Localizing Visual Sounds the Easy Way

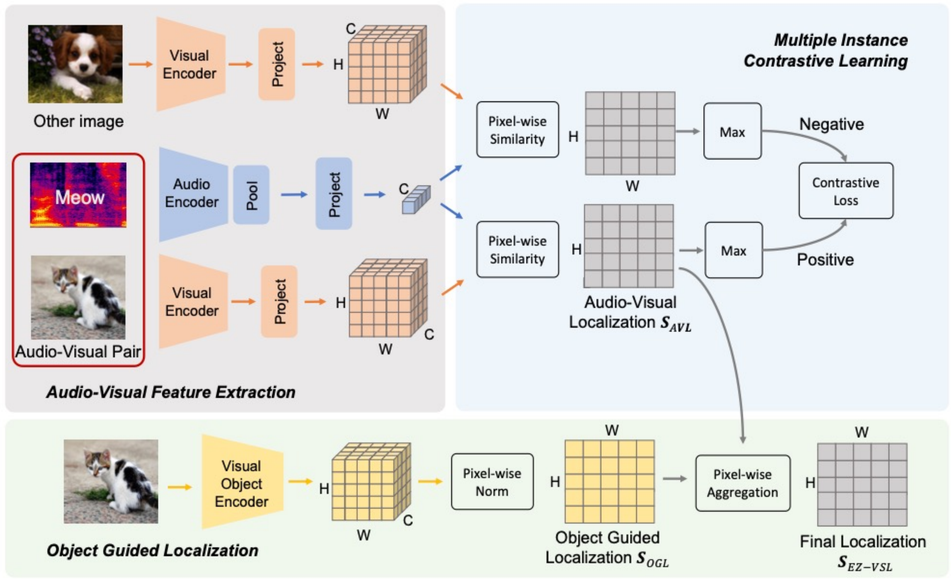
Shentong Mo[1], Pedro Morgado[1,2]
[1]Carnegie Mellon University & [2]University of Wisconsin-Madison

Paper, code, and data are available:
https://github.com/stoneMo/EZ-VSL

Contributions

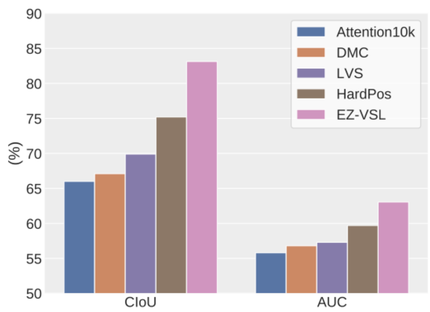
✦ We present a simple yet effective multiple instance learning framework for unsupervised sound source visual localization, which we call **EZ-VSL**.

✦ We propose a novel **object-guided localization** scheme that favors object regions, which are more likely to contain sound sources.

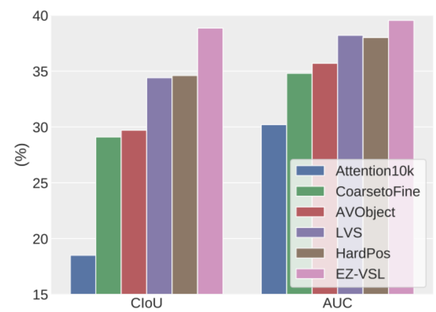


- **Training**: the audio-visual feature extractor computes global audio and localization visual features. Audio-visual alignment is learned by a *multiple instance contrastive learning* objective.
- **Inference**: At inference time, we use another visual encoder pre-trained on object recognition to compute object localization maps, which are combined with audio-visual localization maps for the final prediction.

Comparison with state-of-the-arts



(a) Flick SoundNet

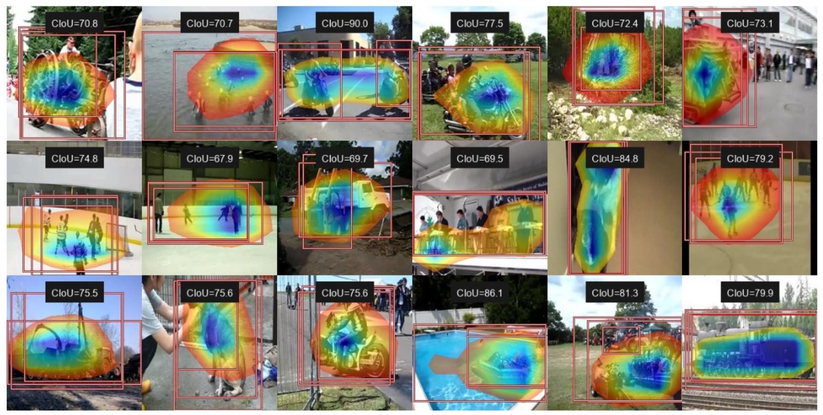


(b) VGG Sound Sources

Ablation Study

AVL	L1-OGL	CLS-OGL	Flickr SoundNet CIoU(%)	Flickr SoundNet AUC(%)	VGG-SS CIoU(%)	VGG-SS AUC(%)
✓			78.31	61.74	35.96	38.20
	✓		78.31	61.17	36.77	38.69
		✓	75.10	58.18	35.13	38.08
✓		✓	81.93	62.50	38.58	39.59
✓	✓		**83.94**	**63.60**	**39.34**	**39.78**

Qualitative Visualizations



Cross-dataset Generalization

Test set	Training set	Method	CIoU(%)	AUC(%)
Flickr SoundNet	VGG-Sound 10k	LVS [6]	61.80	53.60
		EZ-VSL	**78.71**	**61.56**
	VGG-Sound 144k	LVS [6]	71.90	58.20
		EZ-VSL	**84.34**	**63.77**
	VGG-Sound Full	LVS [6]	73.59	59.00
		EZ-VSL	**83.94**	**63.60**
VGG-SS	Flickr 10k	LVS [6]	18.71	30.29
		EZ-VSL	**35.54**	**38.18**
	Flickr 144k	LVS [6]	26.95	34.30
		EZ-VSL	**38.62**	**39.20**

Open Set Source Localization

Test class	Method	CIoU(%)	AUC(%)
Heard 110	LVS [6]	28.90	36.20
	EZ-VSL	**37.25**	**38.97**
Unheard 110	LVS [6]	26.30	34.70
	EZ-VSL	**39.57**	**39.60**

A-V Matching Strategies

AV matching strategy	Flickr SoundNet CIoU(%)	Flickr SoundNet AUC(%)	VGG-SS CIoU(%)	VGG-SS AUC(%)
$\texttt{sim}(\text{MaxPool}_{xy}(V_{xy}), A)$	49.40	48.97	12.72	27.10
$\text{AvgPool}_{xy}(\texttt{sim}(V_{xy}, A))$	33.33	37.56	6.03	19.44
$\text{MaxPool}_{xy}(\texttt{sim}(V_{xy}, A))$	**78.31**	**61.74**	**35.96**	**38.20**

Project Website

Feel free to scan for more details!