Test-time Adaptation against Multi-modal Reliability Bias

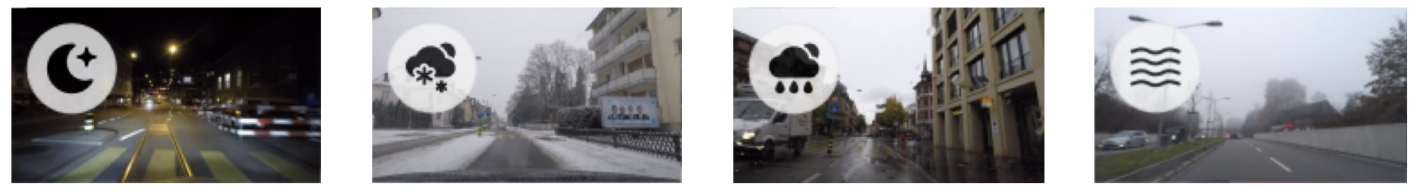
Mouxing Yang[1], Yunfan Li[1], Changqing Zhang[2], Peng Hu[1], Xi Peng[1]*

[1] Sichuan University [2] Tianjin University *Corresponding Author

Background: Distribution Shifts

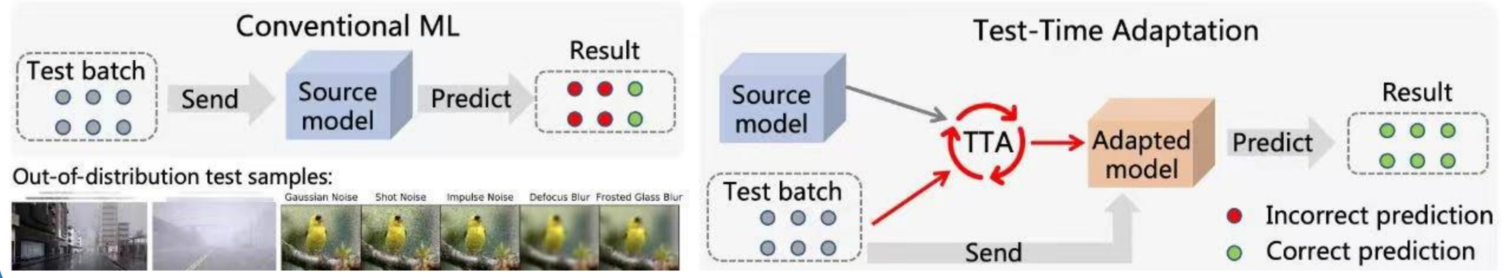
Out of distribution (OOD): the pre-trained models always encounter performance degeneration when the distribution shifts between training and test data emerge in the scenarios of

- Changing weather, e.g., fog, snow
- Degenerated sensors, e.g. defocus, gaussian noise

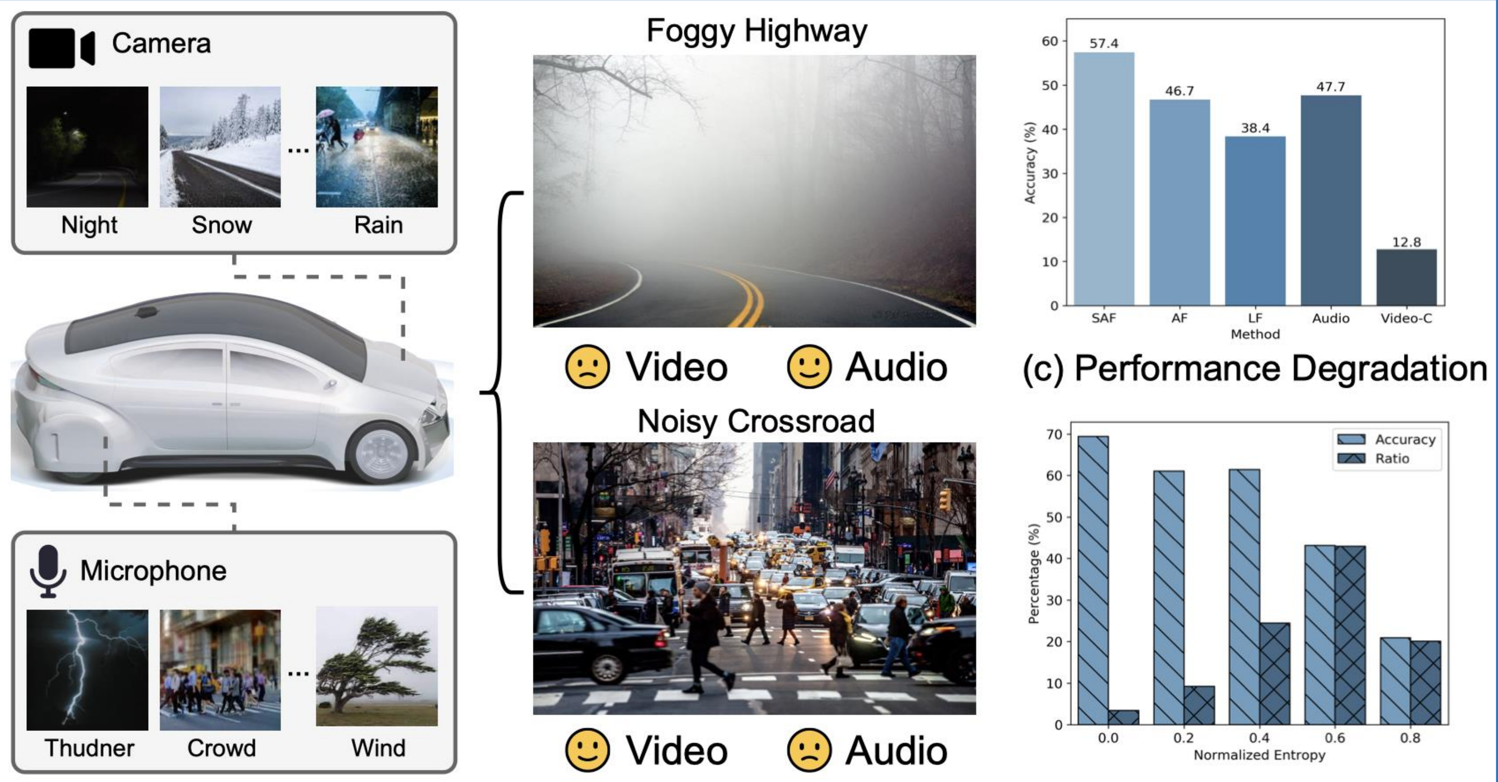


Test-time Adaptation (TTA)

TTA paradigm aims to bridge the gap between domains. To this end, most TTA methods usually work by minimizing the entropy-based objective on the model predictions of unlabeled test samples and updating the normalization layers (LN/BN).



Observations & Motivations



(a) Distribution Shift (b) Multi-modal Reliability Bias (c) Performance Degradation (d) Unstable Entropy

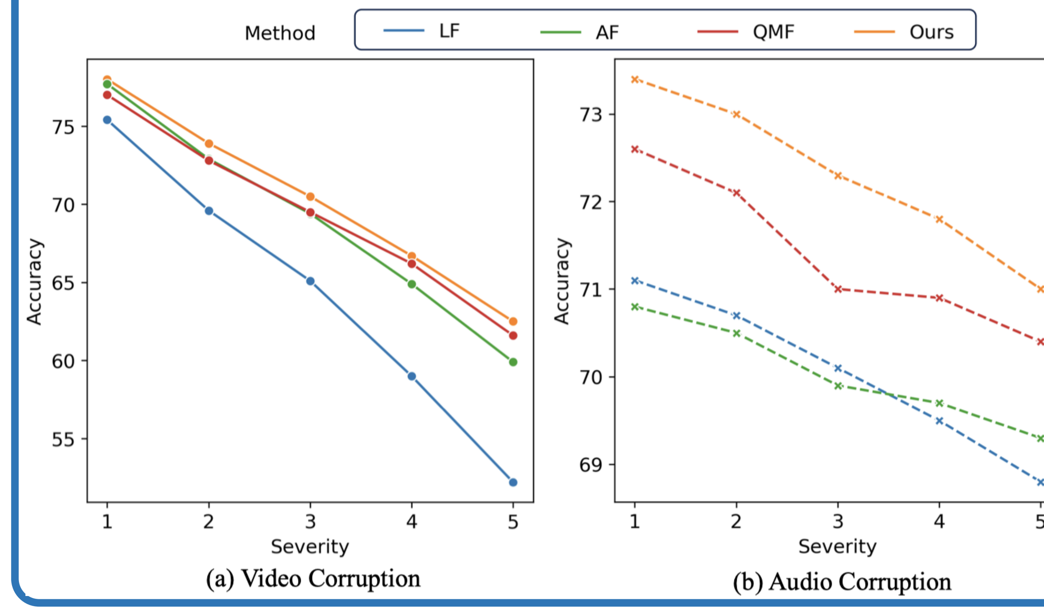
(a) Some modalities will face the distribution shift.

(b) The shifted modalities will lose the task-specific information and suffer from **modality reliability bias.**

(c) Vanilla cross-modal fusion manner with biased modalities ***will give inaccurate predictions***.

(d) The ratio of confident predictions would decrease while the noise might dominate the predictions.

Technical Challenges

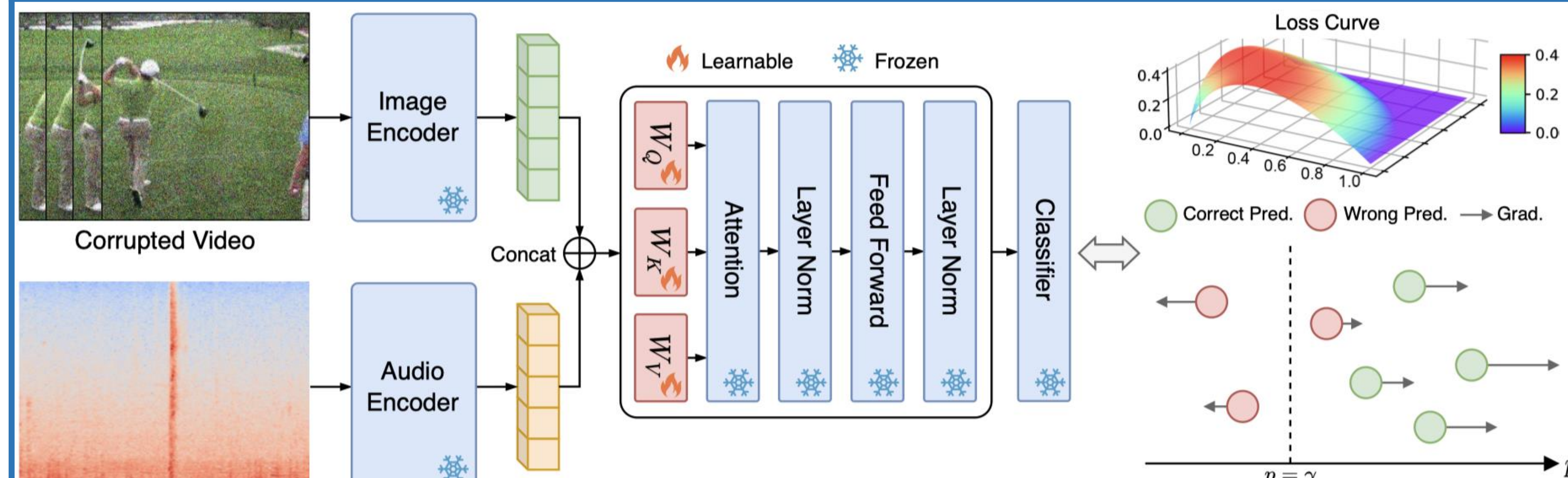


(a) Video Corruption (b) Audio Corruption

Q1: How to achieve the reliable cross-modal fusion for the test stream with modality reliability bias?

Q2: How to achieve robust cross-domain adaptation upon the predictions with heavy noise?

Method: REliable fusion and robust ADaptation (READ)



A1: **Reliable fusion via *Self-adaptive Attention-based Fusion* (SAF) mechanism**. Unlike the existing TTA methods that update LN or BN within the pre-trained models, we propose modulating the attention layer in a self-adaptive manner to achieve the reliable cross-modal information fusion.

A2: **Robust adaptation via a confidence-aware loss function.** The loss will reduce non-monotonously for different predictions. As a result, the high-confident predictions will contribute to optimization while the influence of low-confident predictions will be eliminated.

$$\mathcal{L}_{ra} = \frac{1}{B}\sum_{i=1}^{B} p_i \log\left(\frac{e\gamma}{p_i}\right)$$

Highlights & Contributions

- We reveal a **new problem** for test-time adaptation in multi-modal scenarios, i.e., *modality reliability bias*.
- Extensive experiment results demonstrate that it is **intractable to conquer the modality reliability bias problem using the existing TTA methods and cross-modal fusion mechanisms**.
- We **contribute two benchmarks** (multi-modal action recognition and event classification) for multi-modal TTA with reliability bias.

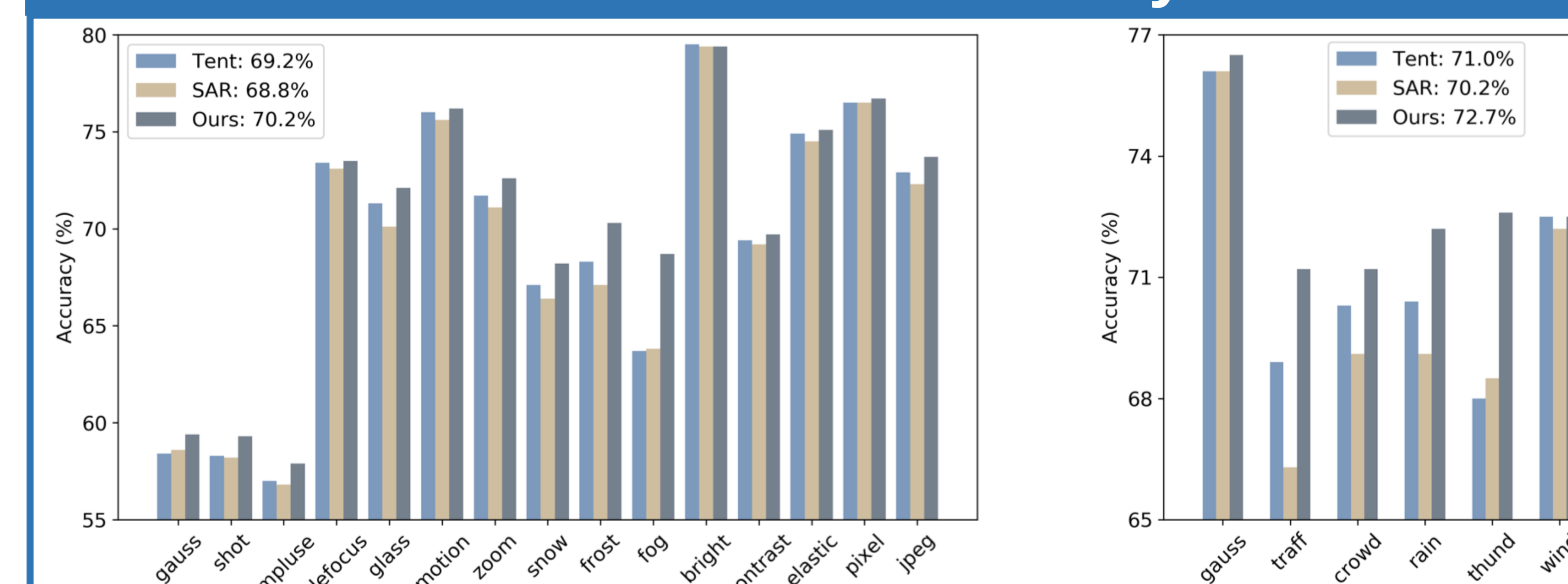
Results under Corrupted Modalities

Methods	Noise: Gauss.	Shot	Impul.	Blur: Defoc.	Glass	Mot.	Zoom	Weather: Snow	Frost	Fog	Brit.	Digital: Contr.	Elas.	Pix.	JPEG	Avg.
Source ((Stat. LN) & LF)	31.8	33.4	31.7	64.0	54.3	67.5	61.9	50.9	54.8	38.4	72.3	44.0	60.2	61.7	56.4	52.2
• MM-TTA (Dyn. LN)	46.2	46.6	46.1	58.8	55.7	62.6	58.7	52.6	54.4	48.5	69.1	49.3	57.6	56.4	54.6	54.5
• Tent (Dyn. LN)	28.6	29.8	28.3	63.4	51.1	67.7	61.7	46.5	51.3	24.5	72.3	38.6	60.7	61.8	54.9	49.4
• EATA (Dyn. LN)	31.8	33.3	31.6	64.2	54.6	67.7	62.2	51.3	54.7	38.1	72.5	44.2	60.4	62.0	57.0	52.4
• SAR (Dyn. LN)	31.9	33.3	31.7	63.8	54.0	67.7	61.8	50.7	54.5	38.8	72.3	44.0	60.3	62.0	56.5	52.2
• READ (Dyn. LN)	34.0	34.5	33.8	65.3	57.7	68.7	64.9	56.1	57.5	41.1	73.2	48.7	62.9	64.6	59.2	54.8
Source (Stat. (LN&AF))	46.8	48.0	46.9	67.5	62.2	70.8	66.7	61.6	60.3	46.7	75.2	52.1	65.7	66.5	61.9	59.9
• Tent (Dyn. LN)	46.3	47.0	46.3	67.2	62.5	71.0	67.6	63.1	61.1	34.9	75.4	51.6	66.8	67.2	62.7	59.4
• EATA (Dyn. LN)	46.8	47.6	47.1	67.2	62.7	70.6	67.2	62.3	60.9	46.7	75.2	52.4	65.9	66.8	62.5	60.1
• SAR (Dyn. LN)	46.7	47.4	46.8	67.0	61.9	70.4	66.4	61.8	60.6	46.0	75.2	52.1	65.7	66.4	62.0	59.8
• READ (SAF)	**49.4**	**49.7**	**49.0**	**68.0**	**65.1**	**71.2**	**69.0**	**64.5**	**64.4**	**57.4**	**75.5**	**53.6**	**68.3**	**68.0**	**65.1**	**62.5**

Methods	Noise: Gauss.	Traff.	Crowd.	Weather: Rain	Thund.	Wind	Avg.	Noise: Gauss.	Traff.	Crowd.	Weather: Rain	Thund.	Wind	Avg.
Source ((Stat. LN) & LF)	71.1	67.8	67.4	67.4	70.6	68.6	68.8	29.5	17.1	22.6	17.3	33.7	20.6	23.5
• MM-TTA (Dyn. LN)	70.8	69.2	68.5	69.0	69.8	69.4	69.4	14.1	5.2	6.4	6.9	8.6	4.5	7.6
• Tent (Dyn. LN)	71.1	68.6	67.8	67.4	71.2	68.9	69.2	6.4	2.1	2.9	1.9	9.5	3.1	4.3
• EATA (Dyn. LN)	71.2	67.9	67.5	67.8	70.9	68.7	69.0	28.8	17.1	22.4	17.4	33.8	20.4	23.3
• SAR (Dyn. LN)	71.1	67.5	67.4	67.4	70.6	68.6	68.8	28.5	16.6	22.4	17.4	33.7	20.2	23.1
• READ (Dyn. LN)	71.3	68.5	68.5	68.4	71.8	69.0	69.6	36.4	25.3	28.9	27.3	35.6	26.6	30.0
Source (Stat. (LN&AF))	73.7	65.5	67.9	70.3	67.9	70.3	69.3	37.0	25.5	16.8	21.6	27.3	25.5	25.6
• Tent (Dyn. LN)	73.9	67.4	69.2	70.4	66.5	70.5	69.6	10.6	2.6	1.8	2.8	5.3	4.1	4.5
• EATA (Dyn. LN)	73.7	66.1	68.5	70.3	67.9	70.1	69.4	39.2	26.1	22.9	26.0	31.7	30.4	29.4
• SAR (Dyn. LN)	73.7	65.4	68.2	69.9	67.2	70.2	69.1	37.4	9.5	11.0	12.1	26.8	23.7	20.1
• READ (SAF)	**74.1**	**69.0**	**69.7**	**71.1**	**71.8**	**70.7**	**71.1**	**40.4**	**28.9**	**26.6**	**30.9**	**36.7**	**30.6**	**32.4**

- TTA methods using late fusion are most sensitive to the reliability bias.
- The attention-based fusion can slightly improve the robustness.
- The proposed loss function can improve both late fusion and attention-based fusion. The proposed SAF with the loss could guarantee noise-resistant thus learning reliable attention for fusion.

Results under Mixed Severity Levels



(a) Video Corruption (b) Audio Corruption

Contact Information

Feel free to drop me an email.
yangmouxing@gmail.com

Code Link Chinese Blog WeChat

The code/benchmarks is available.

https://github.com/XLearning-SCU/2024-ICLR-READ