

Towards Stable Test-Time Adaptation in Dynamic Wild World

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

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



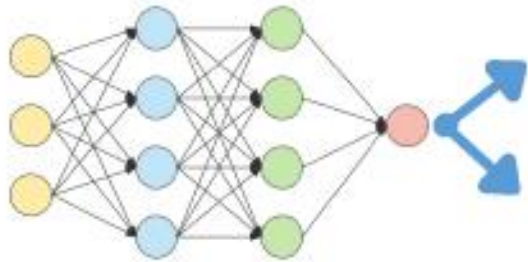
Contents

- 01** Background
- 02** Why Unstable Test-Time Adaptation?
- 03** Our method
- 04** Experimental Results
- 05** Conclusion



Background: Test Data Shifts

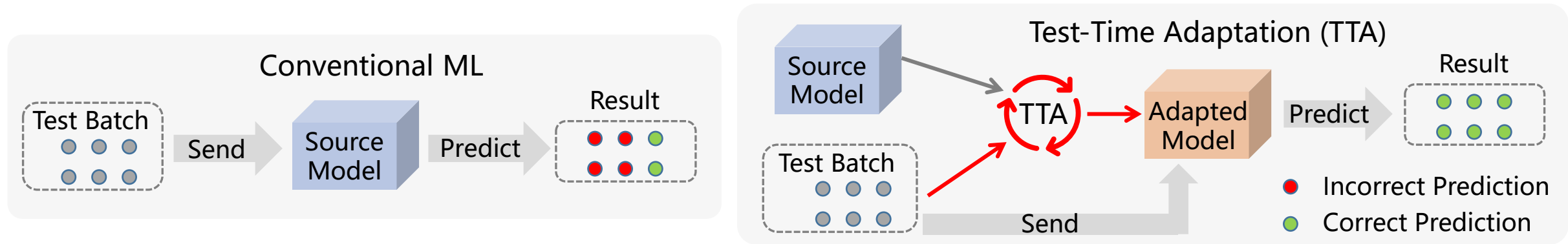
- Deep models are often very sensitive when test samples encountering **natural variations or corruptions** (*also called distribution shifts*):
 - Weather change
 - Unexpected noises



Day-Sunset-Night

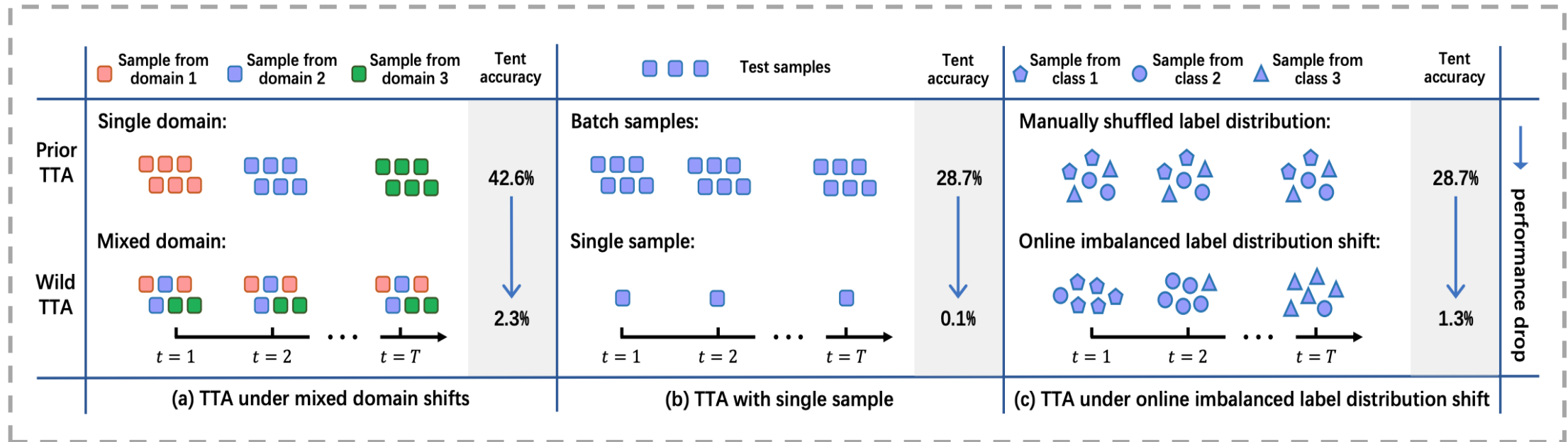
Test-Time Adaptation for Overcoming Data Shifts

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



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

- 01 Background
- 02 Why Unstable Test-Time Adaptation?**
- 03 Our method
- 04 Experimental Results
- 05 Conclusion



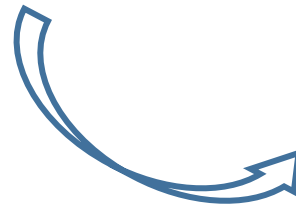
I: What Leads to Unstable TTA?

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

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

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

- **Mixed Shifts:** ideally each domain should have its own \mathbb{E} and Var
- **Single Sample:** it is hard to estimate \mathbb{E} and 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)

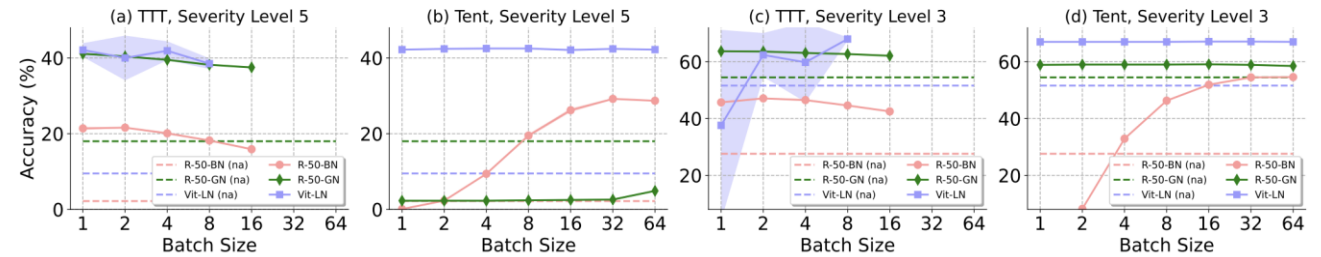
Methods:

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

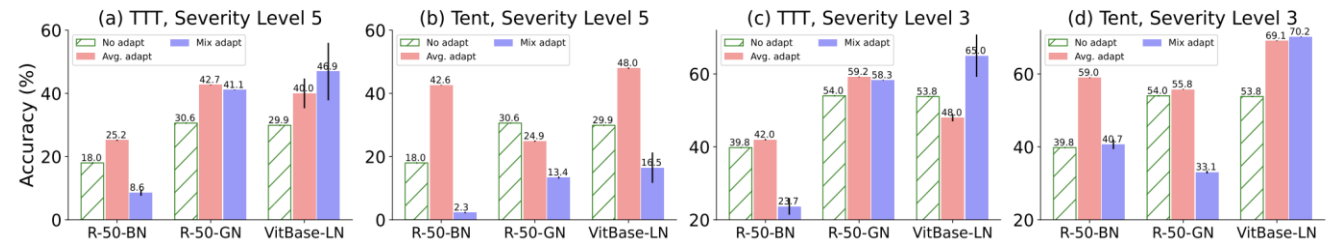
Norms:

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

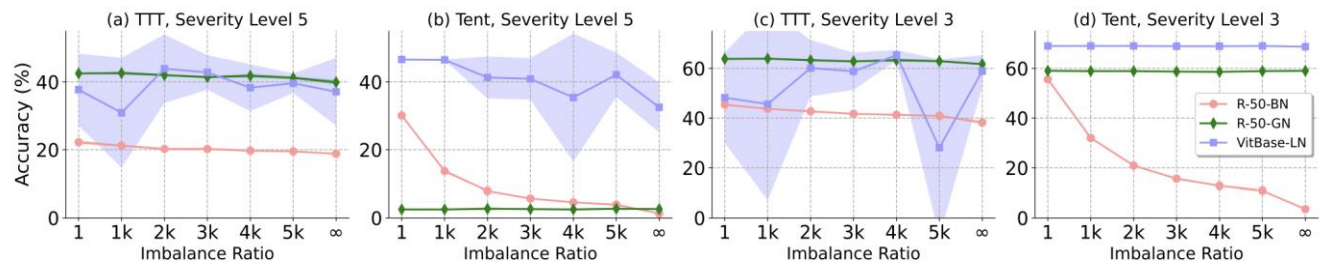
① TTA under small batch sizes



② TTA under mixed domain shifts

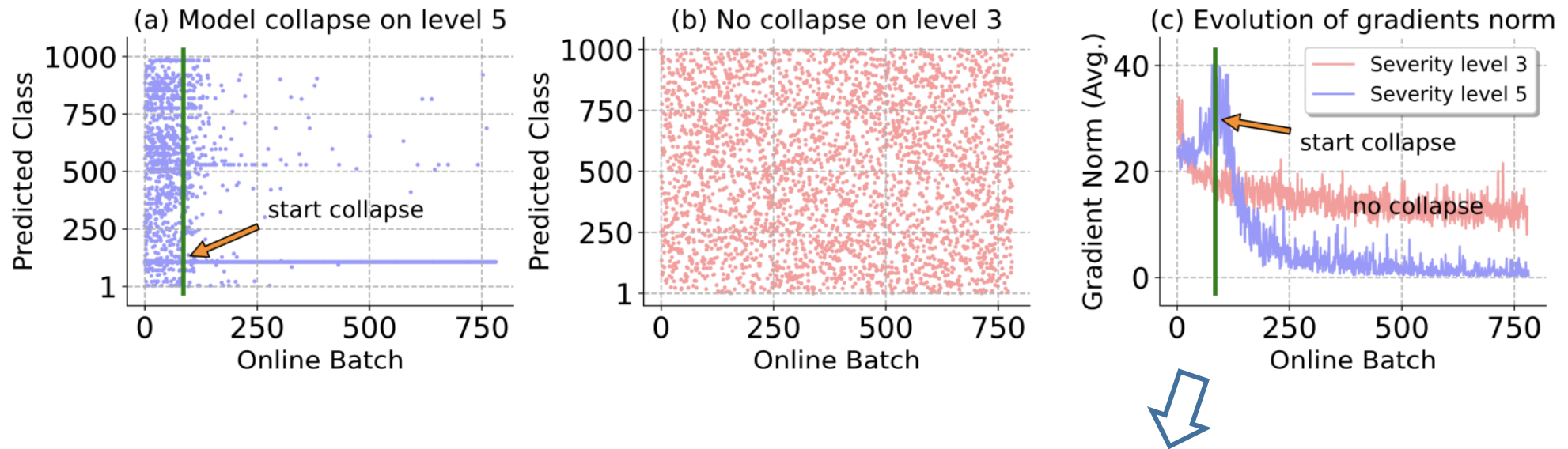


③ TTA under online imbalanced label shifts



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

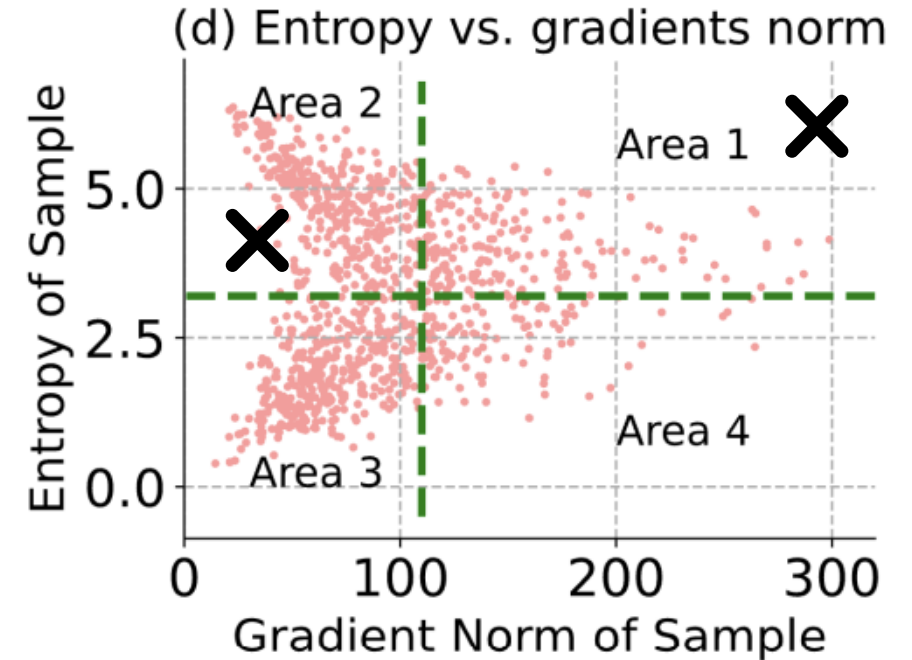
- ◆ 01 Background
- ◆ 02 Why Unstable Test-Time Adaptation?
- ◆ 03 **Our method**
- ◆ 04 Experimental Results
- ◆ 05 Conclusion



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



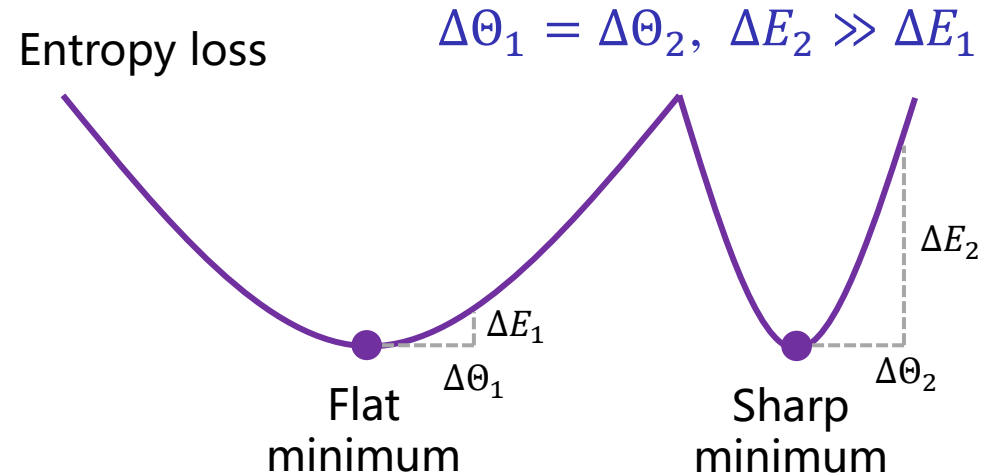
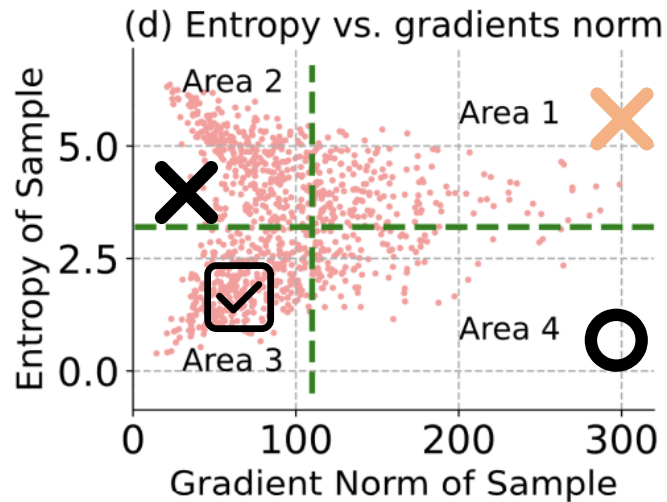
- **Reliable Entropy:**

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

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

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

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

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

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

Contents

- ◆ 01 Background
- ◆ 02 Why Unstable Test-Time Adaptation?
- ◆ 03 Our method
- ◆ **04 Experimental Results**
- ◆ 05 Conclusion



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

Model+Method	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
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)	✗	✗	$50,000 \times 65$	$50,000 \times 64$	AugMix (Hendrycks et al., 2020)	933 minutes
DDA (Gao et al., 2022)	✓	✗	$50,000 \times 2$	0	50,000 diffusion	2,435 minutes
TTT (Sun et al., 2020)	✓	✓	$50,000 \times 21$	$50,000 \times 20$	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)	✗	✓	$50,000 + 12,710 \times 2$	$12,710 \times 2$	Eqn. (5)	115 seconds

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

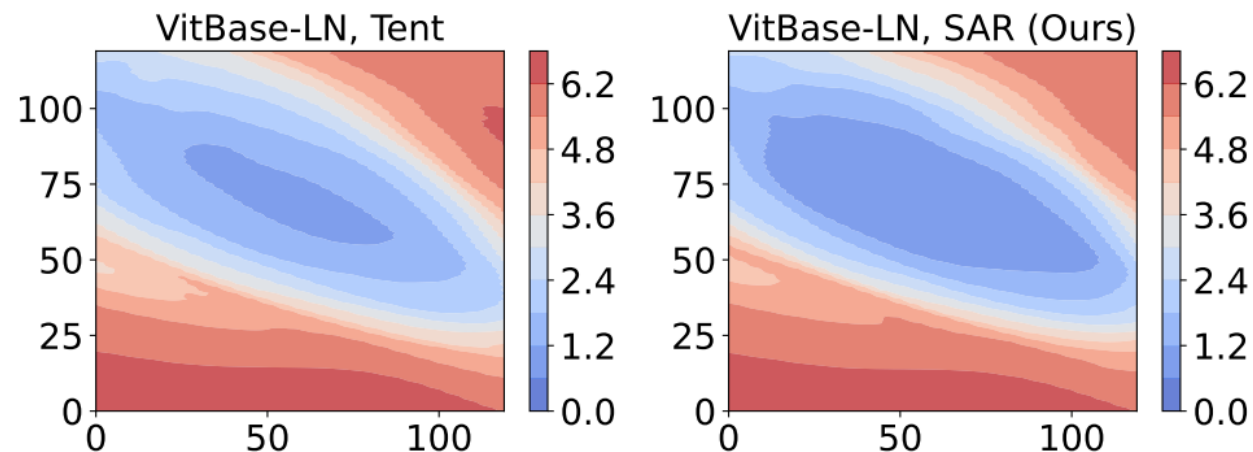


Figure. Loss (entropy) surface.

Contents

- ◆ 01 Background
- ◆ 02 Why Unstable Test-Time Adaptation?
- ◆ 03 Our method
- ◆ 04 Experimental Results
- ◆ 05 **Conclusion**




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>



ICLR
International Conference on
Learning Representations

Towards Stable Test-Time Adaptation in Dynamic Wild World

Shuaicheng Niu*, Jiaxiang Wu*, Yifan Zhang*, Zhiqian Wen, Yaofu Chen, Peilin Zhao, Mingkui Tan




Tencent AI Lab

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.



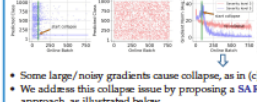
These shifts can significantly impact the performance of the model and cause it to degrade.

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: $y^{(k)} = \gamma^{(k)} \frac{x^{(k)}}{\sigma^{(k)}} + \beta^{(k)}$, $\hat{x}^{(k)} = (x^{(k)} - \mathbb{E}[x^{(k)}]) / \sqrt{\text{Var}[x^{(k)}]}$
- 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. E&Var accurately
 - Imbalanced label shifts: biased by 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

EFFICIENCY COMPARISON


While improving adaptation stability, our SAR maintains high efficiency.

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

Method	GPU time
MEMO (Zhang et al., 2022)	55,980 secs
DDA (Guo et al., 2022)	146,220 secs
TTT (Guo et al., 2020)	3,660 secs
Tent (Wang et al., 2021)	110 secs
EADA (Niu et al., 2022a)	114 secs
SAR (ours)	115 secs

TEST-TIME ADAPTATION (TTA)


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



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

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.

RESULTS UNDER ONLINE IMBALANCED LABEL DISTRIBUTION SHIFTS

Our SAR achieves the best performance over ResNet50-GN and ViTBase-LN.

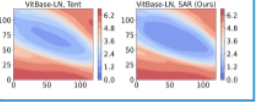
Compare with Tent, OOD accuracy clearly improves, i.e., 22.0% \rightarrow 37.2% on R-50-GN.

Entropy minimization on LN/GN models performs better than that on BN models.

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.



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.


ABLATION STUDIES OF SAR

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

CONTACT INFORMATION

Code: <https://github.com/mr-eggplant/SAR>

Email: shuaichengniu@tencent.com



Thank you!