JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Robust Test-time Adaptation in Dynamic Scenarios

Longhui Yuan, Binhui Xie, Shuang Li*

Beijing Institute of Technology

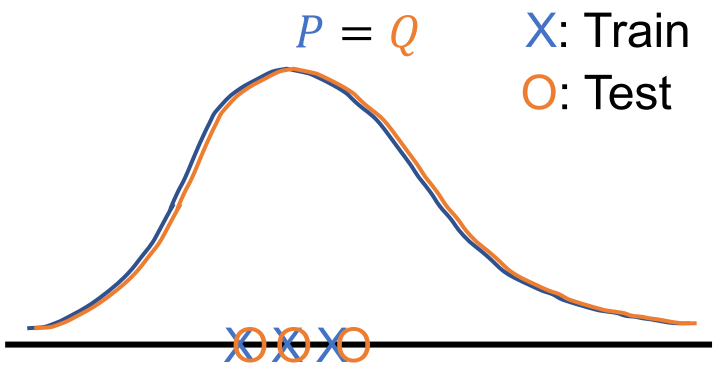
WED-PM-340

Summary

- Problem setup: Practical Test-Time Adaptation (PTTA)
  - distribution change & correlative sampling
- Method: Robust Test-Time Adaptation (RoTTA)
  - robust statistics estimation & category-balanced sampling & time-aware reweighting
- Experiments
  - Performance gain: 5.9% on CIFAR-10-C, 5.5% on CIFAR-100-C, and 2.2% on DomainNet

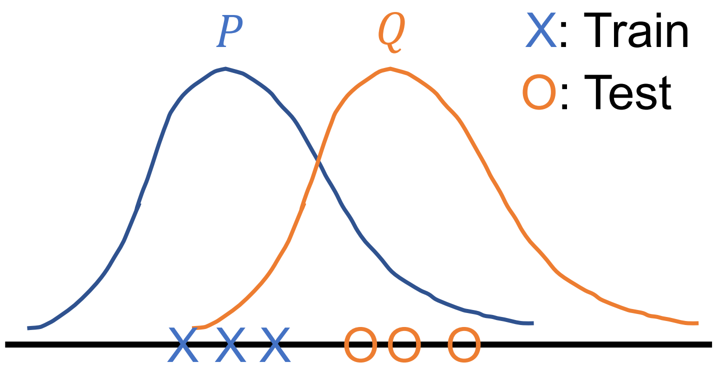
Background

- When training data and test data come from the same distribution, deep learning achieves excellent performance.



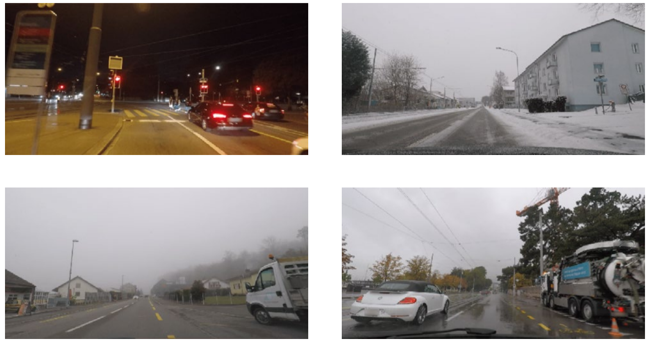
Distribution Shift

- When training data and test data come from the same distribution, deep learning achieves excellent performance.
- **In real world**: distribution shifts exist everywhere.



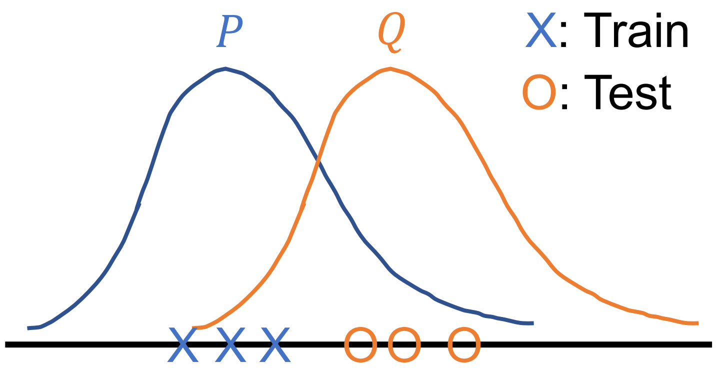
Training data

Test data



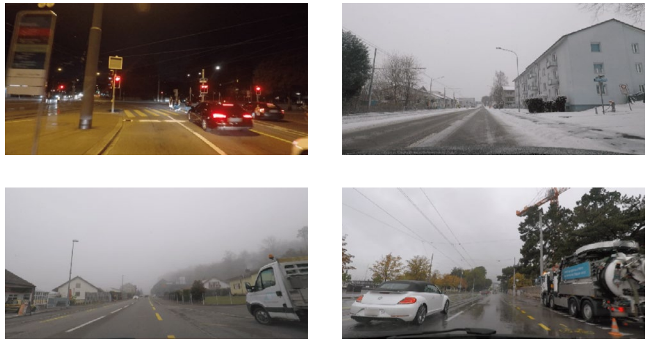
Distribution Shift

- When training data and test data come from the same distribution, deep learning achieves excellent performance.
- **In real world**: distribution shifts exist everywhere.



Training data

Test data



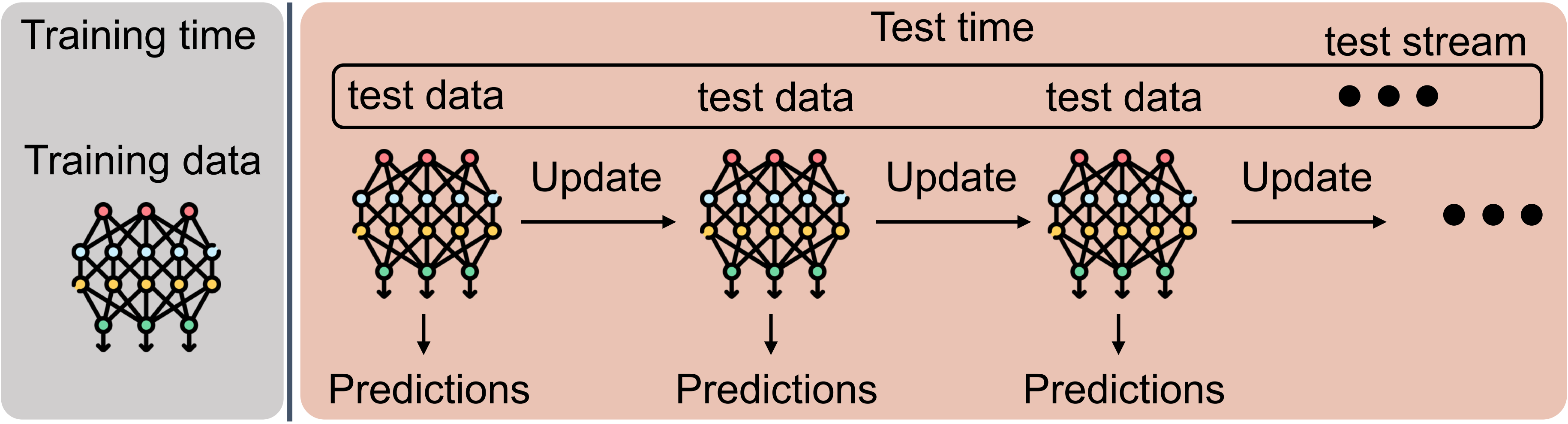
Adapting deep models to new domains is critical

Test-time Adaptation

- Test-time adaptation (TTA) attempts to address distribution shifts during the inference stage.

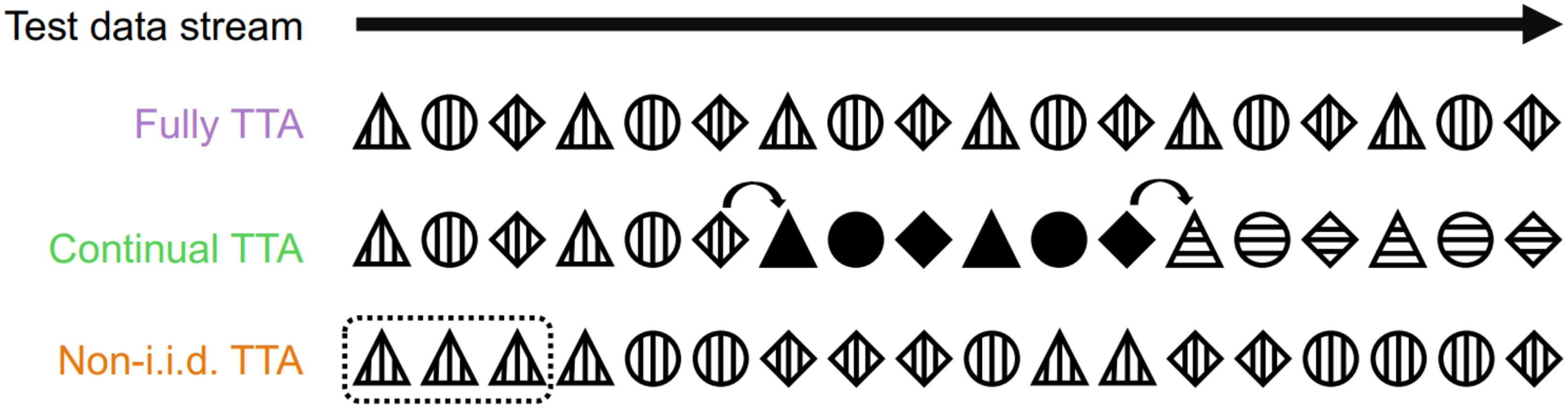
Test-time Adaptation

- Test-time adaptation (TTA) attempts to address distribution shifts during the inference stage.



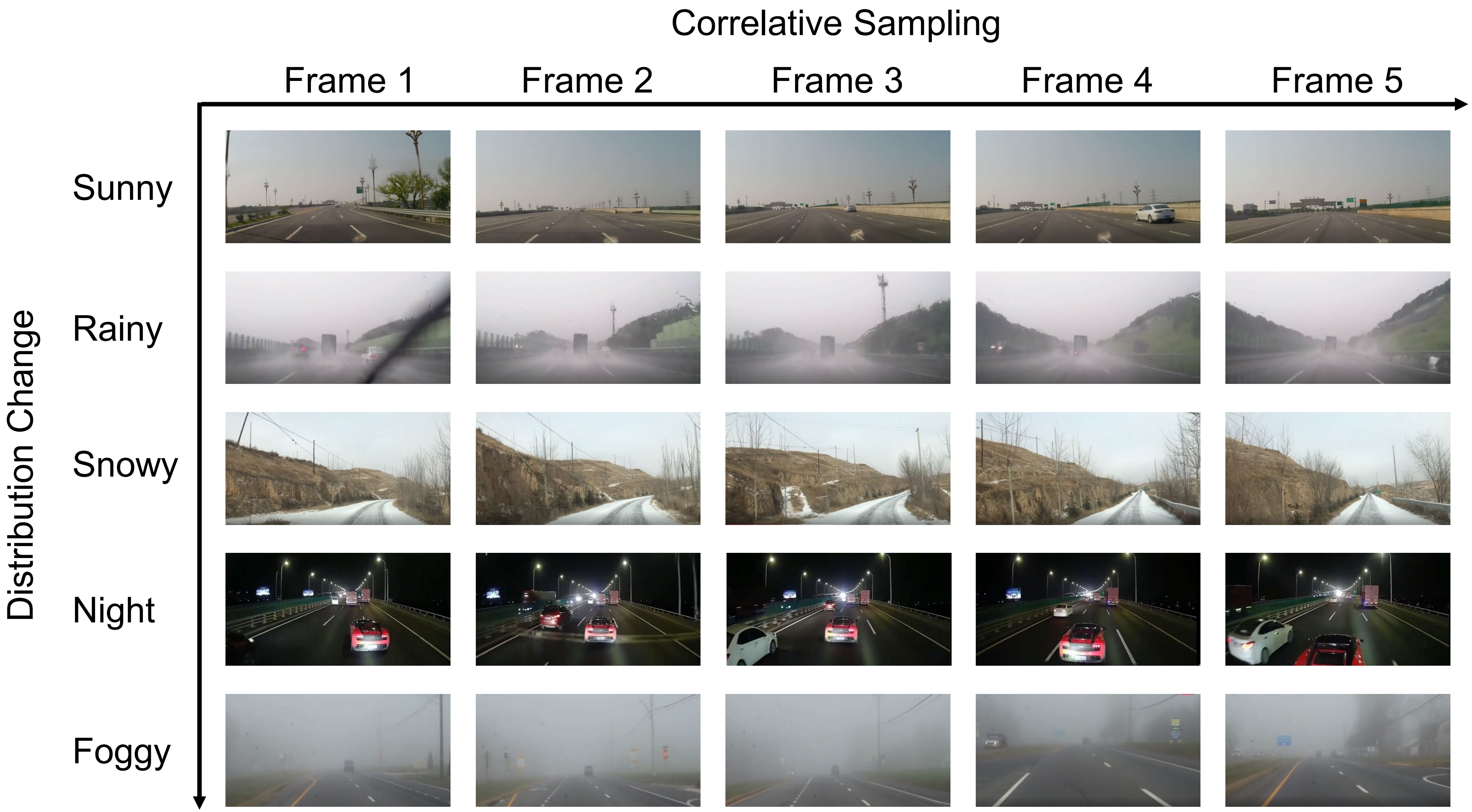
Test-time Adaptation

- Test-time adaptation (TTA) attempts to address distribution shifts during the inference stage.
- Existing setup: Fully TTA, Continual TTA, Non-i.i.d. TTA



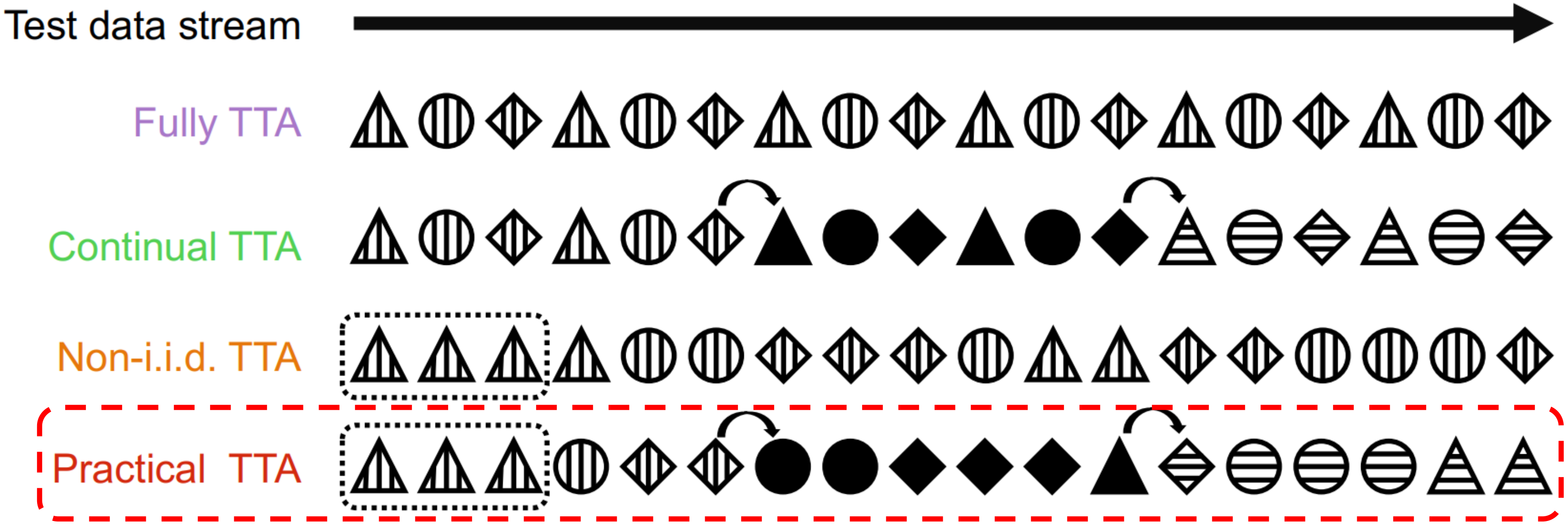
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Test Stream in Real World



Practical Test-time Adaptation

- A more practical test-time adaptation setup where distribution change and correlative sampling occur simultaneously.



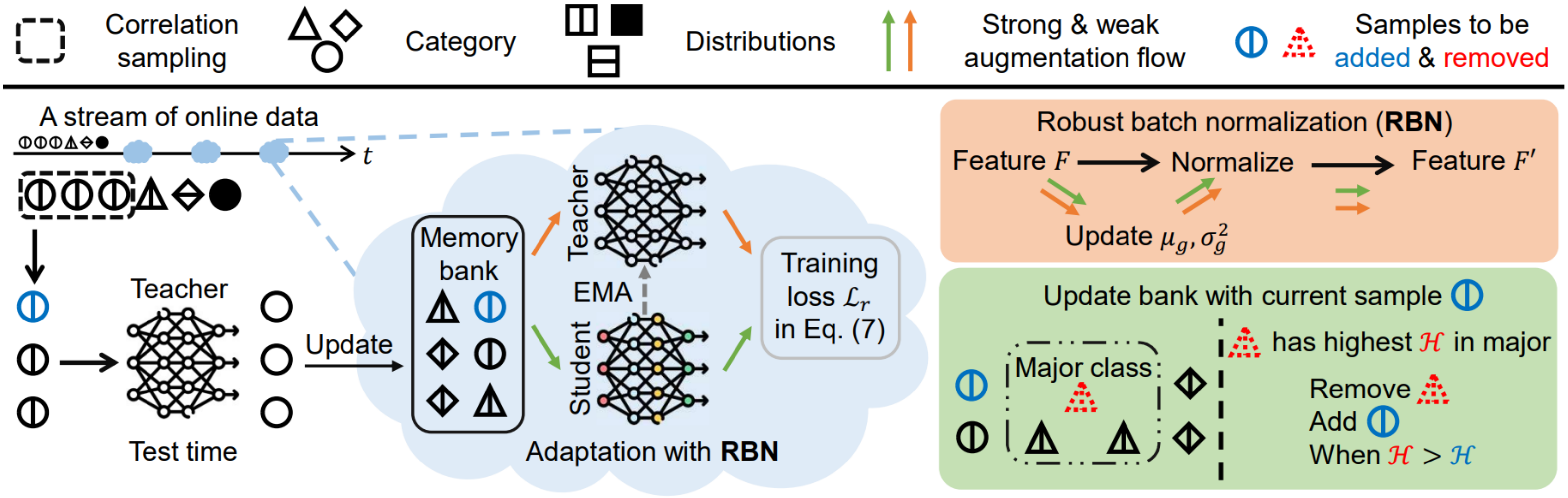
Practical Test-time Adaptation

- A more practical test-time adaptation setup where distribution change and correlative sampling occur simultaneously.
- Challenges:
  - Incorrect estimation in the batch normalization statistics
  - Easily or quickly overfit to the local distribution
  - Error gradients cumulate

Practical Test-time Adaptation

- A more practical test-time adaptation setup where distribution change and correlative sampling occur simultaneously.
- Formulate PTTA as:
  - Given a model f_{θ_0} pre-trained on source domain $\mathcal{D}_{\mathcal{S}} = \{(x_s, y_s)\}$.
  - A stream of unlabeled test samples $\mathcal{X}_0, \cdots, \mathcal{X}_T, \cdots$, where $\mathcal{X}_t$ is a batch of correlated samples from distribution $\mathcal{P}_{test}$ and $\mathcal{P}_{test}$ changes as $\mathcal{P}_0, \mathcal{P}_1, \cdots, \mathcal{P}_\infty$.

Framework



Correlation sampling
Category
Distributions
Strong & weak augmentation flow
Samples to be added & removed
A stream of online data
t
Teacher
Test time
Update
Memory bank
Teacher
EMA
Student
Training loss $\mathcal{L}_r$ in Eq. (7)
Adaptation with **RBN**
Robust batch normalization (**RBN**)
Feature F
Normalize
Feature F'
Update μ_g, σ_g^2
Update bank with current sample
Major class
has highest $\mathcal{H}$ in major
Remove
Add
When $\mathcal{H} > \mathcal{H}$

RoTTA

- The correlation among test samples $\mathcal{X}_t$ at time t leads to a deviation between the observed distribution $\hat{\mathcal{P}}_{test}$ and the test distribution $\mathcal{P}_{test}$.
- Directly adapting on $\hat{\mathcal{P}}_{test}$ leads overfitting ×
- Maintaining a category-balanced memory bank √

RoTTA

- The balance among categories is guaranteed by distributing the capacity of $\mathcal{M}$ equally to each category (refer to lines 5 – 9)
- Heuristic score of timeliness and uncertainty

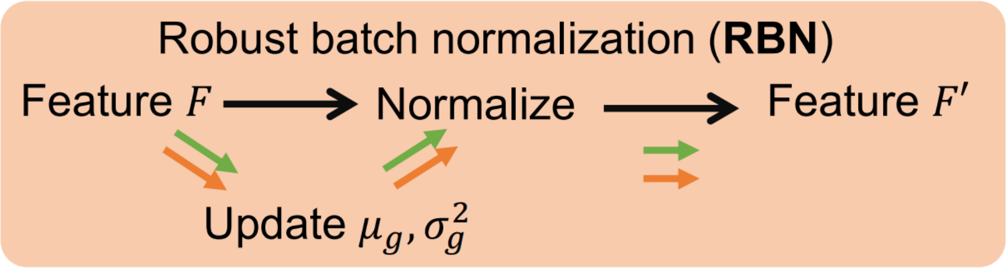
$$\mathcal{H} = \lambda_t \frac{1}{1+\exp(-\mathcal{A}/\mathcal{N})} + \lambda_u \frac{\mathcal{U}}{\log \mathcal{C}}$$

Algorithm 1: CSTU for one test sample.

1 **Input:** a test sample x and the teacher model f_{θ^T}.
2 **Define:** memory bank $\mathcal{M}$ and its capacity $\mathcal{N}$, number of classes $\mathcal{C}$, per class occupation $\mathcal{O} \in \mathbf{R}^{\mathcal{C}}$, total occupation Ω, classes to pop instance $\mathcal{D}$.
3 Infer as $p(y|x) = \text{Softmax}(f_{\theta^T}(x))$.
4 Calculate the predicted category of x as $\hat{y} = \arg\max_c p(c|x)$, the uncertainty as $\mathcal{U}_x = -\sum_{c=1}^{C} p(c|x)\log(p(c|x))$, the age as $\mathcal{A}_x = 0$, and the heuristic score $\mathcal{H}_x$ of x with Eq (6)
5 **if** $\mathcal{O}_{\hat{y}} < \frac{\mathcal{N}}{\mathcal{C}}$ **then**
6   if $\Omega < \mathcal{N}$: Search range $\mathcal{D} = \emptyset$.
7   else: Search range $\mathcal{D} = \{j | j = \arg\max_c \mathcal{O}_c\}$
8 **else**
9   Search range $\mathcal{D} = \{\hat{y}\}$
10 **if** $\mathcal{D}$ *is* $\emptyset$ **then**
11   Add $(x, \hat{y}, \mathcal{H}_x, \mathcal{U}_x)$ into $\mathcal{M}$.
12 **else**
13   Find the instance $(\hat{x}, y_{\hat{x}}, \mathcal{A}_{\hat{x}}, \mathcal{U}_{\hat{x}})$ with the highest value in Eq (6) $\mathcal{H}_{\hat{x}}$ among $\mathcal{D}$.
14   **if** $\mathcal{H}_x < \mathcal{H}_{\hat{x}}$ **then**
15   Remove $(\hat{x}, y_{\hat{x}}, \mathcal{A}_{\hat{x}}, \mathcal{U}_{\hat{x}})$ from $\mathcal{M}$.
16   Add $(x, \hat{y}, \mathcal{H}_x, \mathcal{U}_x)$ into $\mathcal{M}$.
17   **else**
18   Discard x.
19 Increase the age of all instances in $\mathcal{M}$.

RoTTA

- High correlation also makes the wildly used Test-time BN invalid
- Maintain a group of global statistics from the memory bank, and adopt it for inference



$$\mu_g = (1-\alpha)\mu_g + \alpha\mu_b$$

$$\sigma_g^2 = (1-\alpha)\sigma_g^2 + \alpha\sigma_b^2$$

RoTTA

- Training objective at time t:

$$\min_{\theta_{t+1}^S} \mathcal{L}_r = \frac{1}{\Omega}\sum_{i=1}^{\Omega} \mathcal{L}(x_i^{\mathcal{M}}, \mathcal{A}_i; \theta_t^T, \theta_t^S) = \frac{1}{\Omega}\sum_{i=1}^{\Omega} E(\mathcal{A}_i)\ell(x_i', x_i'')$$

$$E(\mathcal{A}_i) = \frac{\exp(-\mathcal{A}_i/\mathcal{N})}{1+\exp(-\mathcal{A}/\mathcal{N})}, \qquad \ell(x_i', x_i'') = -\frac{1}{\mathcal{C}}\sum_{c=1}^{\mathcal{C}} p_T(c|x_i')\log\left(p_S(c|x_i'')\right)$$

- Update teacher model by EMA: $\theta_{t+1}^T = (1-\nu)\theta_t^T + \nu\theta_{t+1}^S$.

Experiments

Table 2. Average classification error of the task CIFAR10 → CIFAR10-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

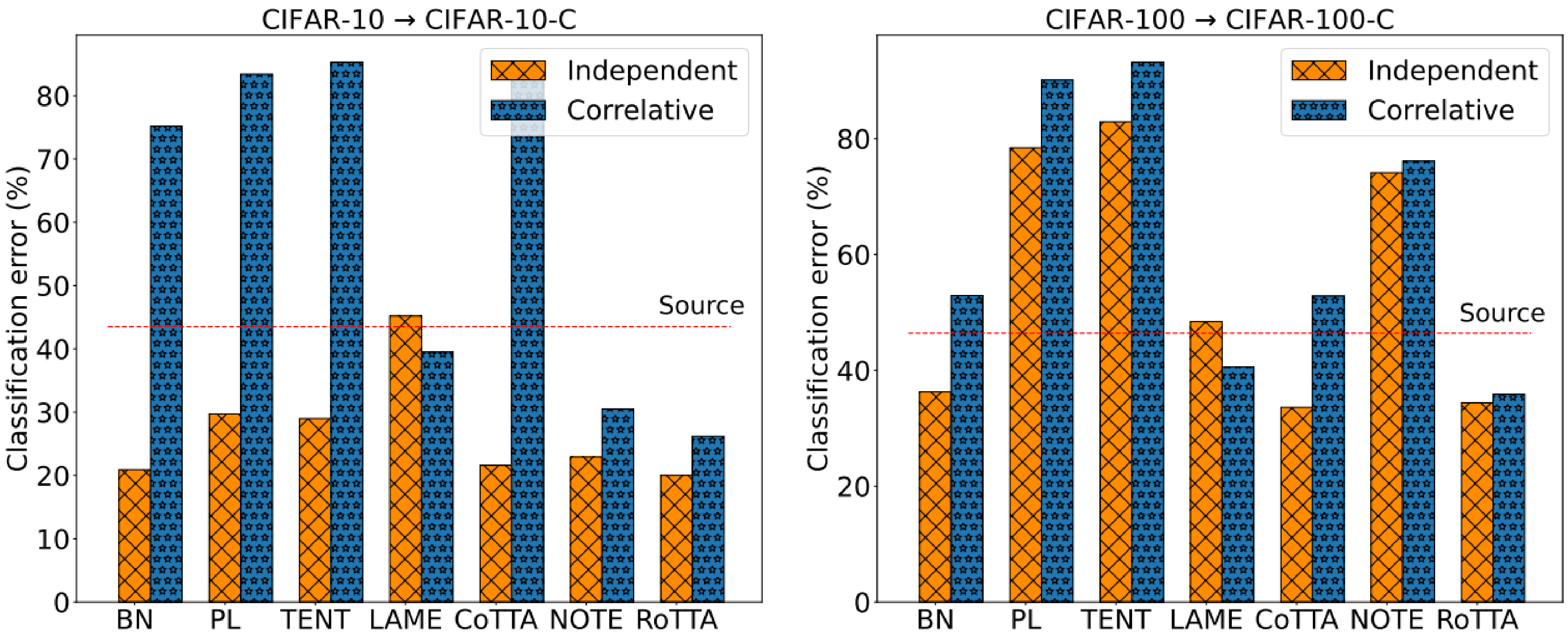
Time	t →															
Method	motion	snow	fog	shot	defocus	contrast	zoom	brightness	frost	elastic	glass	gaussian	pixelate	jpeg	impulse	Avg.
Source	34.8	25.1	26.0	65.7	46.9	46.7	42.0	9.3	41.3	26.6	54.3	72.3	58.5	30.3	72.9	43.5
BN [53]	73.2	73.4	72.7	77.2	73.7	72.5	72.9	71.0	74.1	77.7	80.0	76.9	75.5	78.3	79.0	75.2
PL [39]	73.9	75.0	75.6	81.0	79.9	80.6	82.0	83.2	85.3	87.3	88.3	87.5	87.5	87.5	88.2	82.9
TENT [70]	74.3	77.4	80.1	86.2	86.7	87.3	87.9	87.4	88.2	89.0	89.2	89.0	88.3	89.7	89.2	86.0
LAME [5]	29.5	**19.0**	20.3	65.3	42.4	43.4	36.8	**5.4**	37.2	**18.6**	51.2	73.2	57.0	**22.6**	71.3	39.5
CoTTA [73]	77.1	80.6	83.1	84.4	83.9	84.2	83.1	82.6	84.4	84.2	84.5	84.6	82.7	83.8	84.9	83.2
NOTE [19]	**18.0**	22.1	20.6	35.6	26.9	**13.6**	26.5	17.3	27.2	37.0	48.3	38.8	42.6	41.9	49.7	31.1
RoTTA	18.1	21.3	**18.8**	**33.6**	**23.6**	16.5	**15.1**	11.2	**21.9**	30.7	**39.6**	**26.8**	**33.7**	27.8	**39.5**	**25.2**(+5.9)

Table 3. Average classification error of the task CIFAR100 → CIFAR100-C while continually adapting to different corruptions at the highest severity 5 with correlatively sampled test stream under the proposed setup PTTA.

Time	t →															
Method	motion	snow	fog	shot	defocus	contrast	zoom	brightness	frost	elastic	glass	gaussian	pixelate	jpeg	impulse	Avg.
Source	30.8	39.5	50.3	68.0	29.3	55.1	28.8	29.5	45.8	37.2	54.1	73.0	74.7	41.2	39.4	46.4
BN [53]	48.5	54.0	58.9	56.2	46.4	48.0	47.0	45.4	52.9	53.4	57.1	58.2	51.7	57.1	58.8	52.9
PL [39]	50.6	62.1	73.9	87.8	90.8	96.0	94.8	96.4	97.4	97.2	97.4	97.4	97.3	97.4	97.4	88.9
TENT [70]	53.3	77.6	93.0	96.5	96.7	97.5	97.1	97.5	97.3	97.2	97.1	97.7	97.6	98.0	98.3	92.8
LAME [5]	**22.4**	**30.4**	43.9	66.3	**21.3**	51.7	**20.6**	**21.8**	39.6	**28.0**	48.7	72.8	74.6	**33.1**	**32.3**	40.5
CoTTA [73]	49.2	52.7	56.8	53.0	48.7	51.7	49.4	48.7	52.5	52.2	54.3	54.9	49.6	53.4	56.2	52.2
NOTE [19]	45.7	53.0	58.2	65.6	54.2	52.0	59.8	63.5	74.8	91.8	98.1	98.3	96.8	97.0	98.2	73.8
RoTTA	31.8	36.7	**40.9**	**42.1**	30.0	**33.6**	27.9	25.4	**32.3**	34.0	**38.8**	**38.7**	**31.3**	38.0	42.9	**35.0**(+5.5)

Experiments

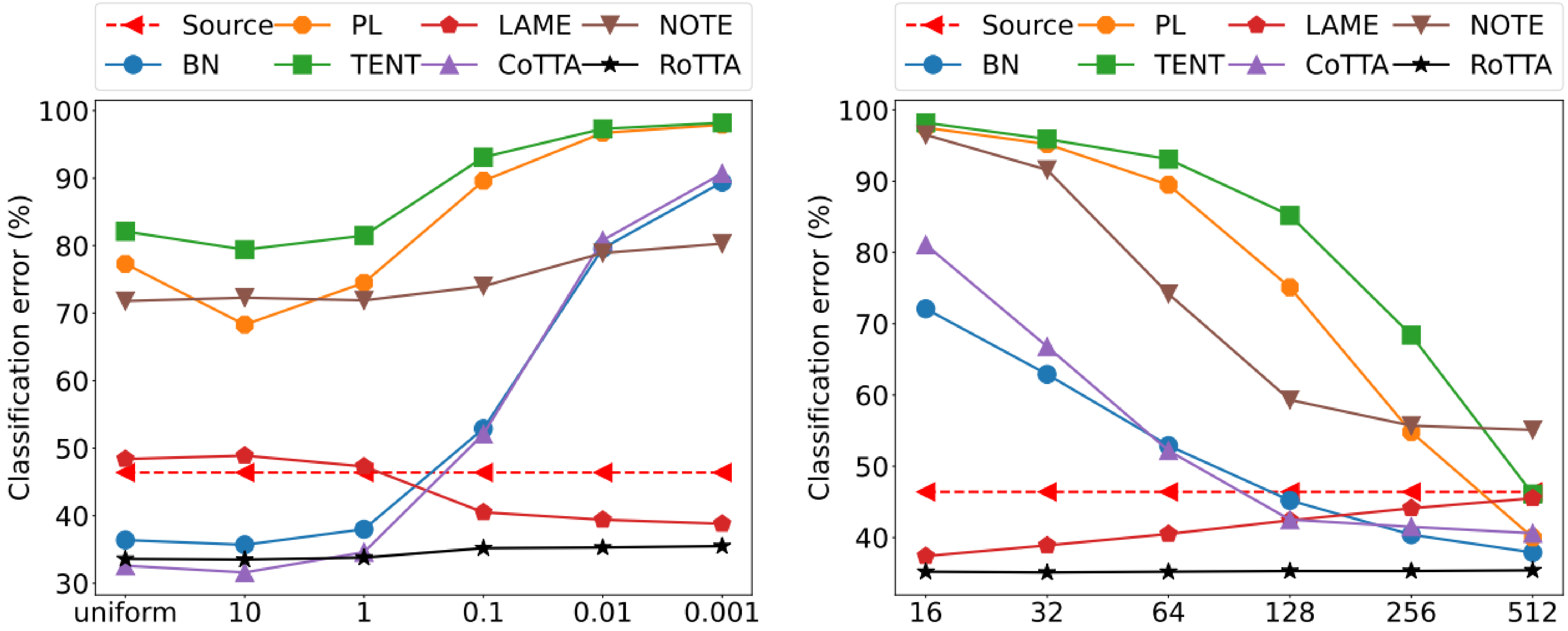
RoTTA achieves excellent results under various setup and distribution change orders



(a) CIFAR10-C. (b) CIFAR100-C.

Experiments

Excellent and stable results prove the stability and effectiveness of RoTTA



(c) δ. (d) Batch size.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Thank You~

Longhui Yuan

arXiv

Code

Home Page

Email