HAL
open science

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

Amandine Brunetto, Hornauer Sascha, Stella Yu, Fabien Moutarde

To cite this version:

Amandine Brunetto, Hornauer Sascha, Stella Yu, Fabien Moutarde. The Audio-Visual BatVision Dataset for Research on Sight and Sound. The 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023), Oct 2023, Detroit, United States. hal-04375380

HAL Id: hal-04375380

https://minesparis-psl.hal.science/hal-04375380v1

Submitted on 5 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

arXiv:2303.07257v2 [cs.RO] 14 Mar 2023

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

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

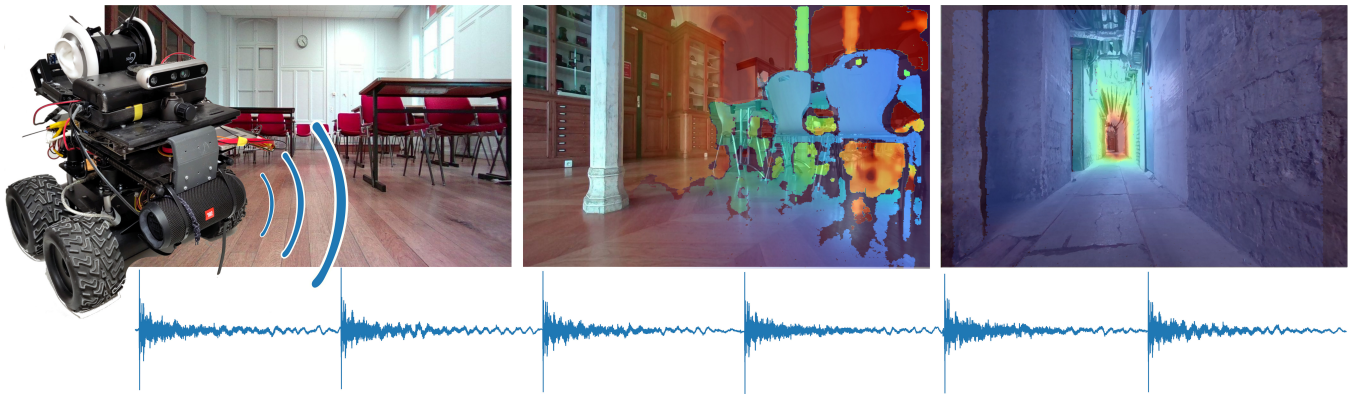

Fig. 1: The BatVision dataset contains large scale audio-visual data from a robots' perspective. For its creation, a robot traversed corridors, offices, lecture halls and driveways at a historic campus and modern office building like a bat, emitting *chirping* sounds with a speaker. A binaural microphone recorded their echoes which carry rich scene information of objects, materials and layout. With this paper we provide echoes, camera images and depth maps, shown overlayed on typical scenes on the right. Echoes are shown under images. This dataset will help investigate fundamental questions on how sound interacts with spaces, how it can be harnessed for robotic navigation and what in general can be understood about a scene from how it sounds.

***Abstract*— Vision research showed remarkable success in understanding our world, propelled by datasets of images and videos. Sensor data from radar, LiDAR and cameras supports research in robotics and autonomous driving for at least a decade. However, while visual sensors may fail in some conditions, sound has recently shown potential to complement sensor data. Simulated room impulse responses (RIR) in 3D apartment-models became a benchmark dataset for the community, fostering a range of audiovisual research. In simulation, depth is predictable from sound, by learning bat-like perception with a neural network. Concurrently, the same was achieved in reality by using RGB-D images and echoes of *chirping* sounds. Biomimicking bat perception is an exciting new direction but needs dedicated datasets to explore the potential. Therefore, we collected the BatVision dataset to provide large-scale echoes in complex real-world scenes to the community. We equipped a robot with a speaker to emit *chirps* and a binaural microphone to record their echoes. Synchronized RGB-D images from the same perspective provide visual labels of traversed spaces. We sampled modern US office spaces to historic French university grounds, indoor and outdoor with large architectural variety. This dataset will allow research on robot echolocation, general audio-visual tasks and sound phænomena unavailable in simulated data. We show promising results for audio-only depth prediction and show how state-of-the-art work developed for simulated data can also succeed on our dataset. The data can be downloaded at `https://forms.gle/W6xtshMgoXGZDwsE7`**

I. Introduction

Large-scale datasets propelled research in past decades, providing first static images and later an abundance of videos for tasks from object detection to activity recognition.

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

Sounds, correlated with visual data from their source, provide exploitable information about an action or context, often at marginal computational overhead.

A novel research direction aims to listen to the environment for improved task performance. Simulated room impulse response (RIR) datasets allow researchers to investigate the interaction of sounds with known space layouts for e.g. depth prediction, obstacle avoidance or to drive towards an alarm beyond the line of sight. They have been used successfully to predict 3D layouts from simulated chirps, similar to how bats find their prey. Robots mastering this echolocation could create instant maps beyond their immediate surroundings without LiDAR, overcoming their field of view and operate even in darkness and smoke.

While some datasets provide recorded RIRs within limitations, there exists no large-scale real dataset to investigate audio-visual 3D scene understanding for robots. This hinders domain adaptation from simulation as well as learning to be robust against or even exploit real world background sounds.

Bats are capable of navigating and localizing prey in flight using echolocation. We showed successfully how to adopt this feat for machine listening, using only audible echoes of chirps for depth prediction [11, 12]. To understand the extend to which this method can complement failing vision sensors and even extend beyond the field of view, more real data in typical robotic scenarios is needed. We therefore present the BatVision Dataset providing publicly available large-scale real data of complex scenes for research on audio-visual 3D scene understanding and robotic echolocation.

Sound-augmented Task Performance. Sound arrives omnidirectional and provides rich information about the space

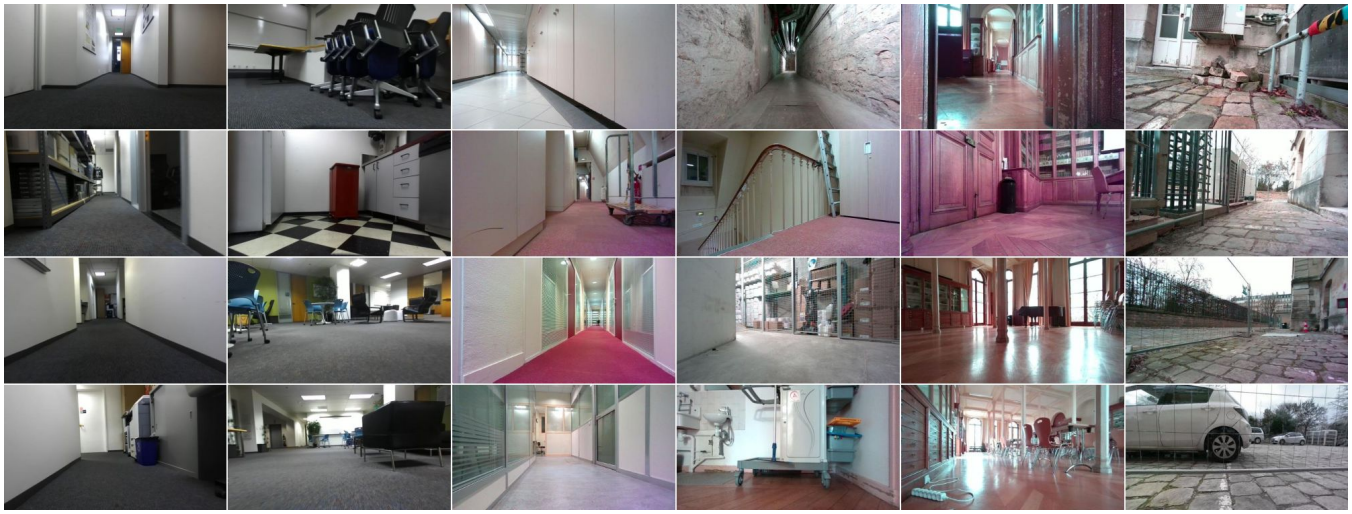

Fig. 2: Example scenes from BV1 (left two columns) and BV2 (right four columns). BV1 contains typical office scenes with many corridors and some open spaces. BV2 columns show a wide variety of corridors, with and without carpet, maintenance areas, antique conference rooms and outdoor scenes.

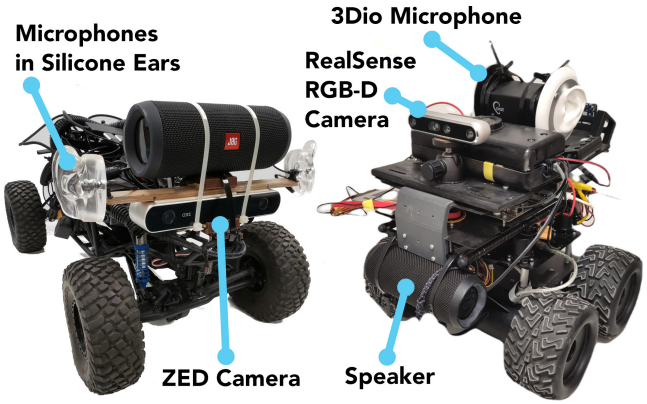

Fig. 3: Recording robot used for BV1 at UC Berkeley (left) and BV2 at Mines Paris (right). Both record binaural audio and RGB-D images yet the hardware setup differs.

it crosses. Due to the correlation of sound and vision it can provide complementary information when visual sensors fail. Cameras and LiDAR struggle with occlusions, reflections and smoke. Depth-from-stereo algorithms produce artifacts in low-light conditions and when objects lack textures. Combining audio-visual information can improve task performance with one modality providing labels for the other [19, 21]. Biomimicking the echolocation principle can even improve 3D scene understanding using sound [11, 13, 28].

Audio-visual data suitable for these tasks is sparse because real-world recordings require complex setups and long sessions [8, 2, 30]. In consequence, researchers often create small datasets tailored to their tasks [19, 31]). For research on robotic echolocation, only very few simple datasets exists.

The BatVision Dataset. Motivated by the success of simulated datasets and the lack of real data for robotic echolocation and audio-visual scene understanding, we recorded the real-life BatVision dataset. We provide large-scale recordings from traversing real campus spaces with a robot, covering a wide range of materials, room shapes and objects, shown in Figure 2. We provide synchronized RGB-D camera images and recorded echoes of *chirping* sounds emitted with a forward facing speaker. Similar to RIRs, echoes can be used to infer the geometry of the room to allow the robot bat-like perception even in darkness, smoke and fog.

Data was collected at UC Berkeley (52,220 instances) and Mines Paris PSL (3,120 instances). Emitted *chirp* signals range between 20Hz and 20kHz. RGB-D images were recorded with an Intel Realsense camera mounted on a robot (Mines Paris) and a ZED camera (UC Berkeley).

In the following we place the BatVision dataset among similar ones and argue its unique utility. We then detail the collection process, data and showcase performance of depth prediction methods trained on it. UC Berkeley data will be referred to as **BV1** and Mines Paris data as **BV2**.

II. Related Work

Audio-visual data is available in many datasets, collected in reality or simulation, suitable for different tasks. However, to our knowledge, no existing dataset supports audio-visual 3D scene understanding with large-scale real data.

Acoustic Room Impulse Response Datasets. RIR datasets are used in room acoustics, speech and audio processing, sound localization and separation and virtual reality. RIRs account for the effect of room acoustics on audio signals and describe sound propagation. By convolving with RIRs, acoustic properties, such as reverb, can be transferred onto arbitrary *dry* sounds. Several datasets exist with RIRs, sampled in real spaces using a complex measurement routine.

The Acoustic Multichannel RIR Dataset [2] contains samples from one rectangular room. Panels, moved between recording sessions, emulate environments such as a small office, meeting or lecture room. The dEchorate dataset [16] provides 1.8k sampled RIRs with annotated early echo timings and 3D positions of microphones and sound sources. For the MIRaGe dataset [8] 371k RIRs are measured in a

dense grid of 4104 source positions in different rooms. [7] estimates RIRs from spherical camera images. RIRs were sampled as ground truth in living room style environments using a custom array of 48 microphones and a soundfield microphone. Object shapes are simple rectangles, aligned towards the main axis of the room.

While these datasets were sampled with high quality equipment and allow room geometry and RIRs estimation their environments are rectangular rooms with cuboid objects and simple materials. In contrast, for BatVision we sampled data in real public spaces, selected for their variety of materials, shape and architectural properties.

Audio-Visual Simulation.

Datasets containing simulated data are widely used, cost-effective, allow large scale generation, controlled conditions and easy data annotation [27], [3]). Simulation of sound propagation has been extensively used in games and AR applications.

SoundSpaces 1.0 [10] allows simulating sound propagation by providing RIR renderings built with bidirectional path-tracing in 3D-scanned apartment models. They were generated for discreet positions and orientations in a grid in two 3D environments, Matterport3D [6]) and Replica[9]). SoundSpaces 2.0 [22] provides continuous on-the-fly rendering and improved the sound propagation. It has since become a widely used benchmark showing impressive performance on a wide range of tasks [29, 13, 29, 28, 18].

While it is a huge contribution to the community, bridging the gap between simulation and the real world stays a significant challenge. Models trained in simulation often overfit to characteristics of the simulator in surprising ways [14]. However, for audio-based 3D scene understanding there is no real-life dataset of a size, comparable to SoundSpaces which motivated our data collection.

Sound Event Localization and Detection Datasets. Sound event localization and detection (SELD) aims to infer the azimuth, elevation and distance of sounds relative to an observer, with or without additional classification. Datasets in the domain contain one or several, moving or static, clear or noisy sounds in different environments. The TUT Acoustic Scenes 2016 dataset [4] consists of 15 real acoustic scenes, such as *City center* or *Metro station*, with annotated sound event classes, onset and offset times, and spatial information. The STARSS22 dataset [26] is a collection of 22 real-world acoustic scene recordings with sound event annotations, captured with a high-resolution spherical microphone array. Annotations cover 13 classes and direction of arrivals.

For 3D scene understanding and especially reconstruction, SELD datasets often lack sufficient spatial ground truth information of the environments the sounds occur in. For the BatVision dataset forward facing camera images and depth information is provided to allow exploring the correlations in the audio and visual modality.

Real-World Audio-Visual Datasets.

The huge EGO 4D dataset [23] contains videos from the perspective of persons performing a wide range of activities. Parts contain audio and 3D meshes of the environment. It is an impressive effort and contribution for many research areas such as activity recognition. In contrast, for BatVision we emit chirps into recorded scenes to allow inferring RIRs and finally the 3D scene. It is magnitudes smaller but better suited for our tasks.

[5] proposed "The Greatest Hits" dataset. By filming objects being struck with a drumstick, they aim to study physical interactions with a visual scene and synthesize plausible impact sounds. It includes 46,577 actions of hitting and scratching objects. The research is a step towards understanding the link between material, action and emitted sound. Similarly, with our recorded echoes we collect interactions between chirps and scene materials.

The authors in [19] predict approaching vehicles at blind intersections from sound before they enter the line-of-sight. They captured crossing vehicles at intersections with a custom microphone array and a front-facing camera mounted on a car. [15] recorded the "Omni Auditory Perception Dataset" standing next to streets with eight binaural microphones and a 360° camera. Pseudo ground truth depth is predicted from monocular images. The authors emphasize its mid-size and the great effort required to create it.

[17] introduced the "Quiet Campus Dataset" of ambient sounds from a variety of quiet indoor scenes. Paired with RGB-D images, they predict distances to walls from the whirling of a fan or noise coming through a window. While this unique idea shows impressive task performance with passive observations, we focus on active sounds which are humanly audible for improved performance.

Recording interactions of chirps with spaces has a recent history. The BatVision depth-from-binaural audio idea [11] inspired [20] to collect similar data with added 360° LiDAR scans for depth and four ear-shaped microphones. With 5,000 samples the dataset falls between BV1 and BV2. [31] predict an occupancy map from conversations in spaces. Beyond using SoundSpaces they also captured real data in a mock-up apartment, citing the lack of publicly available real world data. To compute RIRs they capture *chirps* from a speaker with an Eigenmike. The "Studio Dataset" [24] records 1478 samples of *chirp* echoes with four microphones and RGB-D images with a RealSense camera. Everything is fixed into one metal frame facing in one direction. While also testing in SoundSpaces, their real-world recordings are done in one rectangular room showing simple cuboid objects.

The recording of so many datasets, tailored to specific needs show the potential impact that more and larger publicly available datasets can have. Presented tasks from depth prediction to occupancy mapping can be investigated with our BatVision dataset. We will describe in the following how and give details on the extend of our data collection.

III. The Audio-Visual BatVision Dataset

A. Dataset Overview

We collected the Audio-Visual BatVision Dataset at various locations at the UC Berkeley (ICSI), and Ecole des Mines Paris using small robots, carrying all sensors. The sites

TABLE I: Overview of Real-World Dataset. Mic. stands for microphone, A. for active and P. for passive sound. Here, by *chirp* we mean frequency swept signals with different range and parameters. Passive sound means no controlled sound is emitted during the recording. Instances are given either by their number or by the total amount of hours or videos recorded.

Dataset	# Instances	Locations	Visual Labels	Perspective	Audio Labels	Audio Sensors	Sound Type
BV1	52,220	Hallways, open areas, conference rooms, office spaces	Monocular RGB Depth from stereo clipped at 12m	Robot	72.5ms long binaural separated audio (44.1kHz)	2 Hear-shaped Mic.	A. (*chirp*)
BV2	3,120	Hallways, outdoors, narrow underground corridors, conference rooms	Monocular RGB Depth from infrared at maximum 64m	Robot	0.45s long binaural audio (44.1kHz)	Hear-shaped binaural Mic.	A. (*chirp*)
Acoustic Room Impulse Response Datasets							
[2]	234	Rectangular room. Panels emulate typical acoustic environments	✗	✗	8 channels RIR (48kHz)	Mic. array	A. (*chirp*)
[8]	371k	Rectangular room. Panels emulate typical acoustic environments	✗	✗	RIR (48kHz)	6 Mic. arrays, One additional mic.	A. (*chirp*, white noise)
[16]	≈ 1.8k	Cuboïd room with different wall configurations	✗	✗	RIR (48kHz)	Mic. Array	A. (*chirp*, white noise, anechoïc speech)
[7]	?	Living room, controlled acoustic environment	Stereo 360° RGB, Depth from stereo	?	RIR	Mic. circle	A. (*chirp*)
Sound Event Localization and Detection Datasets							
[4]	≈77k	15 outdoor and indoor acoustic scenes	✗	✗	30s long binaural audio (44.1kHz) annotated in sound classes	Binaural mic.	P.
[26]	≈5h	Indoor rooms	360° RGB	✗	2x 4-channels spatial format audio (24kHz) annotated in 13 sound classes	Spherical Mic. Array (SMA)	P.
Real-World Audio-Visuals Datasets							
[23]	2,535h of videos with audio	Wide range of activities, cities and locations	RGB, 3D scans and other	Person	?	Various	P.
[5]	46,577	Indoor and outdoor scenes	RGB	Person	35s long audio	Mic. attached to camera	A. (object stuck with drumstick)
[19]	411 video recordings	Urban environments	RGB	Car	1s long audio (48kHz)	Mic. array	P.
[31]	?	Mock-up apartment	RGB, Depth from mono, Occupancy map	Rig	RIR and conversation (16kHz)	SMA	A. (*chirp*, speaker conversation)
[15]	54,250 video segments	Streets of a city covering 165 locations	360° RGB	Rig	3s long 8-channels audio (96kHz)	4 hear-shaped binaural mic.	P.
[24]	1,478	Near-rectangular reverberant studio	RGB, Depth	Rig	1s long audio (16kHz)	4 omnidirectional mic.	A. (*chirp*)
[17]	≈15h	Classroom and hallways	RGB, Depth	Robot	2 channels audio (16kHz)	Stereo mic.	P.

provide varied architectural styles, room shapes, materials and therefore acoustic impressions.

We crossed lecture halls, corridors, offices and cobblestone paths recording while emitting humanly audible sine swept signals (*chirps*) with a forward facing speaker (Figure 3). We recorded their echoes, forward facing camera images and provide depth-maps from the same perspective. *Chirps* ascend from 20Hz to 20kHz in 3 ms. Their use is motivated by their counterparts in nature, helping animals echolocalize, but they are also common in sound engineering to record RIRs [1]. Binaural microphones record audio at 44.1kHz with 24 bits to keep the full frequency range of the chirps. Each *instance* is one *chirp* synchronized with one 1280x720 RGB-D image. An overview of this dataset and other real-

world dataset is available in Table I

The distribution of the collected depth can be seen at fig. 4 and fig. 5. By comparing the average depth value per instances it shows BV2 is more long-tailed. BV1 depth values are clipped to the BV2 max depth value to remove outliers of the stereovision algorithm used. Pixel-wise average depth shows a more complex scene distribution in BV2.

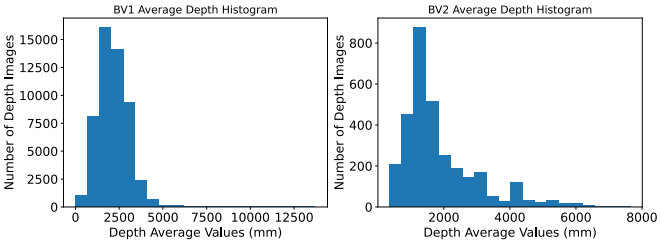

Fig. 4: Histogram of average depth per instance. BV2 depth distribution is more long-tailed than BV1, which is consistent with the variety of data.

Data collected at the UC Berkeley (BV1). In Berkeley, 52,220 instances were collected at two floors of an institute containing hallways, open areas, conference rooms and offices. At two distinct areas of one floor, 39,564 and 7,618 instances were collected for training and validation and 5,038 instances on another floor for testing. While similar, the floors' spatial layout, furniture, occupancy, and decorations are different, see Figure 2. For details, please see [11] showcasing the initial depth prediction-from-audio idea.

A JBL Flip4 Bluetooth speaker emitted chirps while two USB Lavalier MAONO AU-410 microphones, embedded into silicone ears, recorded their echoes. They were mounted 23.5cm apart while the speaker sat between (Figure 3). We excluded motor sounds by pushing the robot on a trolley.

We used a ZED stereo camera to record images of the scene ahead and calculated depth maps with the camera API. Depth-from-stereo fails for some pixel which are *NAN*. We provide depth maps with RGB images from the left camera.

Designed for smaller spaces, audio recordings were cut at 72.5ms, including echoes from objects at 12m distance. This trades-off perception at a relevant distance while excluding later noisy reverberations. We also clip depth to 12m during training even though further distances are available.

Data collected in Mines Paris (BV2). In Ecole des Mines Paris, we collected 3,120 instances with large visual and acoustic variety (See Figure 2).

We split data into sets of 1,911 train, 625 validation and 584 test instances. Given the multi-modal scene distribution we aim to balance task difficulty. We split instances by time of recording when moving through rooms, avoiding loops. That way, our incremental coverage of rooms will lead to poses being sufficiently separated in the sets. Even if we revisit parts of rooms we chose different trajectories thereby avoiding repeated poses. Tasks on this split will be harder than if instances were randomly assigned, which could result in neighbouring poses ending up in training and test. Simple interpolation will not yield best performance. Admittedly our split is easier than separating complete rooms as done for BV1. We observed outdoor reconstruction performance is acceptable but sub-par suggesting domain shift. If very different features, i.e. outdoors, carpeted rooms or stone corridors, are not in the training data further domain adaptation would be necessary which is left for future work.

We provide monocular RGB images and depth from an Intel RealSense D455 camera. The *chirp-emitting* speaker is identical to BV1 but we recorded echoes with a binaural 3Dio Free Space microphone. A robot carried all hardware and an Nvidia Jetson TX-1 to run all recording software (see Figure 3). We excluded motor noise by switching between stopping and driving and filtered out instances while moving.

Synchronization was achieved using ROS timestamps and *chirps* detection with manual checks. Audio data instances were cut to be 0.45s long which includes echoes from objects up to 75m away. Because of larger spaces in BV2, we keep the long tail to sense far away reflection. The maximum depth value of the camera is ≈64m.

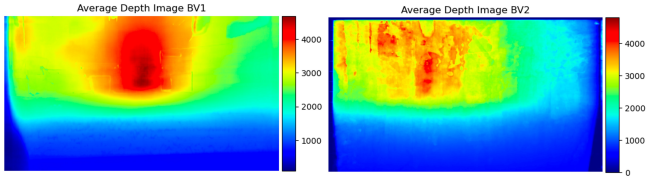

Fig. 5: Average depth per pixel. In BV1 corridors are often centered, in BV2 depth distribution is more complex.

B. Limitations

Unlike in simulation, physically recording introduces a number of limitations. Echoes were recorded in real rooms next to noisy streets and with typical sounds of busy academic institutions. Some data may therefore contain audible noise, typical for the recording context. Models trained will need to learn to be robust to these noise profiles.

Some approaches predicting RIRs need independently changing emitter and microphone positions to record sounds on a direct path [16, 25]. In our data collection by driving, the robot carries all the equipment so the emitter is fixed close to the microphone and only the echoes change.

After filtering instances in motion few remain from the same pose. We kept these to allow learning a noise-model but they can be filtered by thresholding optical flow.

Collection with different hardware for BV1 and 2 leads to different formats. Audio instances collected at UC Berkeley are shorter. Some very bright images from Mines Paris shows slight purple discoloration due to an issues with the Intel RealSense D455 RGB-D camera. Finally, our consumer-grade speaker can not produce the full frequency spectrum.

IV. Depth Prediction on BatVision Data

The data can be used for depth-prediction from audio-visual data with approaches developed for real or simulated data. For illustration we trained a U-Net audio-only depth prediction baseline and compared with one state-of-the-art audio-visual approach developed by its authors in simulation.

Beyond Image to Depth. Recent work improves audio-visual depth prediction using a pre-trained material classifier

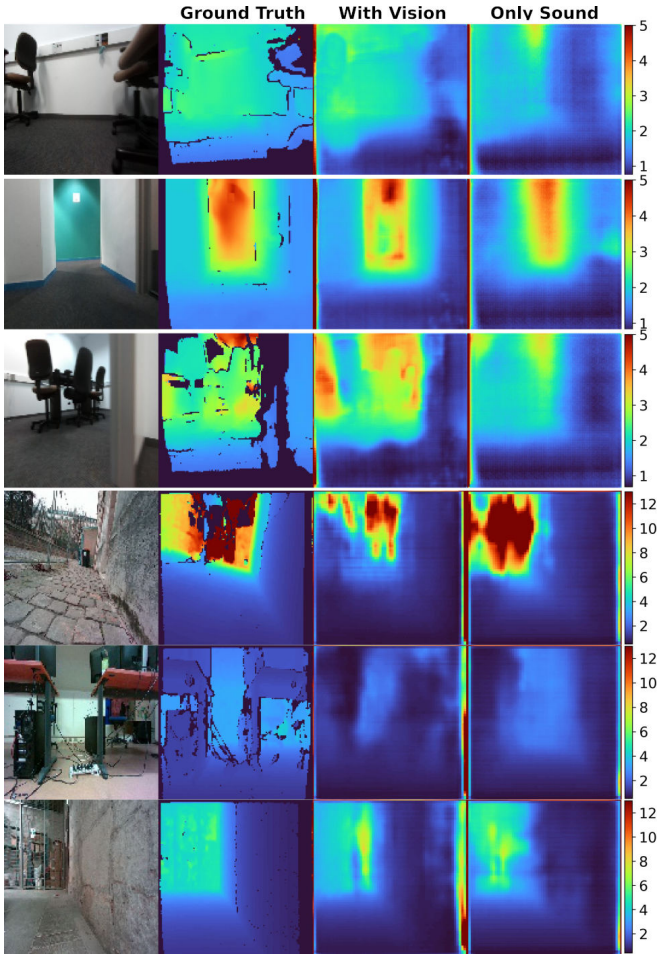

Fig. 6: Test set results when training *Beyond Image to Depth* on BatVision V1 (top) and BatVision V2 (bottom), depth in meters. Same hyperparameters as in simulation show well visible general layout and obstacles. Fifth row shows free space between two desks even based on audio-only. Fine structures such as cables are still hard to reconstruct.

to decide which modality to pay attention to. The authors [18] extract latent material features from SoundSpaces 1.0 camera images and generate attention maps for the visual and audio-stream in the network. They convolve a *dry-chirp* sound with SoundSpace RIRs to simulate emitting it into spaces from different poses. Captured echoes and RGB-D images make up their dataset with which they predict depth from RGB images and audio.

Training their code on the BatVision dataset shows similar performance as on SoundSpaces (see Table II). Apart from changing the spectrogram resolution, no further adjustments were needed when switching to real data. We ablate results in a study using the audio signal only by setting the RGB images to zero. That way we keep the architecture unchanged.

Qualitatively, fine structures are harder to predict which may stem from our coarser ground truth depth compared to simulation (See Figure 6). Measuring depth in reality, either from stereo (ZED camera) or active stereo (RealSense camera), limits the accuracy depending on the object distance.

Comparable results show that an approach developed for

TABLE II: Depth prediction results from Beyond Image to Depth [18], trained on BatVision and simulated data (Replica and Matterport). Similar V1 and Matterport results suggest comparable difficulty of tasks. Our simple U-Net baseline slightly outperforms [18] when using audio-only (AO).

	RMSE↓	REL↓	log10↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
Replica [18]	**0.249**	**0.118**	**0.046**	**0.869**	**0.943**	**0.970**
Matterport3D [18]	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 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

simulated data can be adapted and works with the same hyperparameters on our real dataset.

U-Net Baseline. The extend to which audio can complement visual sensor data is not clear. Related work showed repeatedly the quality of the visual signal strongly dominates task performance. The compared work [18] investigates material-based attention to let the network choose the influence of each modality more explicitly. We investigate the claim that audio can help in bad visual conditions with a baseline, based on the audio signal alone. This serves to investigate the contribution of the audio signal in isolation with clear quality attribution. However, in all practical applications, the visual signal should be used as well.

We train a U-Net from audio only, similar to related work [11, 13] and isolated on BV2 and BV1. We achieve solid results, correctly predicting free space, obstacles and the general room layout (see Figure 7). Without further tweaks such as GCC-Phat features or using a GAN [11, 12], the performance shows the exploitable quality of the data itself.

Implementation Details. The only input are spectrograms, generated from waveforms with 512 frequency bins (nfft), 64 window and 16 hop length, resized to 256x256.

For BV2, best performance is obtained with depth clipped to 30m and audio cut accordingly. This can be explained as audio energy decays with distance so far traveling echoes are hard to distinguish from noise. For BV1, depth is clipped to 12m and the audio accordingly. Ground truth depth is always normalized using the chosen max depth value.

We use an 8 block U-Net with skip connection. Each encoder block is composed of 2D convolutions, batch normalization and leaky ReLUs. Each decoder block is composed of 2D transposed convolution, batch normalization and ReLUs (see Figure 8). We found skip-connections improve performance though contrary to a segmentation task their contribution is not yet clearly understood. We train with 256 batch size, 0.002 learning rate for BV2 and 0.001 learning rate for BV1 with the AdamW optimizer and L1 loss.

Results on Depth Prediction. With this simple baseline we retrieve the general geometry of the space (see Table II and fig. 7). On BV2, complex obstacles such as chairs are visible. When the visual sensors fails (e.g on glass), audio gives correct information about the depth. The network generalizes well between different acoustic environment. It

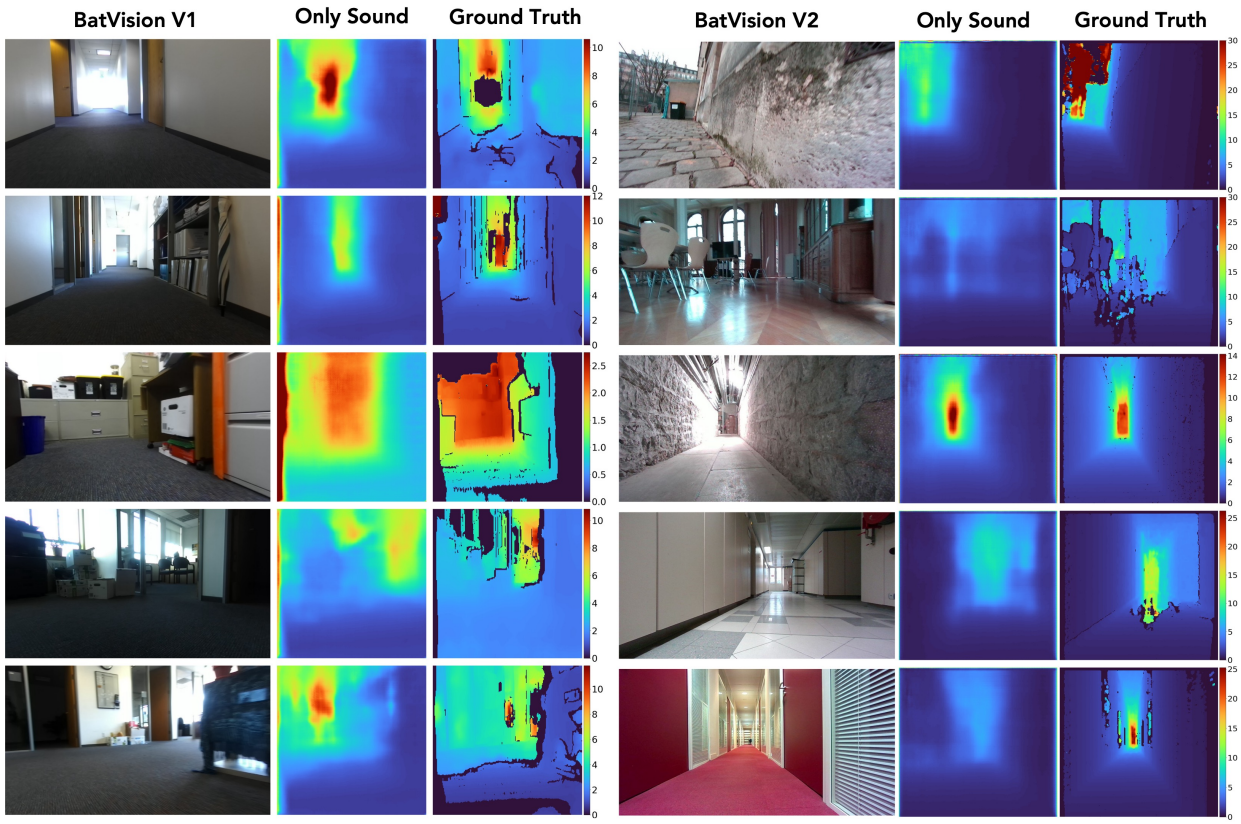

Fig. 7: U-Net Baseline depth predictions on BatVision V1 (left) and BatVision V2 (right) ('*turbo*' colormap). Columns are f.l.t.r. RGB image, depth prediction and ground truth depth. Units in meters. Left, beyond showing correct structures, results improved quantitatively in this larger dataset. Right, obstacles like chairs are reconstructed and corridor layouts are correct, even though wall materials, and hence echoes, vary strongly from carpet to stone.

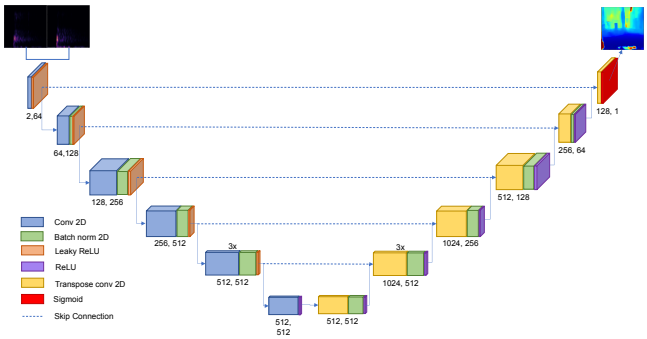

Fig. 8: U-Net Architecture. Input spectrograms are processed to depth maps. Skip-connections used show better performance even though the spatial correspondence between input and output features is not clear.

reproduces corridors robustly built with various material (e.g carpet floor and glass wall, stone and tiled floor, see Figure 7). Outdoors the performance is diminished with the network underestimating depth systematically. This shows some inevitable bias of the data having a majority of indoor data. Trained on BV1 data, corridors and obstacles are well reconstructed even though finer structures are equally lost.

V. Conclusion

Recent success in using simulated audio-visual data for scene understanding shows the potential of the audio modality. Depth prediction from sound alone or in addition to vision is possible, allowing to perceive the environment like a bat. The BatVision dataset will support the research community with large-scale real audio-visual data to improve task performance and uncover novel uses. We present an audio-only depth-prediction baseline as starting point and obtain good results when training a state-of-the-art approach on our data. Future datasets could include ultrasound *chirps* to enable inaudible human-robot collaboration.

VI. Acknowledgment

We thank all our collaborators, especially Daniel Lin for data collection, Jesper Haahr Christensen for insights into sonar and collaboration on the original idea, Karl Zipser for conception of the robot for BV1. Finally David Mazouz and Jacky Lech for advice and help building the robot for BV2. The authors acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-22-CE94-0003 (projet Omni-BatVision).

References

[1] Angelo Farina. "Simultaneous measurement of impulse response and distortion with a swept-sine technique". In: *Audio engineering society convention 108*. Audio Engineering Society. 2000.

[2] Elior Hadad et al. "Multichannel audio database in various acoustic environments". In: *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE. 2014, pp. 313–317.

[3] Adrien Gaidon et al. "Virtual worlds as proxy for multi-object tracking analysis". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4340–4349.

[4] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. "TUT database for acoustic scene classification and sound event detection". In: *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE. 2016, pp. 1128–1132.

[5] Andrew Owens et al. "Visually indicated sounds". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2405–2413.

[6] Angel Chang et al. "Matterport3D: Learning from RGB-D Data in Indoor Environments". In: *2017 International Conference on 3D Vision (3DV)*. IEEE. 2017.

[7] Hansung Kim et al. "Acoustic room modelling using a spherical camera for reverberant spatial audio objects". In: *Audio Engineering Society Convention 142*. Audio Engineering Society. 2017.

[8] Jaroslav Čmejla et al. "Mirage: Multichannel database of room impulse responses measured on high-resolution cube-shaped grid in multiple acoustic conditions". In: *arXiv preprint arXiv:1907.12421* (2019).

[9] Julian Straub et al. "The Replica dataset: A digital replica of indoor spaces". In: *arXiv preprint arXiv:1906.05797* (2019).

[10] Changan Chen et al. "Soundspaces: Audio-visual navigation in 3d environments". In: *Computer Vision–ECCV 2020, Proceedings*. Springer. 2020.

[11] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. "Batvision: Learning to see 3d spatial layout with two ears". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020.

[12] Jesper Haahr Christensen, Sascha Hornauer, and Stella Yu. "BatVision with GCC-PHAT Features for Better Sound to Vision Predictions". In: *Sight & Sound, CVPR Workshops* 2020 (2020).

[13] Ruohan Gao et al. "Visualechoes: Spatial image representation learning through echolocation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 658–676.

[14] Abhishek Kadian et al. "Sim2real predictivity: Does evaluation in simulation predict real-world performance?" In: *IEEE Robotics and Automation Letters* 5.4 (2020), pp. 6670–6677.

[15] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. "Semantic object prediction and spatial sound super-resolution with binaural sounds". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer. 2020, pp. 638–655.

[16] Diego Di Carlo et al. "dEchorate: a calibrated room impulse response dataset for echo-aware signal processing". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2021 (2021), pp. 1–15.

[17] Ziyang Chen, Xixi Hu, and Andrew Owens. "Structure from silence: Learning scene structure from ambient sound". In: *arXiv preprint arXiv:2111.05846* (2021).

[18] 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, pp. 8268–8277.

[19] Yannick Schulz et al. "Hearing what you cannot see: Acoustic vehicle detection around corners". In: *IEEE Robotics and Automation Letters* 6.2 (2021).

[20] Ethan Tracy and Navinda Kottege. "Catchatter: Acoustic perception for mobile robots". In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 7209–7216.

[21] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[22] Changan Chen et al. "Soundspaces 2.0: A simulation platform for visual-acoustic learning". In: *arXiv preprint arXiv:2206.08312* (2022).

[23] Kristen Grauman et al. "Ego4d: Around the world in 3,000 hours of egocentric video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18995–19012.

[24] Go Irie, Takashi Shibata, and Akisato Kimura. "Co-Attention-Guided Bilinear Model for Echo-Based Depth Estimation". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 4648–4652.

[25] Andrew Luo et al. "Learning Neural Acoustic Fields". In: *arXiv preprint arXiv:2204.00628* (2022).

[26] Archontis Politis et al. "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events". In: *arXiv preprint arXiv:2206.01948* (2022).

[27] Tao Sun et al. "SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[28] Chenghao Zhang et al. "Stereo Depth Estimation with Echoes". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*. Springer. 2022.

[29] Lingyu Zhu, Esa Rahtu, and Hang Zhao. "Beyond Visual Field of View: Perceiving 3D Environment with Echoes and Vision". In: *arXiv preprint arXiv:2207.01136* (2022).

[30] Dengxin Dai et al. "Binaural SoundNet: Predicting Semantics, Depth and Motion With Binaural Sounds".

In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2023), pp. 123–136.

[31] Sagnik Majumder et al. “Chat2Map: Efficient Scene Mapping from Multi-Ego Conversations”. In: *arXiv preprint arXiv:2301.02184* (2023).