OCTOBER 1 - 5, 2023

IEEE/RSJ International Conference on Intelligent Robots and Systems

The Audio-Visual BatVision Dataset for Research on Sight and Sound

Amandine Brunetto[1], Sascha Hornauer[1]*, Stella X. Yu[2], Fabien Moutarde[1]*

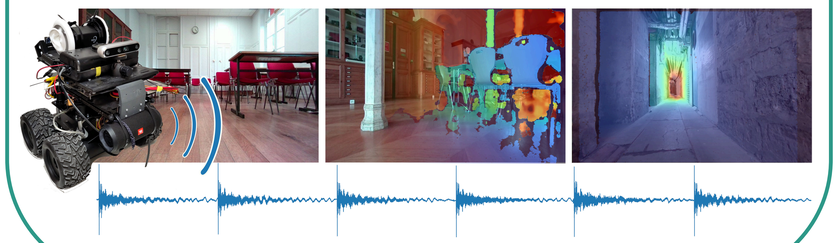
[1] Center for Robotics, Mines Paris – PSL University, Paris, France
[2] University of Michigan, Ann Arbor, United States of America
* Equal Contribution

Introduction

- **Sound** is a promising modality able to **complement visual sensors** especially when failing.
- Some species use sound to navigate: **Echolocation**
- No large-scale real-world audio-visual dataset available **for robotic echolocation** and **scene understanding**.

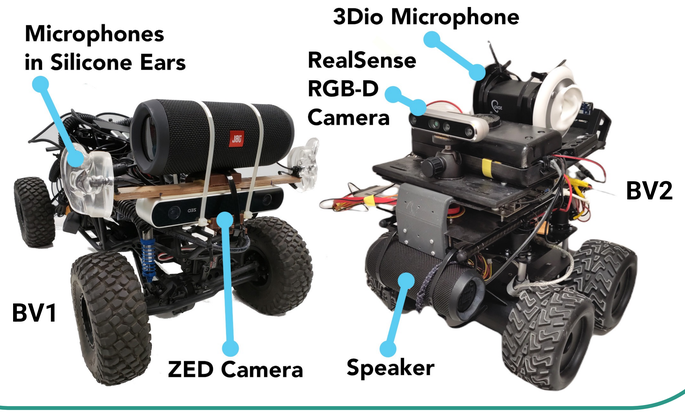
The BatVision Dataset

55K Real-world synchronized echoes and RGB-D images from a robot perspective recorded at UC Berkeley (BV1) and Mines Paris PSL (BV2).



Method

- **Linear frequency sweep signals** (*chirps*) between 20Hz-20kHz ascending in 0.3ms every 0.5s.
- **Echoes** are recorded with **a binaural microphone**.
- Sound & Vision synchronization is done using **ROS**.



Dataset

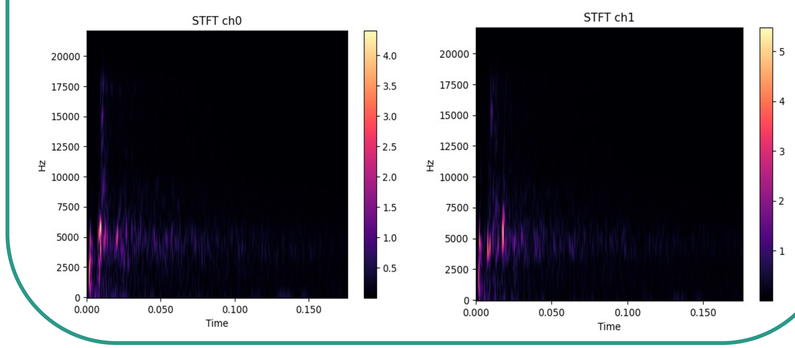
- **Diverse environments**: varied architecture styles, room shapes, materials and objects.
- Mainly **indoor** data but BV2 contains **outdoors**.

Depth Prediction

- **Baseline: U-Net** with audio as the only input.
- ***Sota* method developed in simulation**: **BITD**[3]. Successfully trained and tested on BatVision.

Input

- Echoes clipped according to depth (BV2 at 30m, BV1 at 12m) and processed using STFT.
- For BITD, echoes, RGB and a material network are used.

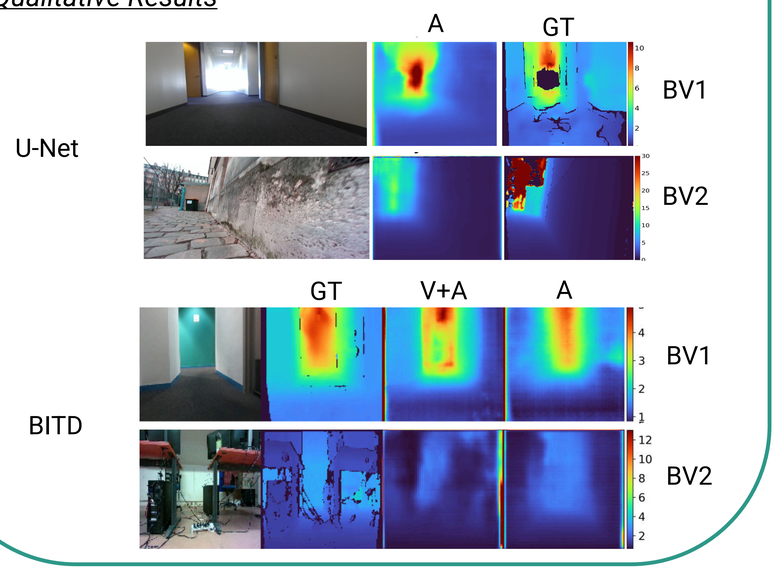


Results

Quantitative Results

		RMSE↓	REL↓	log10↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
BITD	Replica [19]	**0.249**	**0.118**	**0.046**	**0.869**	**0.943**	**0.970**
	Matterport3D [19]	0.950	0.175	0.079	0.733	0.886	0.948
	BV1	**0.901**	**0.234**	**0.097**	**0.688**	**0.888**	**0.942**
	BV2	2.286	0.323	0.119	0.647	0.834	0.901
	BV1 AO	1.350	0.453	0.159	0.441	0.707	0.843
	BV2 AO	2.878	0.521	0.197	0.430	0.629	0.765
U-Net	U-Net BV1 AO	1.336	0.361	0.147	0.508	0.738	0.856
	U-Net BV2 AO	2.676	0.432	0.160	0.497	0.717	0.835

Qualitative Results



Conclusion

- We provide **publicly available large-scale real audio-visual data** to improve task performance and uncover novel uses.
- **Depth prediction** from sound alone or in addition to vision is possible on those real data.
- Data and code available:

[3] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. “Beyond image to depth: Improving depth prediction using echoes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021,