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

AAAI-22

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.

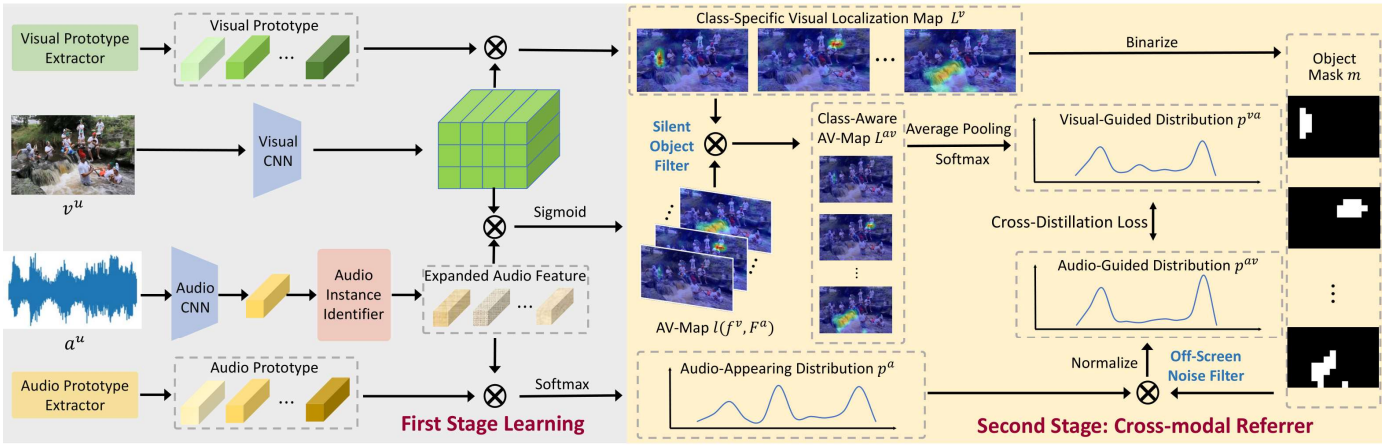
Key Notations

Audio-visual Pairs: $\{(a_i, v_i)|i = 1,2,\dots,N\}$

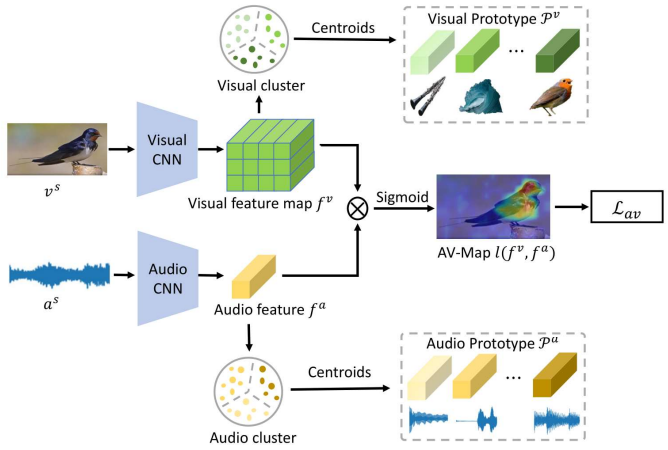
Audio-visual Prototype: $\mathcal{P} \in R^{K\times C}$ $y_i \in \{0,1\}^K$

Cross-Distillation Loss: $D_{KL}(p^{va}||p^{av}) + D_{KL}(p^{av}||p^{va})$

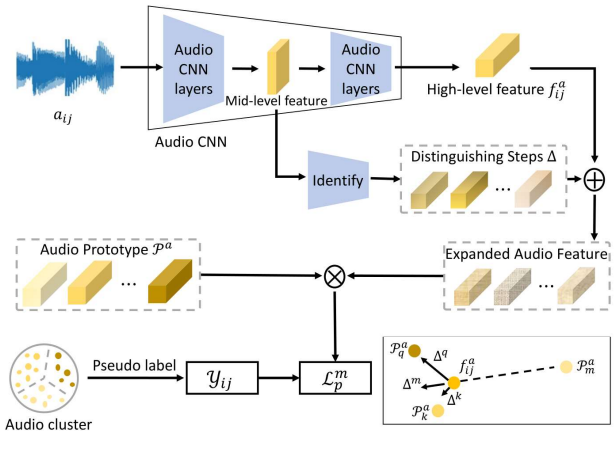
Interference Eraser (IEr) Framework



Left: Audio-visual Prototype Extraction



Right: Audio-Instance-Identifier



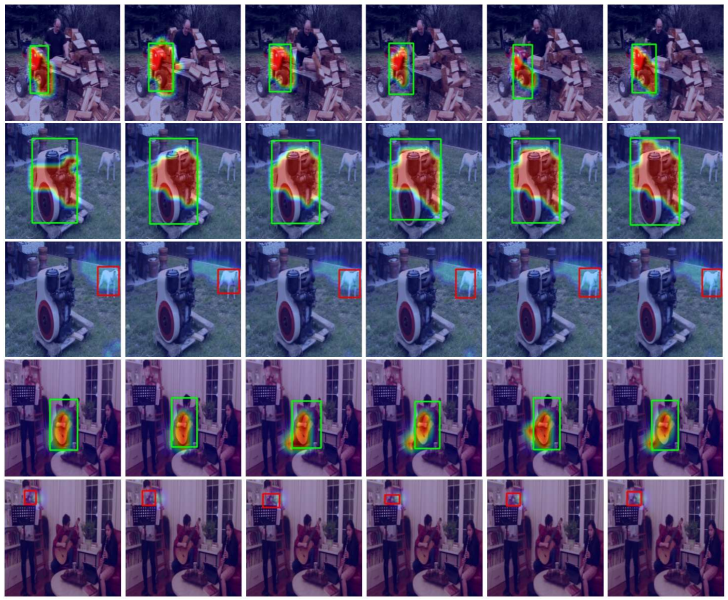
Experiments

Scenario	Single Sound Scene (a)				General in-the-Wild Scene (b)							
Dataset	MUSIC		VGGSound		MUSIC-Syn.		MUSIC-Duet		MUSIC-Un.		VGG-Un.	
Methods\Metrics	IoU	AUC	IoU	AUC	CIoU	AUC	CIoU	AUC	CIoU	AUC	CIoU	AUC
Object-that-sound	26.1	35.8	48.4	46.1	3.7	10.2	13.2	18.3	0.1	6.8	7.8	15.1
Sound-of-pixel	40.5	43.3	42.5	45.1	8.1	11.8	16.8	16.8	7.5	11.6	7.9	14.4
DSOL	51.4	43.6	49.3	45.8	32.3	23.5	30.2	22.1	3.2	7.3	8.1	12.2
Interference Eraser	**53.9**	**50.7**	**51.3**	**46.9**	**47.6**	**29.8**	**52.9**	**33.8**	**15.6**	**15.3**	**12.8**	**17.6**

Visual Sound Localization results in (a) Single Sound Scene; (b) General in-the-Wild Scene.

Dataset	MUSIC				VGGSound			
Method	NMI	Prec.	Rec.	mAP	NMI	Prec.	Rec.	mAP
w/o **Audio-Instance-Identifier**	0.692	0.372	0.335	0.355	0.410	0.218	0.076	0.189
Audio-Instance-Identifier w/o curriculum	0.758	0.441	0.489	0.403	0.414	0.298	0.160	0.231
Audio-Instance-Identifier w/ curriculum	**0.809**	**0.461**	**0.715**	**0.433**	**0.436**	**0.346**	**0.232**	**0.283**

Ablation study on Audio-Instance-Identifier with curriculum learning strategy.



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.