



# Localizing Visual Sounds the Easy Way

Shentong Mo<sup>1</sup>, Pedro Morgado<sup>1,2</sup>

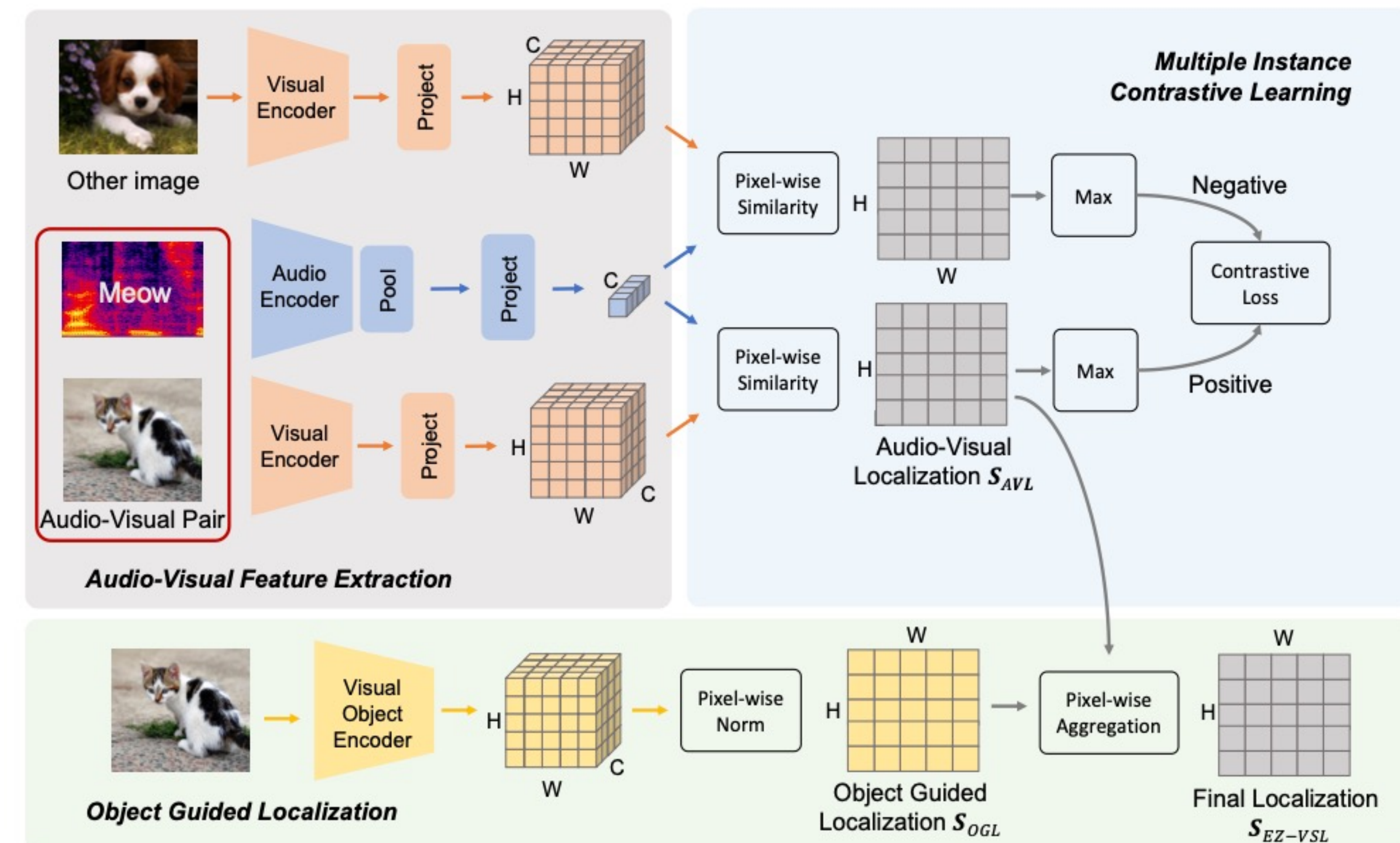
<sup>1</sup>Carnegie Mellon University & <sup>2</sup>University of Wisconsin-Madison

Paper, code, and data are available:

<https://github.com/stoneMo/EZ-VSL>

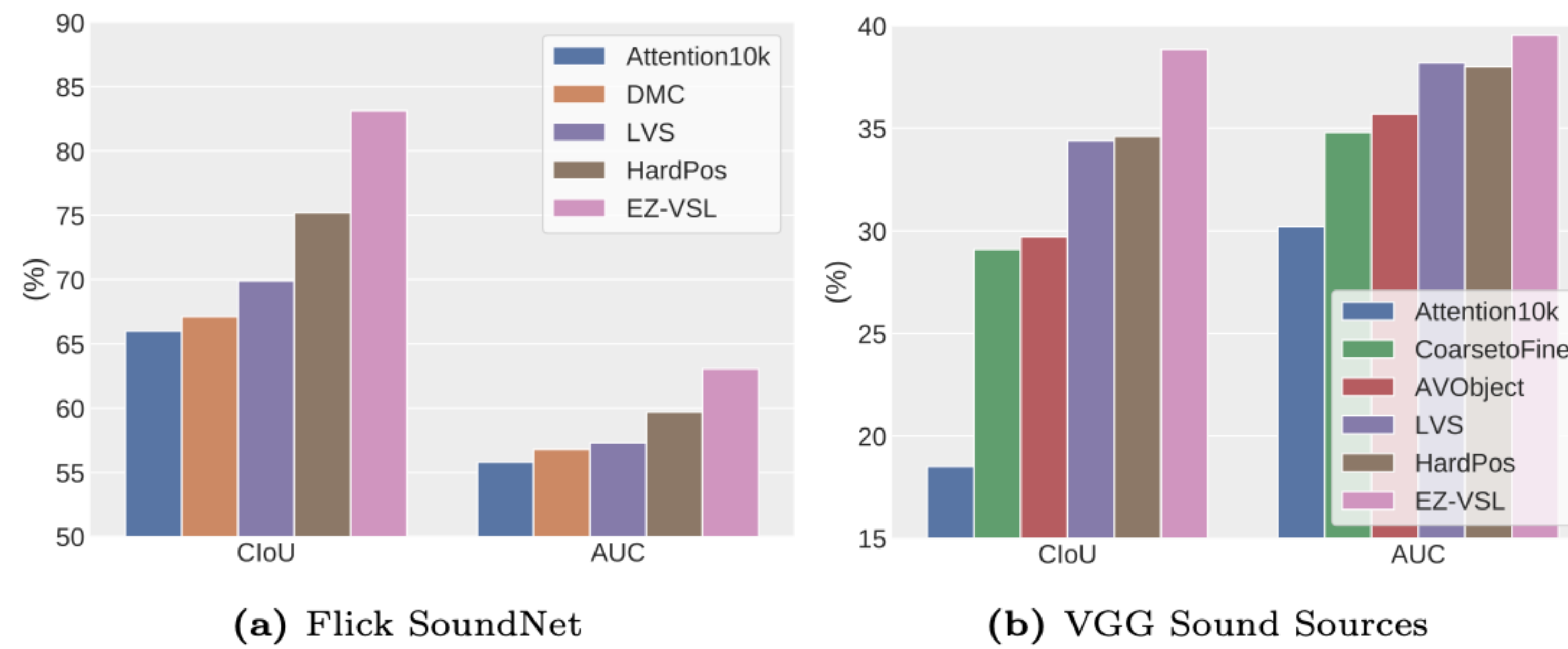
## Contributions

- ◆ 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.

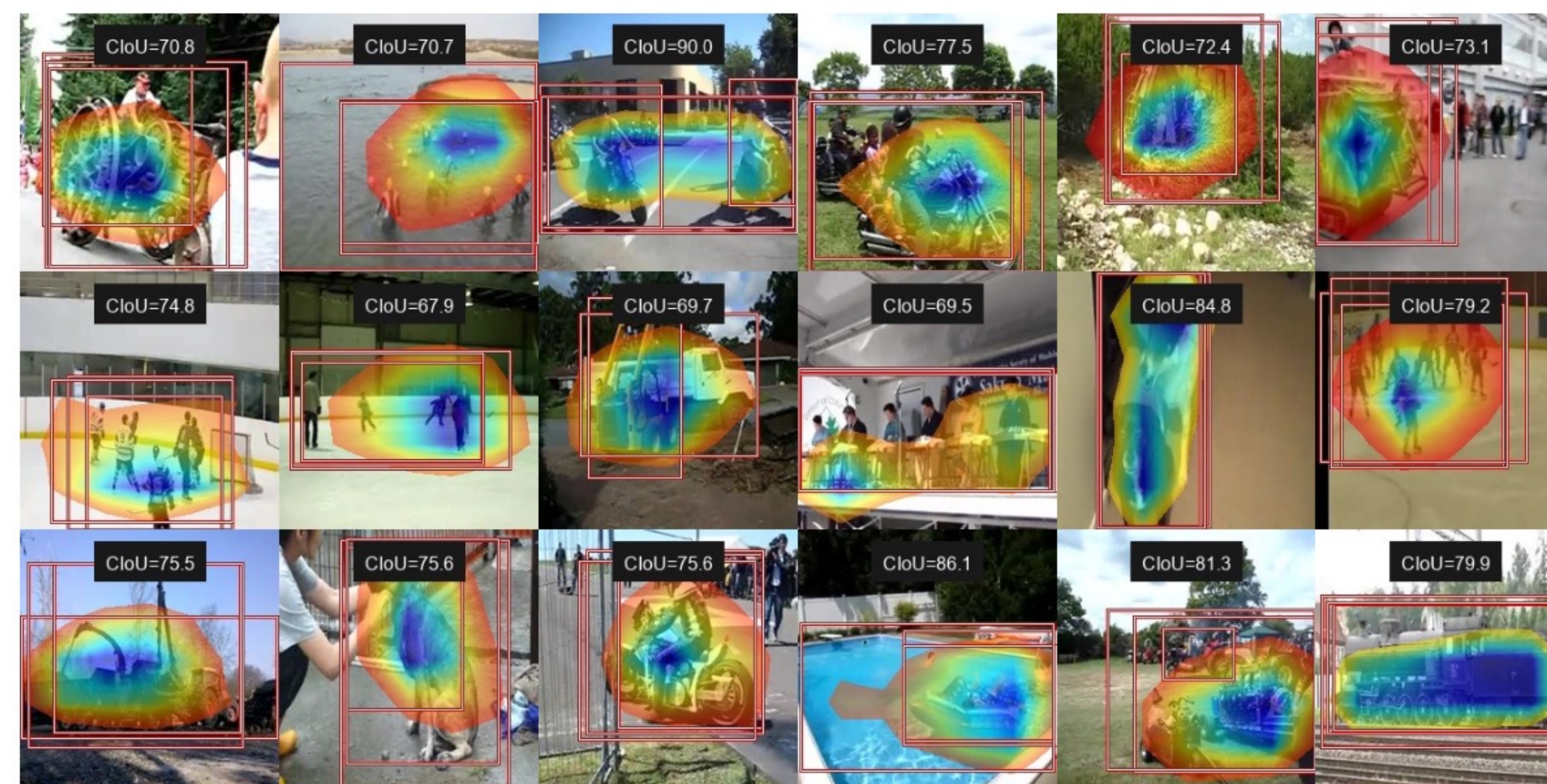
## Comparison with state-of-the-arts



## Ablation Study

AVL	L1-OGL	CLS-OGL	Flickr SoundNet		VGG-SS	
			CIoU(%)	AUC(%)	CIoU(%)	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
✓	✓		<b>83.94</b>	<b>63.60</b>	<b>39.34</b>	<b>39.78</b>

## 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	<b>78.71</b>	<b>61.56</b>
	VGG-Sound 144k	LVS [6]	71.90	58.20
		EZ-VSL	<b>84.34</b>	<b>63.77</b>
	VGG-Sound Full	LVS [6]	73.59	59.00
		EZ-VSL	<b>83.94</b>	<b>63.60</b>
VGG-SS	Flickr 10k	LVS [6]	18.71	30.29
		EZ-VSL	<b>35.54</b>	<b>38.18</b>
	Flickr 144k	LVS [6]	26.95	34.30
		EZ-VSL	<b>38.62</b>	<b>39.20</b>

## Open Set Source Localization

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

## A-V Matching Strategies

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

## Project Website



Feel free to scan for more details!



October 23-27, 2022, Tel Aviv

TEL AVIV 2022