# Towards Stable Test-Time Adaptation in Dynamic Wild World

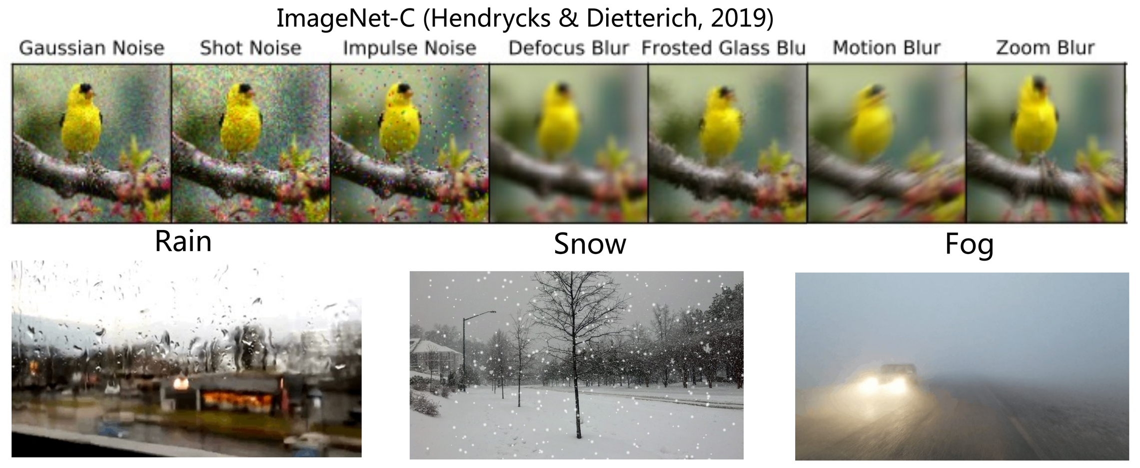
Shuaicheng Niu*, Jiaxiang Wu*, **Yifan Zhang***, Zhiquan Wen, Yaofo Chen, Peilin Zhao, Mingkui Tan

National University of Singapore

## Background: Data Shifts

Distribution shifts: when using a pre-trained model, the test samples may encounter natural variations or corruptions that were not present in training data:

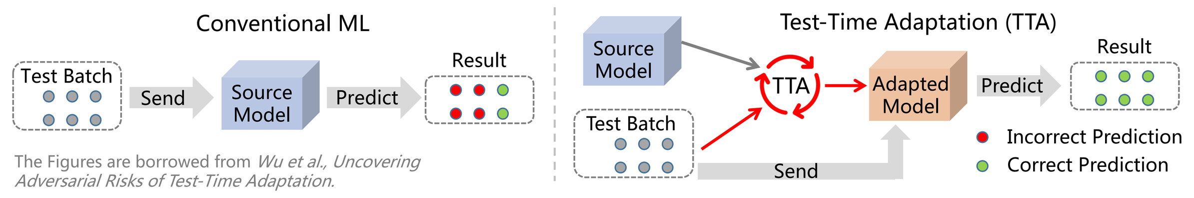
- *Lighting changes* due to weather change
- *Noises* due to sensor degradation, etc.




These shifts can significantly impact the performance of the model and cause it to degrade

## Test-Time Adaptation (TTA)

- TTA aims to address data shifts by adapting the trained model on test data before prediction



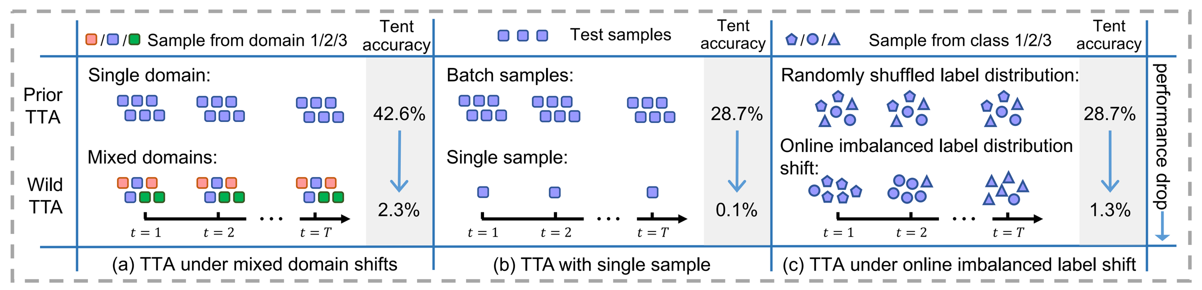

*The Figures are borrowed from Wu et al., Uncovering Adversarial Risks of Test-Time Adaptation.*

- Fully TTA adapts models online with only $x_{test}$

| Setting | Source data | Target data | Training loss | Testing loss | Offline | Online |
|---|---|---|---|---|---|---|
| Fine-tuning | × | $x^t, y^t$ | $\mathcal{L}(x^t, y^t)$ | -- | √ | × |
| Unsupervised domain adaptation | $x^s, y^s$ | $x^t$ | $\mathcal{L}(x^s, y^s) + \mathcal{L}(x^s, x^t)$ | -- | √ | × |
| Test-time training [ICML 20] | $x^s, y^s$ | $x^t$ | $\mathcal{L}(x^s, y^s) + \mathcal{L}(x^s)$ | $\mathcal{L}(x^t)$ | × | √ |
| Fully test-time adaptation [ICLR 21] | × | $x^t$ | × | $\mathcal{L}(x^t)$ | × | √ |

## Problem: TTA in the Wild

**Limitation:** online TTA is unstable under wild test scenarios (such as mixed domain shifts, single data, and imbalance), leading to severe model collapse




**Goal:** we aim to figure out the reason why TTA is unstable in the wild world, and then boost its stability

## Main Contributions

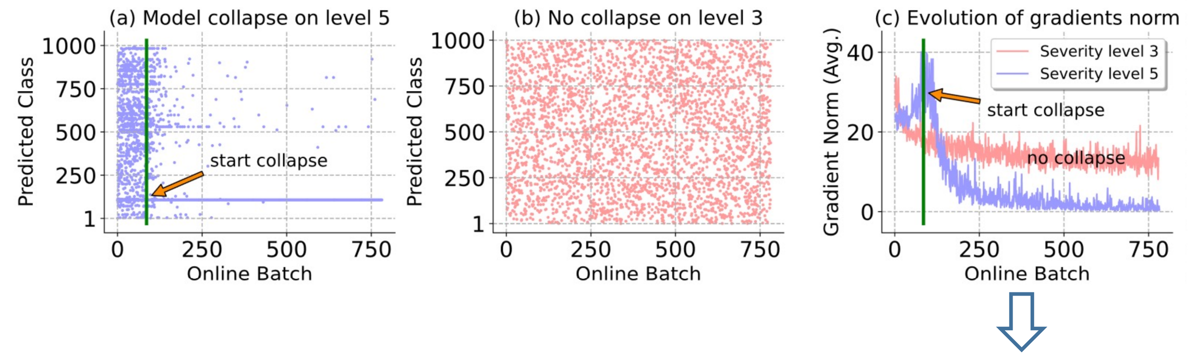
- We find that batch-agnostic norm layers (i.e., GN and LN) are more beneficial to stable TTA than BN under wild test settings
- We propose a simple yet effective SAR, which addresses the model collapse of online TTA and makes it more stable under wild test settings

## I: What Leads to Unstable TTA?

- Batch Normalization (BN) is a crucial factor hindering TTA stability under the wild test settings
- Most TTA methods are built upon test-time BN statistics adaptation: $y^{(k)} = \gamma^{(k)}\widehat{x}^{(k)} + \beta^{(k)}$, $\widehat{x}^{(k)} = \left(x^{(k)} - \mathbb{E}\left[x^{(k)}\right]\right) / \sqrt{\mathrm{Var}\left[x^{(k)}\right]}$
- However, the $\mathbb{E}$ and Var estimation under wild settings would be inaccurate:
  - *Mixed domain shifts*: ideally each domain should have its own statistics
  - *Single sample*: hard to estimate $\mathbb{E}$&Var accurately
  - *Imbalanced label shifts*: biased to specific classes
- Observation: models with **batch-agnostic** norm layer (*e.g.*, layer norm) are **more suitable** for TTA

## II: What Leads to Unstable TTA

- TTA on models with GN/LN layers do not always succeed, and still suffer from failure cases
- Online entropy minimization tends to result in collapsed trivial solutions, *i.e.*, predicting all samples to the same class, as shown in (a) vs. (b)




- Some large/noisy gradients cause collapse, as in (c)
- We address this collapse issue by proposing a **SAR** approach, as illustrated below

## SAR: Sharpness-Aware and Reliable Entropy Minimization

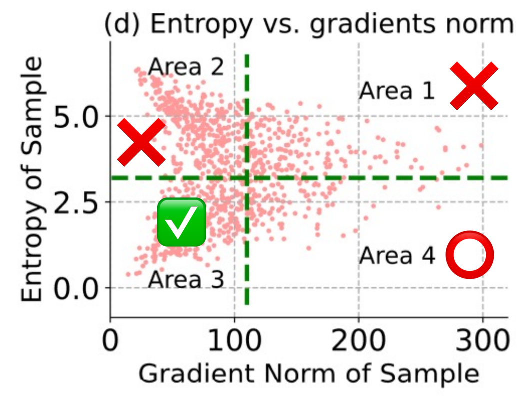
- Directly filtering out noisy gradients via gradients norm is infeasible, since the threshold is hard to set
- We seek to filter samples via an alternative metric, and investigate the relation of entropy *vs.* gradients norm

❶ **Reliability:** discard partial large/noisy gradients via entropy

- Remove samples in Areas 1 and 2:

$$\min_{\Theta} S(\mathbf{x})E(\mathbf{x};\Theta), \quad \text{where} \quad S(\mathbf{x}) \triangleq \mathbb{I}_{\{E(\mathbf{x};\Theta)<E_0\}}(\mathbf{x})$$

- Samples in Area 1 have large gradients
- Samples in Area 2 are unconfident (Niu et al., 2022)

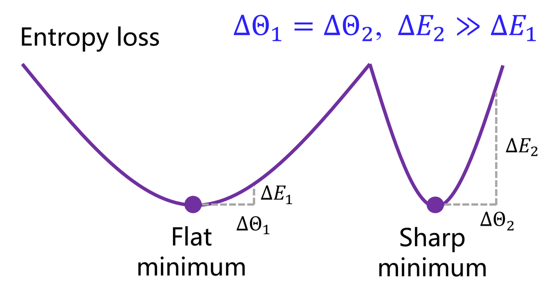



❷ **Sharpness-Aware:** make the update robust to remaining large/noisy gradients

- Alleviate the effects of samples in Area 4
- Constrain the entropy surface to be flat:

$$\min_{\Theta} E^{SA}(\mathbf{x};\Theta), \quad \text{where} \quad E^{SA}(\mathbf{x};\Theta) \triangleq \max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} E(\mathbf{x};\Theta+\boldsymbol{\epsilon})$$

- Following SAM (Foret et al., 2020) to solve this optimization problem




## Results under Online Imbalanced Label Distribution Shifts

- Our SAR achieves the best performance over ResNet50- GN and VitBase-LN
- Compare with Tent, OOD accuracy clearly improves, i.e., 22.0% → 37.2% on R-50-GN
- Entropy minimization on LN/GN models perform better than that on BN models

Results on ImageNet-C with severity level 5 regarding Corruption Accuracy (%)

| Model+Method | Gauss. | Shot | Impul. | Defoc. | Glass | Motion | Zoom | Snow | Frost | Fog | Brit. | Contr. | Elastic | Pixel | JPEG | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ResNet50 (BN) | 2.2 | 2.9 | 1.8 | 17.8 | 9.8 | 14.5 | 22.5 | 16.8 | 23.4 | 24.6 | 59.0 | 5.5 | 17.1 | 20.7 | 31.6 | 18.0 |
| • MEMO | 7.4 | 8.6 | 8.9 | 19.8 | 13.2 | 20.8 | 27.5 | 25.6 | 28.6 | 32.3 | 60.8 | 11.0 | 23.8 | 33.2 | 37.7 | 24.0 |
| • DDA | 32.2 | 33.1 | 32.0 | 14.6 | 16.4 | 16.6 | 24.4 | 20.0 | 25.5 | 17.2 | 52.2 | 3.2 | 35.7 | 41.8 | 45.4 | 27.2 |
| • Tent | 1.2 | 1.4 | 1.4 | 1.0 | 0.9 | 1.2 | 2.6 | 1.7 | 1.8 | 3.6 | 5.0 | 0.5 | 2.6 | 3.2 | 3.1 | 2.1 |
| • EATA | 0.3 | 0.3 | 0.3 | 0.2 | 0.2 | 0.5 | 0.9 | 0.8 | 0.9 | 1.8 | 3.5 | 0.2 | 0.8 | 1.2 | 0.9 | 0.9 |
| ResNet50 (GN) | 17.9 | 19.9 | 17.9 | **19.7** | 11.3 | 21.3 | 24.9 | 40.4 | **47.4** | 33.6 | 69.2 | 36.3 | 18.7 | 28.4 | 52.2 | 30.6 |
| • MEMO | 18.4 | 20.6 | 18.4 | 17.1 | 12.7 | 21.8 | 26.9 | **40.7** | 46.9 | 34.8 | 69.6 | 36.4 | 19.2 | 32.2 | 53.4 | 31.3 |
| • DDA | **42.5** | **43.4** | **42.3** | 16.5 | 19.4 | 21.9 | 26.1 | 35.8 | 40.2 | 13.7 | 61.3 | 25.2 | **37.3** | 46.9 | 54.3 | 35.1 |
| • Tent | 2.6 | 3.3 | 2.7 | 13.9 | 7.9 | 19.5 | 17.0 | 16.5 | 21.9 | 1.8 | 70.5 | 42.2 | 6.6 | 49.4 | 53.7 | 22.0 |
| • EATA | 27.0 | 28.3 | 28.1 | 14.9 | 17.1 | 24.4 | 25.3 | 32.2 | 32.0 | 39.8 | 66.7 | 33.6 | 24.5 | 41.9 | 38.4 | 31.6 |
| • SAR (ours) | $33.1_{\pm1.0}$ | $36.5_{\pm0.4}$ | $35.5_{\pm1.1}$ | $19.2_{\pm0.4}$ | $\mathbf{19.5}_{\pm1.2}$ | $\mathbf{33.3}_{\pm0.5}$ | $\mathbf{27.7}_{\pm4.0}$ | $23.9_{\pm5.1}$ | $45.3_{\pm0.4}$ | $\mathbf{50.1}_{\pm1.0}$ | $\mathbf{71.9}_{\pm0.1}$ | $\mathbf{46.7}_{\pm0.2}$ | $7.1_{\pm1.8}$ | $\mathbf{52.1}_{\pm0.5}$ | $\mathbf{56.3}_{\pm0.1}$ | $\mathbf{37.2}_{\pm0.6}$ |
| VitBase (LN) | 9.4 | 6.7 | 8.3 | 29.1 | 23.4 | 34.0 | 27.0 | 15.8 | 26.3 | 47.4 | 54.7 | 43.9 | 30.5 | 44.5 | 47.6 | 29.9 |
| • MEMO | 21.6 | 17.4 | 20.6 | 37.1 | 29.6 | 40.6 | 34.4 | 25.0 | 34.8 | 55.2 | 65.0 | 54.9 | 37.4 | 55.5 | 57.7 | 39.1 |
| • DDA | 41.3 | 41.3 | 40.6 | 24.6 | 27.4 | 30.7 | 26.9 | 18.2 | 27.7 | 34.8 | 50.0 | 32.3 | 42.2 | 52.5 | 52.7 | 36.2 |
| • Tent | 32.7 | 1.4 | 34.6 | 54.4 | 52.3 | 58.2 | 52.2 | 7.7 | 12.0 | 69.3 | 76.1 | 66.1 | 56.7 | 69.4 | 66.4 | 47.3 |
| • EATA | 35.9 | 34.6 | 36.7 | 45.3 | 47.2 | 49.3 | 47.7 | **56.5** | **55.4** | 62.2 | 72.2 | 21.7 | 56.2 | 64.7 | 63.7 | 49.9 |
| • SAR (ours) | $\mathbf{46.5}_{\pm3.0}$ | $\mathbf{43.1}_{\pm7.4}$ | $\mathbf{48.9}_{\pm0.4}$ | $\mathbf{55.3}_{\pm0.1}$ | $\mathbf{54.3}_{\pm0.2}$ | $\mathbf{58.9}_{\pm0.1}$ | $\mathbf{54.8}_{\pm0.2}$ | $53.6_{\pm7.1}$ | $46.2_{\pm3.5}$ | $\mathbf{69.7}_{\pm0.3}$ | $\mathbf{76.2}_{\pm0.1}$ | $\mathbf{66.2}_{\pm0.3}$ | $\mathbf{60.9}_{\pm0.3}$ | $\mathbf{69.6}_{\pm0.1}$ | $\mathbf{66.6}_{\pm0.1}$ | $\mathbf{58.0}_{\pm0.5}$ |

## Ablation Studies of SAR

- Reliable and Sharpness-aware entropy, in conjunction, yield stable TTA

Corruption Accuracy (%) on ImageNet-C (level 5) under online imbalanced label distribution shifts

| | Noise | | | Blur | | | | Weather | | | | Digital | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model+Method | Gauss. | Shot | Impul. | Defoc. | Glass | Motion | Zoom | Snow | Frost | Fog | Brit. | Contr. | Elastic | Pixel | JPEG | Avg. |
| ResNet50 (GN) | 3.2 | 4.1 | 4.0 | 17.1 | 8.5 | 27.0 | 24.4 | 17.9 | 25.5 | 2.6 | 72.1 | 45.8 | 8.2 | 52.2 | 56.2 | 24.6 |
| • reliable | 34.5 | 36.8 | 36.2 | 19.5 | 3.1 | 33.6 | 14.5 | 20.5 | 38.3 | 2.4 | 71.9 | 47.0 | 8.3 | 52.1 | 56.4 | 31.7 |
| • reliable+sa | 33.8 | 35.9 | 36.4 | 19.2 | 18.7 | 33.6 | 24.5 | 23.5 | 45.2 | 49.3 | 71.9 | 46.6 | 9.2 | 51.6 | 56.4 | **37.0** |
| • reliable+sa+reset | 33.6 | 36.1 | 36.2 | 19.1 | 18.6 | 33.9 | 24.7 | 22.5 | 45.7 | 49.0 | 71.9 | 46.6 | 9.2 | 51.5 | 56.3 | **37.0** |

## Efficiency Comparison

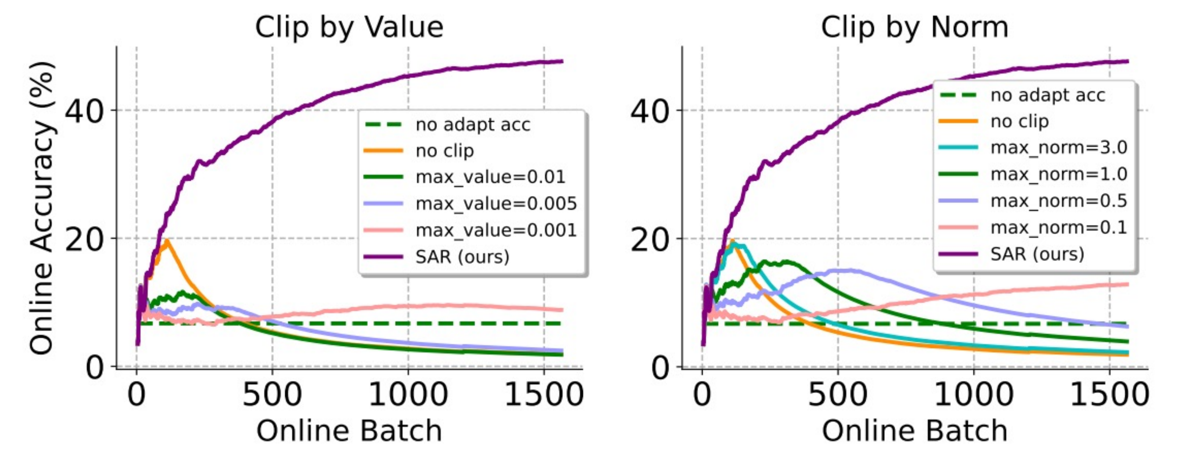
- While improving adaptation stability, our SAR maintains high efficiency

Time for processing 50,000 images (Gaussian noise, level 5 on ImageNet-C) via a single V100 GPU on ResNet50-GN

| Method | GPU time |
|---|---|
| MEMO (Zhang et al., 2022) | 55,980 secs |
| DDA (Gao et al., 2022) | 146,220 secs |
| TTT (Sun et al., 2020) | 3,600 secs |
| Tent (Wang et al., 2021) | 110 secs |
| EATA (Niu et al., 2022a) | 114 secs |
| SAR (ours) | 115 secs |

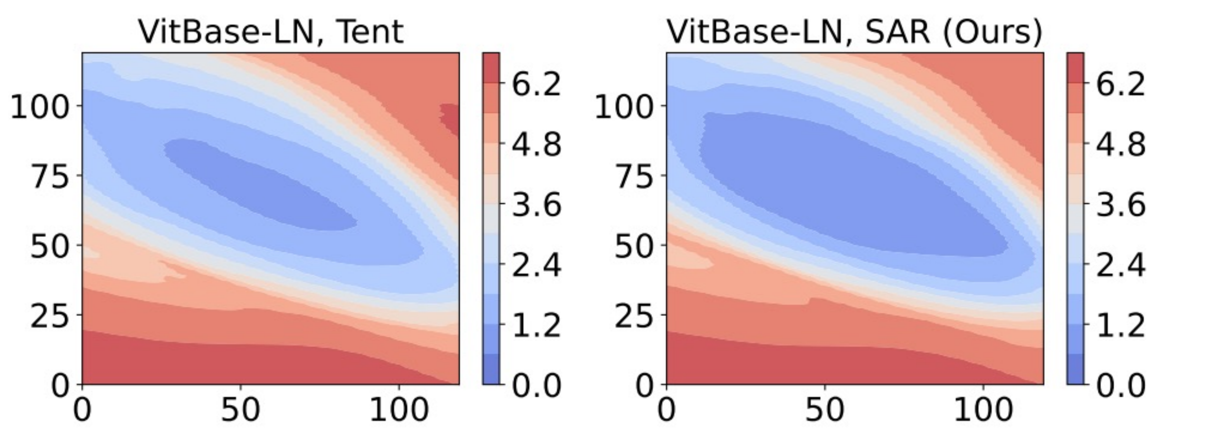
## Comparison with Gradient Clip

- Large $\delta$ of clip: cannot prevent model collapse
- Small $\delta$ of clip: leading to limited learning ability and biased gradient directions
- Our SAR does not need to tune such a parameter and yields better performance




## Entropy Surface Visualization

- Results on ImageNet-C (Gaussian noise, level 5)
- The area (the deepest blue) within the lowest loss contour line of our SAR is larger than Tent
- Our SAR has a flatter entropy surface, and thus is more insensitive to noisy updates




## Contact Information

**Code** `https://github.com/mr-eggplant/SAR`

**Email** Yifan Zhang
`yifan.zhangg@u.nus.edu`