Visual Sound Localization in the Wild by Cross-Modal Interference Erasing

Xian Liu 1,2,*, Rui Qian 1,3,*, Hang Zhou 1*, Di Hu 4, Weiyao Lin 3, Ziwei Liu 5, Bolei Zhou 1, Xiaowei Zhou 2
1 The Chinese University of Hong Kong 2 Zhejiang University 3 Shanghai Jiao Tong University 4 Renmin University of China 5 Nanyang Technological University

Motivation

- In the localization task, sound sources of different volumes should be evenly identified.
- The off-screen sound and background noise will compromise the procedure of audio-visual modality matching.

Key Notations

Audio-visual Pairs: \( (a_i, v_i) | i = 1, 2, ..., N \)

Audio-visual Prototype: \( P \in \mathbb{R}^{Y \times K} \quad y_i \in \{0,1\}^Y \)

Cross-Distillation Loss: \( D_{KL}(p^{af}||p^{aaf}) + D_{KL}(p^{pa}||p^{aaf}) \)

Experiments

Localization results on realistic and synthetic videos. The green box indicates target sounding object area, while the red box means this class of object is silent and its activation value should be low.

Conclusions

- We introduce a novel framework Interference Eraser (IEr) to enhance robust visual sound localization for in-the-wild scenes.
- Our proposed Audio-Instance-Identifier learns the distinguishing-step to achieve volume agnostic mixed sound perception.
- We propose the Cross-modal Referrer to eliminate the interference of visible but silent objects and audible but off-screen sounds.