SLT 2022 · Languages of the world · 9th to 12th January 2023 · Doha, Qatar · Poster ID 1-1-18-MMP

# Push-Pull: Characterizing the Adversarial Robustness for Audio-Visual Active Speaker Detection


Xuanjun Chen [1*]  Haibin Wu [12*]  Helen Meng [2†]  Hung-Yi Lee [1†]  Jyh-Shing Roger Jang [1†]

[1]National Taiwan University  [2]The Chinese University of Hong Kong


## Introduction

### Audio-Visual Active Speaker Detection (AVASD)

- **Goal:** Determine if visible person in the video is speaking.
- **TalkNet [5]:** One of SOTA models for AVASD, which is shown in Figure 1 (a).
- **Applications:** An indispensable front-end for several applications, such as user authentication.
- **Challenges:** The adversarial robustness of AVASD models hasn't been investigated.

### Takeaways

- We first expose that AVASD models are highly susceptible to multi-modal adversarial attacks.
- We propose the audio-visual interaction loss (AVIL) to enlarge the inter-class difference and intra-class similarity, resulting in more robust AVASD models.
- The AVIL outperforms the adversarial training by 33.14% mAP (%) under multi-modal attacks.

Paper  Xuanjun Chen  Haibin Wu

## Multi-Modal Adversarial Attacks

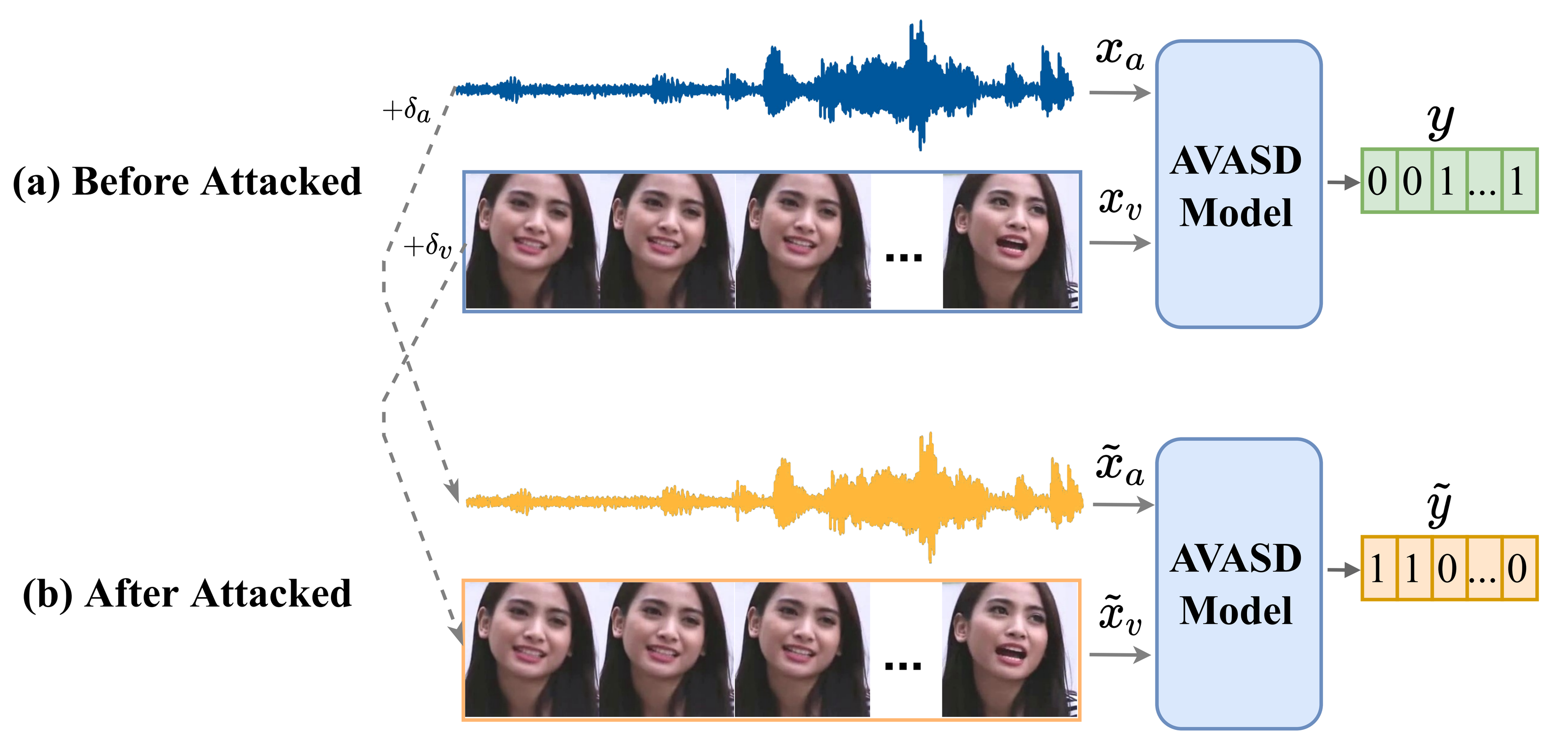



Figure 1. The multi-modal adversarial attack framework.

### Multi-Modal Adversarial Attacks Objective Function

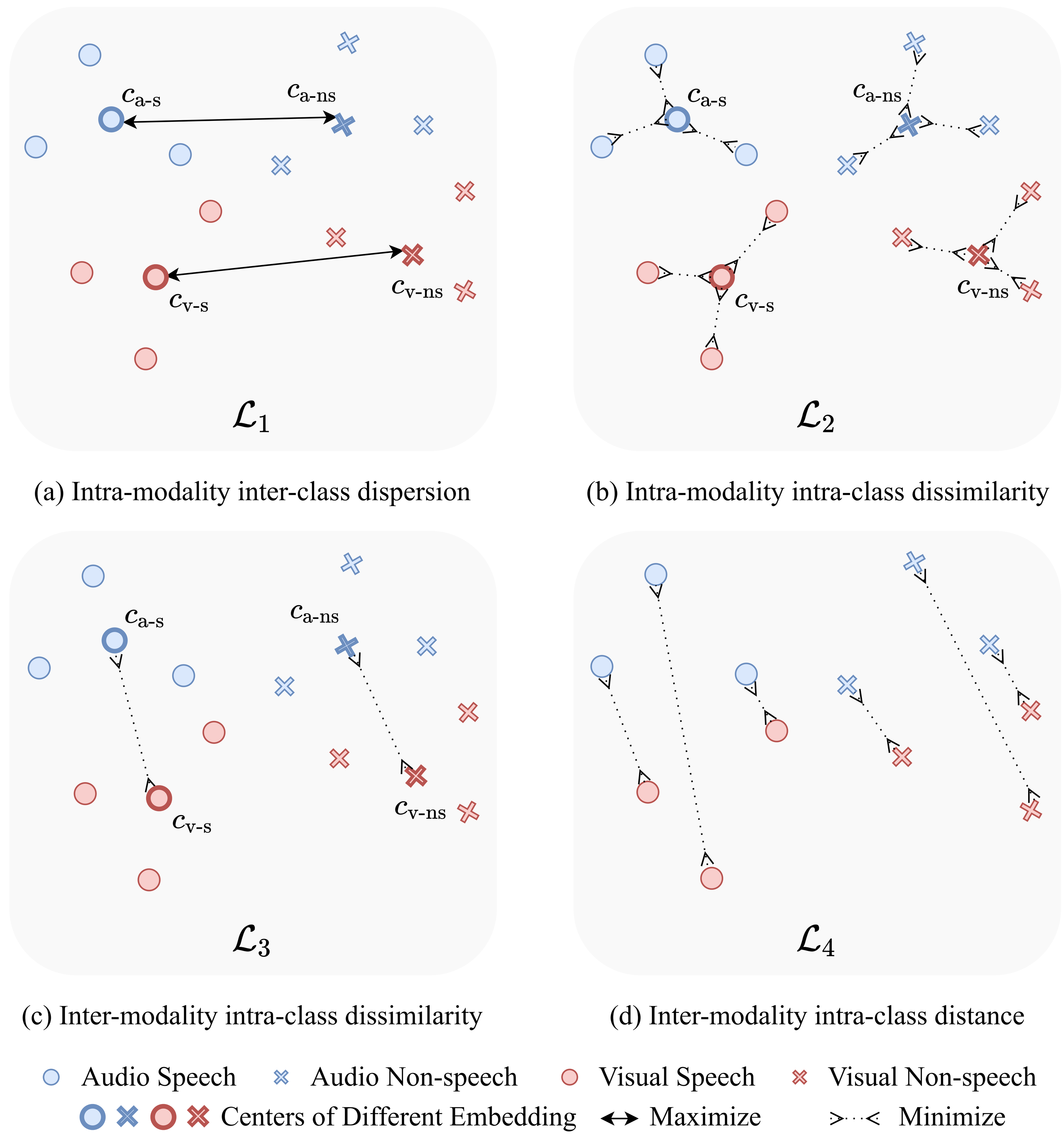
- **Goal:** Generate some imperceptible perturbation to fool model into making wrong predictions.
- **Perturbation:** maximize cross entropy loss $\mathcal{L}_{CE_{all}}$ difference between $y$ and $\tilde{y}$ via function:

$$\arg\max_{\delta_a,\delta_v} \mathcal{L}_{CE_{all}}(\tilde{x}_a, \tilde{x}_v, y), s.t. \ ||\delta_a||_p \le \epsilon_a, \quad ||\delta_v||_p \le \epsilon_v,$$

### Notations

- $\mathcal{L}_{CE_{all}}$ contains $\mathcal{L}_{CE_a}$, $\mathcal{L}_{CE_v}$, $\mathcal{L}_{CE_{av}}$, which corresponding to different prediction classifiers.
- $x_a$ and $x_v$ are the audio and visual samples, $y$ is ground-truth for the input $\{x_a, x_v\}$.
- $\delta_a$ and $\delta_v$ are the adversarial perturbations for $x_a$ and $x_v$; $||\cdot||_p$ is the $p$-norm.
- $\epsilon_{av}$, $\epsilon_a$, $\epsilon_v$ are attack budget: $\epsilon_a = \epsilon_{av} \times 10^{-4}$ and $\epsilon_v = \epsilon_{av} \times 10^{-1}$.
- $\tilde{y}$ is the prediction for the adversarial samples $\{\tilde{x}_a, \tilde{x}_v\}$.

## Attacks Defense by Audio-Visual Interaction Loss (AVIL)




(a) Intra-modality inter-class dispersion

(b) Intra-modality intra-class dissimilarity

(c) Inter-modality intra-class dissimilarity

(d) Inter-modality intra-class distance

Figure 2. The Audio-Visual Interaction Loss.

### Training Objective Function

- Summing cross entropy loss $\mathcal{L}_{CE_{all}}$ (i.e., $\mathcal{L}_{CE_a}$, $\mathcal{L}_{CE_v}$, $\mathcal{L}_{CE_{av}}$) and AVILs during training

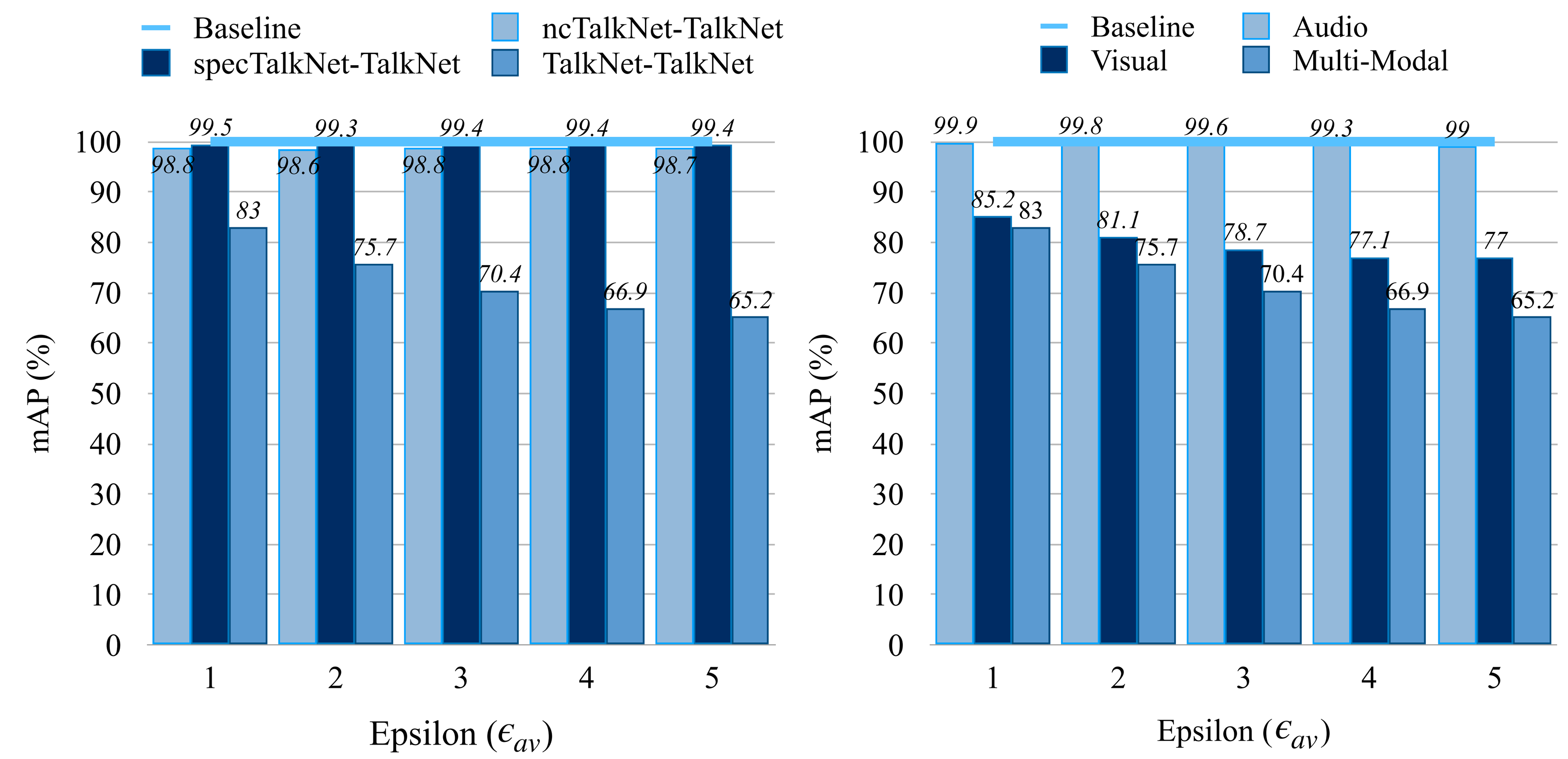
### Rationale of AVILs

- Minimizing $\mathcal{L}_1$ will equip the model with better discrimination capacity between speech and non-speech embeddings, resulting in higher inter-class differences.
- Maximizing $\mathcal{L}_2$, $\mathcal{L}_3$ and minimizing $\mathcal{L}_4$ will force the model to render compact intra-class features.
- Incorporating $\mathcal{L}_1$-$\mathcal{L}_4$ in the training process, we can simultaneously urge the model to learn both discriminative inter-class features, and compact intra-class features, leading the model less susceptible to adversarial attacks.

## Experimental Setup

- **Dataset:** AVA-ActiveSpeaker [4]; **Evaluation Metric:** Mean average precision (mAP (%)).
- **Black-box attacker:** specTalkNet, ncTalkNet; **White-box attacker:** TalkNet.

## Experiment




(a) Black-box attacker V.S. White-box attacker

(b) Single-modal attack V.S. Multi-modal attack

Figure 3. Adversarial attack performance of AVASD models under PGD [3] method.

| | Model | Adversarial training [2] | Clean mAP (%) | MIM [1] mAP (%) | PGD [3] mAP (%) |
|---|---|---|---|---|---|
| (A) | $\mathcal{L}_{CE_{all}}$ | ✗ | 92.58 | 49.30 | 47.79 |
| (B1) | $\mathcal{L}_{CE_{all}}$ | MIM | 91.34 | 52.18 | 54.23 |
| (B2) | $\mathcal{L}_{CE_{all}}$ | PGD | 91.68 | 58.3 | 56.06 |
| (D1) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_2$ | ✗ | 92.46 | 67.89 | 64.11 |
| (D2) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_3$ | ✗ | 92.20 | 47.92 | 49.27 |
| (D3) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$ | ✗ | 91.81 | 93.34 | 93.15 |
| (D4) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_2 + \mathcal{L}_3$ | ✗ | 92.27 | 63.36 | 61.54 |
| (D5) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_2 + \mathcal{L}_4$ | ✗ | 91.93 | 66.28 | 67.75 |
| (D6) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_3 + \mathcal{L}_4$ | ✗ | 91.70 | 92.48 | 91.01 |
| (E1) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$ | MIM | 91.70 | 99.98 | 99.97 |
| (E2) | $\mathcal{L}_{CE_{all}} + \mathcal{L}_1 + \mathcal{L}_4$ | PGD | 91.88 | 97.47 | 98.67 |

Table 1. AVASD mAP(%) of different models under MIM and PGD attack algorithms. We get the data with correct prediction for model (A)-(E2) and do intersection to get the testing data.

### Attacker Perspective

- **Figure 3 (a):** TalkNet is vulnerable to white-box attacks but robust to black-box attacks.
- **Figure 3 (b):** TalkNet is vulnerable to multi-modal and visual attacks but robust to audio attacks.

### Defense Perspective

- **Table 1:** Combining AVIL with adversarial training can leverage their complementary to reach the best adversarial robustness.

## References


[1] Yinpeng Dong et al. Boosting adversarial attacks with momentum. arxiv preprint. *arXiv preprint arXiv: 1710.06081*, 2017.
[2] Ian J Goodfellow et al. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
[3] Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
[4] Joseph Roth et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP*.
[5] Ruijie Tao et al. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021.