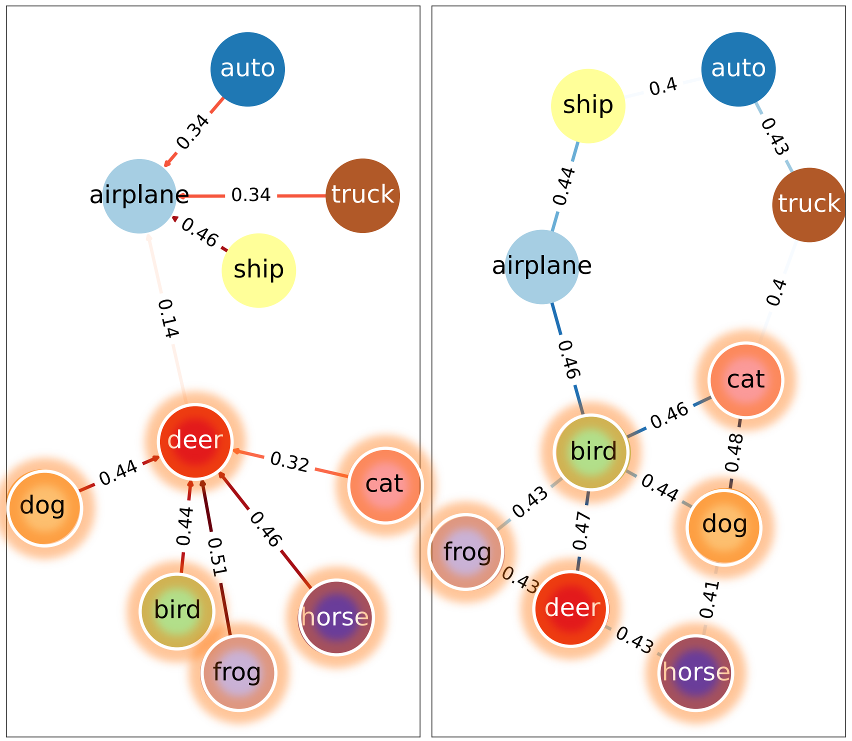
## Confusion Matrix (The last visit)




Confusion matrix of the two models at the last (20th) recurring TTA visit

## Intra-category Similarity

Predicting collapsing behavior




Misclassification rate of collapsed RoTTA

Inter-category cosine similarity (source model)

Force-directed graph showing the most *prone to misclassification* pairs *have highest similarity* in the feature space of the source model

[1] Each column on these plots shows the histogram of model prediction (class labels are color-coded). CIFAR-10-C has an equal number of images for 10 classes. Hence, predictions from an ideal model should follow a uniform distribution.

# Quantitative Results of PeTTA & Ablation Studies

We evaluate our PeTTA and five other comparable TTA methods in recurring TTA setting on ImageNet-C dataset

| 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** |

PeTTA achieves the lowest average error

PeTTA shows a *persisting performance* across 20 recurring TTA visits

## Ablation Studies:

| Method: Regularizer | Method: Fisher | CIFAR-10-C | CIFAR-100-C | DomainNet |
|---|---|---|---|---|
| L2 | ✗ | 23.0 | 35.6 | 43.1 |
| | ✓ | 22.7 | 36.0 | 43.9 |
| Cosine | ✗ | 23.0 | **35.2** | **42.5** |
| | ✓ | **22.6** | 35.9 | 43.3 |

PeTTA favors various choices of regularizer $\mathcal{R}(\theta)$

| Method | CIFAR-10-C | CIFAR-100-C | DomainNet |
|---|---|---|---|
| Baseline w/o $\mathcal{R}(\theta)$ | 42.6 | 63.0 | 77.9 |
| $\mathcal{R}(\theta)$ fixed $\lambda = 0.1\lambda_0$ | 43.3 | 65.0 | 80.0 |
| $\mathcal{R}(\theta)$ fixed $\lambda = \lambda_0$ | 42.0 | 64.6 | 66.6 |
| PeTTA - $\lambda_t$ | 27.1 | 55.0 | 59.7 |
| PeTTA - $\lambda_t + \alpha_t$ | 23.9 | 41.4 | 44.5 |
| PeTTA - $\lambda_t + \mathcal{L}_{AL}$ | 26.2 | 36.3 | 43.2 |
| PeTTA - $\lambda_t + \alpha_t + \mathcal{L}_{AL}$ | **23.0** | **35.2** | **42.5** |

Without using/ fixed regularization coefficients does not address the performance degradation

To maintain persistence, utilizing all components is suggested in PeTTA

# Conclusions: Persistent Test-time Adaptation (PeTTA)

Introducing a new testing scenario – *recurring TTA* for demonstrating the performance degradation of existing continual TTA methods

Conducting theoretical analysis on performance degradation of TTA on $\epsilon$*–GMMC*, indicating factors that contribute to model collapse

Introducing a new baseline – *persistent TTA (PeTTA).* PeTTA strikes a balance between two objectives: adaptation and collapse prevention

For more information, visit our project page at 👉
See you at:
***POSTER SECTION 4 (Thursday Afternoon)***

PROJECT