# Towards Stable Test-Time Adaptation in Dynamic Wild World

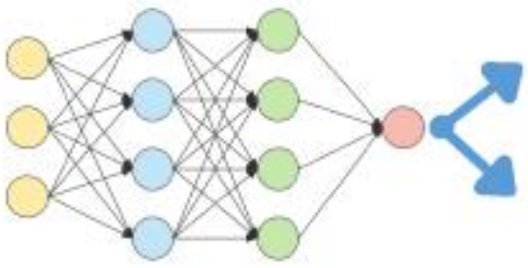
Shuaicheng Niu*, Jiaxiang Wu*, Yifan Zhang*, Zhiquan Wen, Yaofo Chen, Peilin Zhao and Mingkui Tan

South China University of Technology, Tencent AI Lab, National University of Singapore

# Contents

# Background: Test Data Shifts

- Deep models are often very sensitive when test samples encountering **natural variations or corruptions** (*also called distribution shifts*):
  - Weather change
  - Unexpected noises

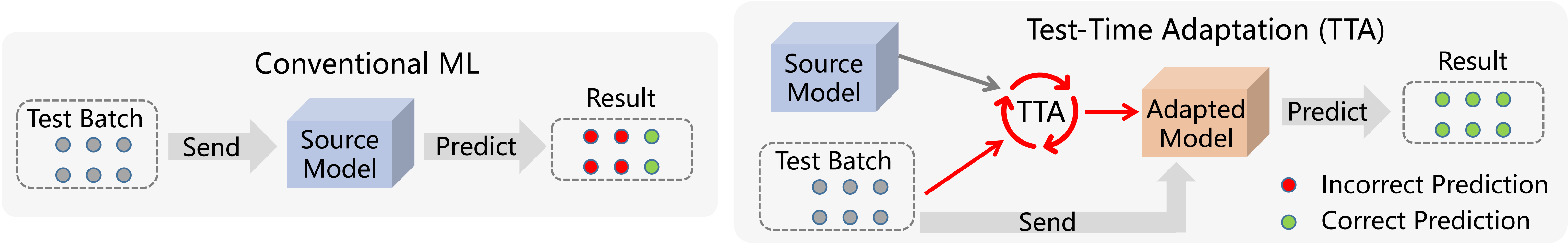



Day-Sunset-Night

ImageNet-C (Hendrycks et al., 2019)

# Test-Time Adaptation for Overcoming Data Shifts

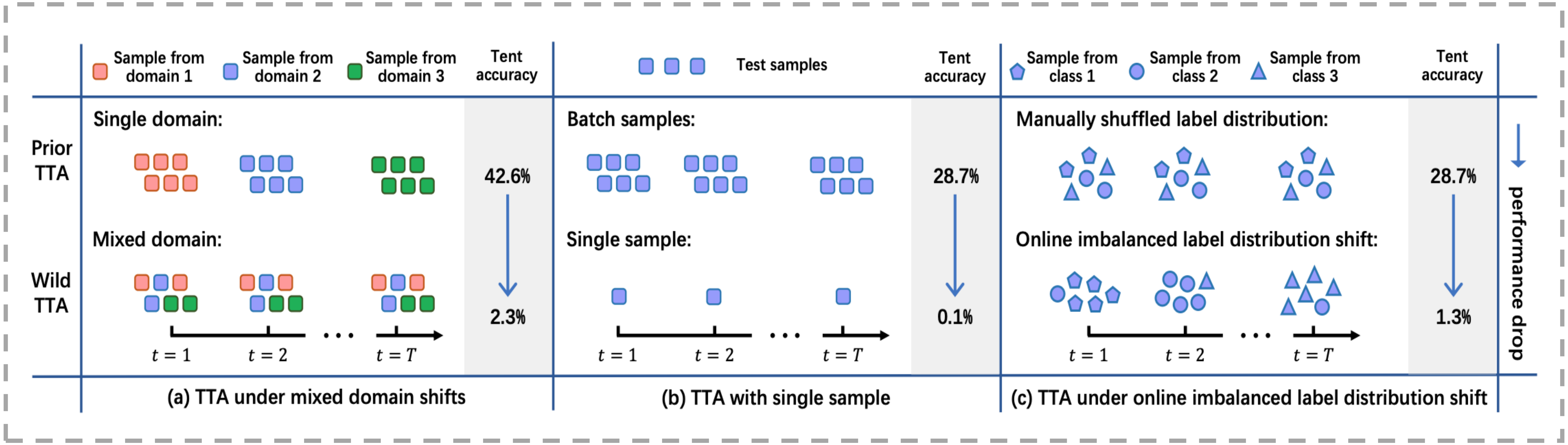
- Goal: TTA aims to adapt model to test-data domain before prediction
  - Adapt online with only unlabeled test data




The Figures are borrowed from *Uncovering Adversarial Risks of Test-Time Adaptation.*

# Problem: Test-Time Adaptation in Wild World

- **Limitation**: TTA is unstable under wild scenarios
  - severe performance degradation, or even model collapse




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

# Contents

# I: What Leads to Unstable TTA?

Batch Normalization (BN) is a crucial factor hindering TTA stability under the above wild test settings

$$y^{(k)} = \gamma^{(k)}\widehat{x}^{(k)} + \beta^{(k)}, \quad \text{where} \quad \widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}\left[x^{(k)}\right]}{\sqrt{\mathrm{Var}\left[x^{(k)}\right]}}.$$

BN statistics estimation would be inaccurate when test data stream has:

- Mixed Shifts: ideally each domain should have its own $\mathrm{E}$ and $\mathrm{Var}$
- Single Sample: it is hard to estimate $\mathrm{E}$ and $\mathrm{Var}$ accurately
- Online Imbalanced Label Shifts: will bias to some specific class

Our claim: models with batch-agnostic norm layers are more suitable for TTA

# I: What Leads to Unstable TTA?--Empirical Study

GN and LN models performs more stably than BN models (but still suffer several failure cases)

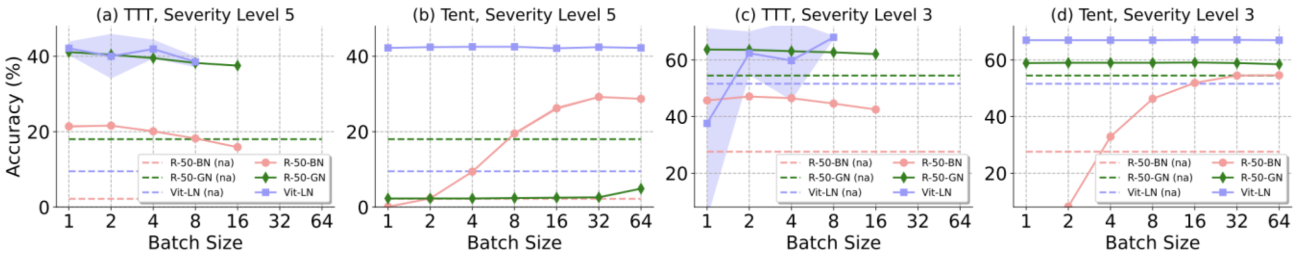
Methods:

- TTT (Sun et al., 2020)
- Tent (Wang et al., 2021)

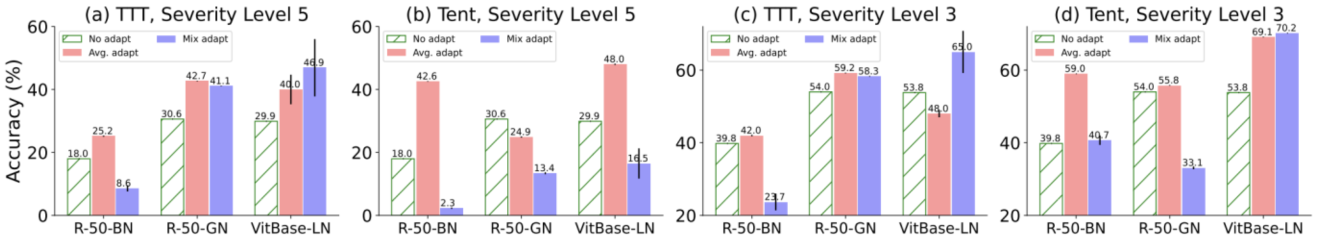
Norms:

- GN (group norm)
- LN (layer norm)
- BN (batch norm)

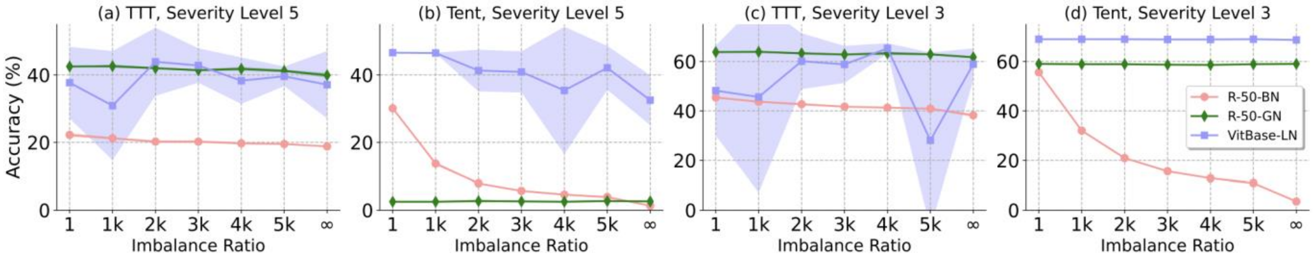
① TTA under small batch sizes




② TTA under mixed domain shifts

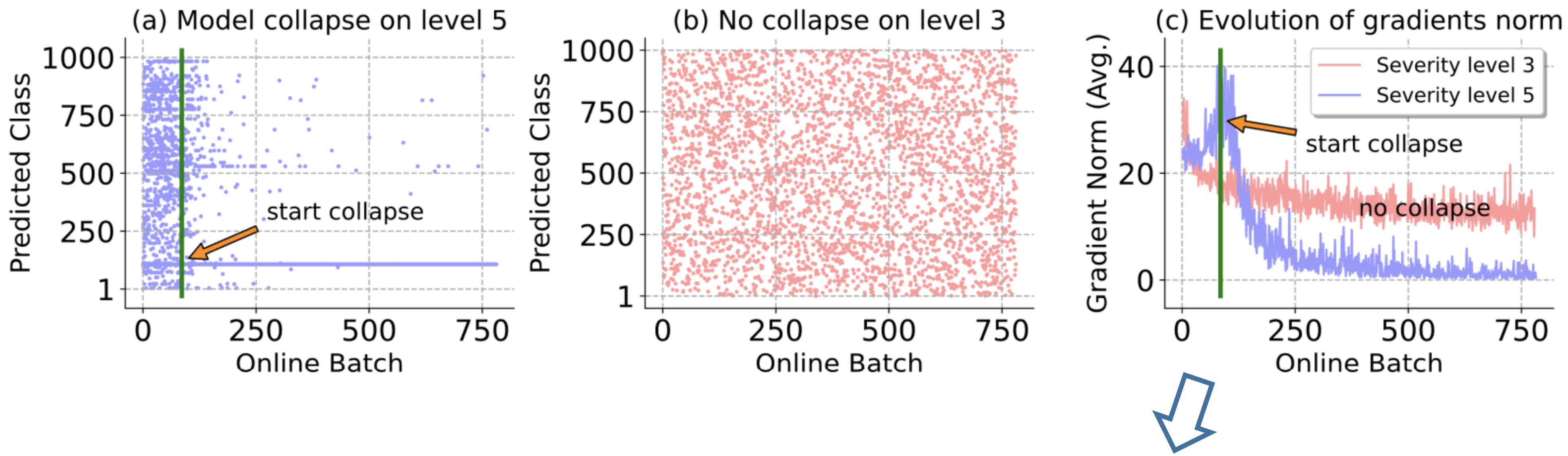



③ TTA under online imbalanced label shifts



# II: What Leads to Unstable TTA?

Online entropy minimization tends to result in collapsed trivial solutions, i.e., predict all samples to the same class



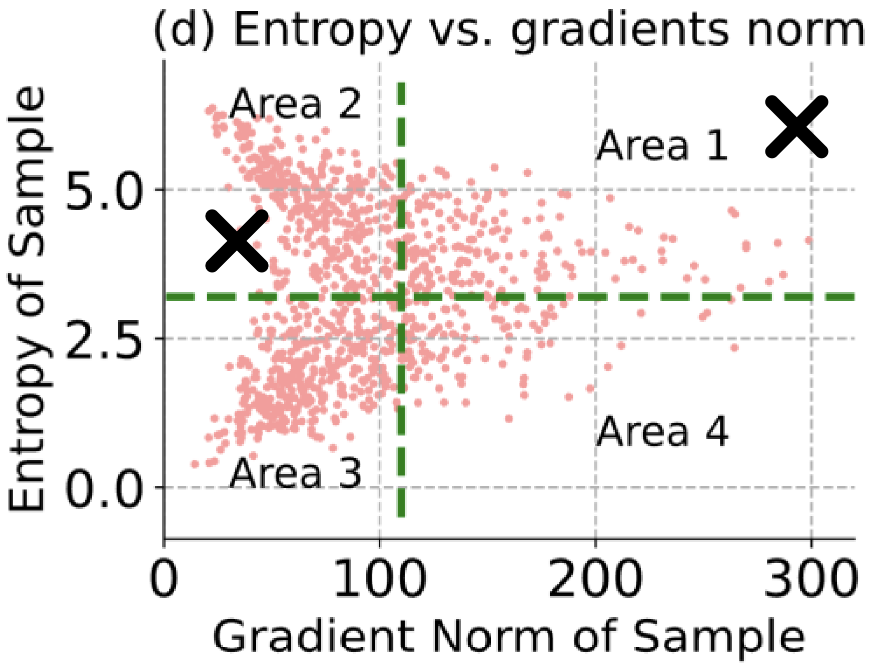

Some large/noisy gradients cause collapse

# Contents

# SAR: Sharpness-Aware and Reliable Entropy Minimization

Motivation:

- We find that removing noisy gradients via gradient norm filtering is infeasible, since its threshold is hard to select
- We instead use entropy for filtering, which is easier to select threshold

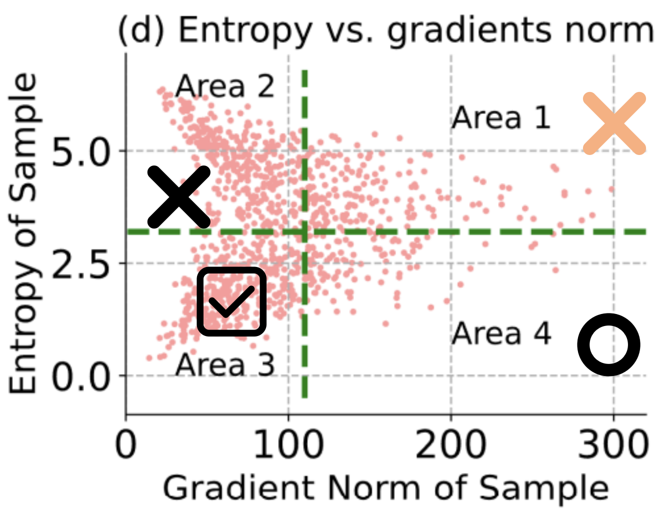


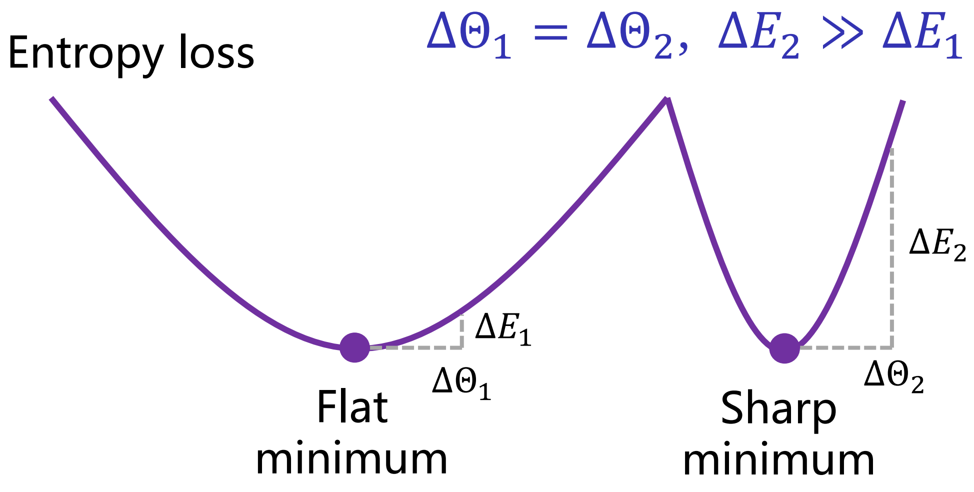

- **Reliable Entropy:**
  - Remove samples in Areas 1 (large gradients) and Area 2 (unconfident):

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

where the threshold $E_0 \in (0, \ln C]$, and $C$ is the class number

# SAR: Sharpness-Aware and Reliable Entropy Minimization




- **Sharpness-Aware:** make the model more robust to large/noisy gradients in Area 4

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

- We use SAM (Foret et al. 2021) to address the optimization, leading to the final objective:

$$\min_{\tilde{\Theta}} S(\mathbf{x})E^{SA}(\mathbf{x};\Theta)$$

The sharpness solution is inspired by Foret et al., Sharpness-aware minimization for efficiently improving generalization

# Contents

# Results under Online Imbalanced Label Distribution Shifts

- SAR achieves the best performance over ResNet50-GN and VitBase-LN
  - Compare to Tent, SAR leads to 15.2% gains on R-50-GN and 10.7% gain on Vit-B-LN

| | Noise | | | Blur | | | | Weather | | | | Digital | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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$_{\pm0.4}$ | **19.5**$_{\pm1.2}$ | **33.3**$_{\pm0.5}$ | **27.7**$_{\pm4.0}$ | 23.9$_{\pm5.1}$ | 45.3$_{\pm0.4}$ | **50.1**$_{\pm1.0}$ | **71.9**$_{\pm0.1}$ | **46.7**$_{\pm0.2}$ | 7.1$_{\pm1.8}$ | **52.1**$_{\pm0.5}$ | **56.3**$_{\pm0.1}$ | **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) | **46.5**$_{\pm3.0}$ | **43.1**$_{\pm7.4}$ | **48.9**$_{\pm0.4}$ | **55.3**$_{\pm0.1}$ | **54.3**$_{\pm0.2}$ | **58.9**$_{\pm0.1}$ | **54.8**$_{\pm0.2}$ | 53.6$_{\pm7.1}$ | 46.2$_{\pm3.5}$ | **69.7**$_{\pm0.3}$ | **76.2**$_{\pm0.1}$ | **66.2**$_{\pm0.3}$ | **60.9**$_{\pm0.3}$ | **69.6**$_{\pm0.1}$ | **66.6**$_{\pm0.1}$ | **58.0**$_{\pm0.5}$ |

# Efficiency Comparison and Ablations

- While improving adaptation stability, SAR maintains high efficiency

| Method | Need source data? | Online update? | #Forward | #Backward | Other computation | GPU time (50,000 images) |
|---|---|---|---|---|---|---|
| MEMO (Zhang et al., 2021) | ✗ | ✗ | 50,000×65 | 50,000×64 | AugMix (Hendrycks et al., 2020) | 933 minutes |
| DDA (Gao et al., 2022) | ✓ | ✗ | 50,000×2 | 0 | 50,000 diffusion | 2,435 minutes |
| TTT (Sun et al., 2020) | ✓ | ✓ | 50,000×21 | 50,000×20 | rotation augmentation | 61 minutes |
| Tent (Wang et al., 2021) | ✗ | ✓ | 50,000 | 50,000 | n/a | 110 seconds |
| EATA (Niu et al., 2022) | ✓ | ✓ | 50,000 + 26,196 | 26,196 | regularizer | 114 seconds |
| SAR (ours) | ✗ | ✓ | 50,000 + 12,710×2 | 12,710×2 | Eqn. (5) | 115 seconds |

- Visualization of entropy loss surface
  - SAR is flatter, and more insensitive to noisy gradients

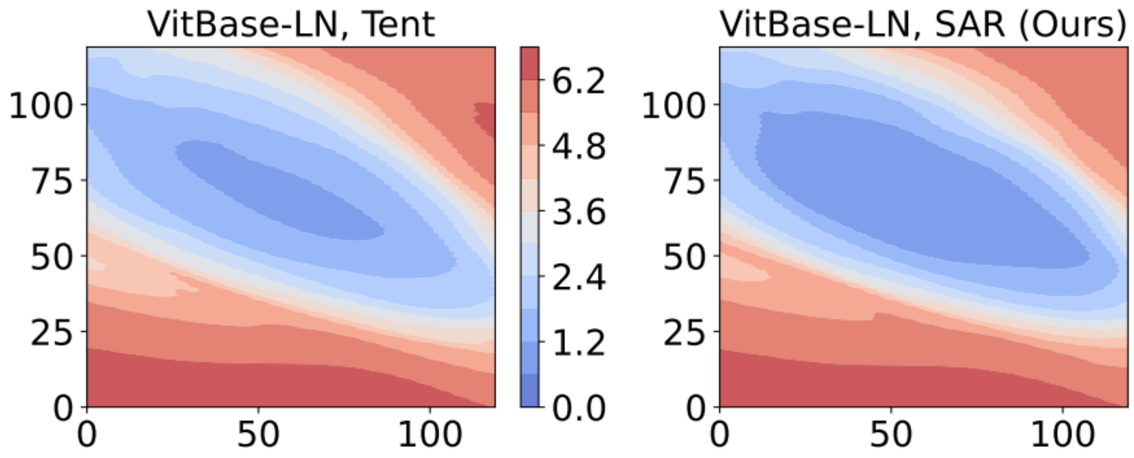



Figure. Loss (entropy) surface.

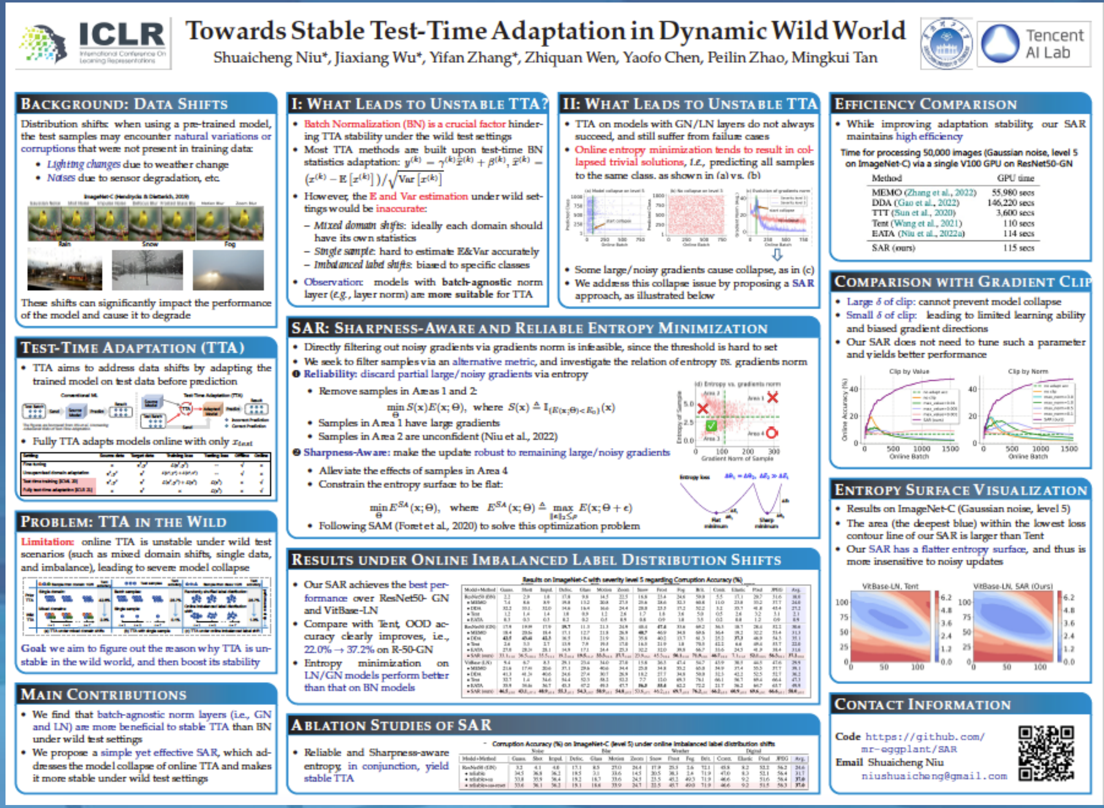
# Contents

# Conclusion

- We find that batch-agnostic norm layers (i.e., GN and LN) are more effective than BN for stable TTA under wild test settings
- We propose to use GN/LN models for stable TTA in the wild
- We further enhance the stability of online TTA for GN/LN models via a simple yet effective SAR method

**Please use our github repository:**

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




# Thank you!