Audio-Visual Depth and Material Estimation for Robot Navigation

Justin Wilson[1] and Nicholas Rewkowski[2] and Ming C. Lin[2]
https://cs.unc.edu/%7Ewilson/EchoCNN/

***Abstract*— Reflective and textureless surfaces such as windows, mirrors, and walls can be a challenge for scene reconstruction, due to depth discontinuities and holes. We propose an audio-visual method that uses the reflections of sound to aid in depth estimation and material classification for 3D scene reconstruction in robot navigation and AR/VR applications. The mobile phone prototype emits pulsed audio, while recording video for audio-visual classification for 3D scene reconstruction. Reflected sound and images from the video are input into our audio (EchoCNN-A) and audio-visual (EchoCNN-AV) convolutional neural networks for surface and sound source detection, depth estimation, and material classification. The inferences from these classifications enhance 3D scene reconstructions containing open spaces and reflective surfaces by depth filtering, inpainting, and placement of unmixed sound sources in the scene. Our prototype, demos, and experimental results from real-world with challenging surfaces and sound, also validated with virtual scenes, indicate high success rates on classification of material, depth estimation, and closed/open surfaces, leading to considerable improvement in 3D scene reconstruction for robot navigation.**

I. INTRODUCTION

Scenes containing open and reflective surfaces, such as windows and mirrors, are central to robot simultaneous localization and mapping (SLAM). They can also enhance AR/VR immersion in terms of both graphics and sound. However, they present a unique set of challenges. First, they are difficult to detect, map, and reconstruct due to their transparency and high reflectivity. Distinguishing between glass (e.g. window) and an opening in the space is an important part of the audio-visual experience for robot navigation, but also AR/VR engagement. In addition, illumination, background objects, and min/max depth ranges can be confounding factors.

Scene reconstructions for robot navigation and SLAM have led to advances in detection [1], segmentation [2], [3], and semantic understanding [4] and they are used to generate large-scale, labeled datasets of object [5] and scene [6], [7] geometric models to further aid training and sensing in a 3D environment. Advances have also been made to account for challenging surfaces [8], [9], [10]. Yet, scenes containing open and reflective surfaces, such as windows and mirrors, remain an open research area. Our work augments existing vision-based methods by adding audio context of surface detection, depth, and material estimation for recreating a digital scene from a real one.

Previous work has used sound to better understand objects in scenes. For instance, impact sounds from interacting with objects in a scene to perform segmentation [3] and to emulate the sensory interactions of human information processing [11]. Audio has also been used to compute material [12], object [11], scene [13], and acoustical [14] properties. Moreover, using both audio and visual sensory inputs has proven more effective; for example, multi-modal learning for object classification [15], [16] and object tracking [17].

Previous work has used sound to better understand objects in scenes. For instance, impact sounds from interacting with objects in a scene to perform segmentation [3] and to emulate the sensory interactions of human information processing [11]. Audio has also been used to compute material [12], object [11], scene [13], and acoustical [14] properties. Moreover, using both audio and visual sensory inputs has proven more effective; for example, multi-modal learning for object classification [15], [16] and object tracking [17].

Fusing multiple modalities, such as vision and sound, provide a wider range of possibilities than either single modality alone. In this work, we show that augmenting vision-based techniques with audio, called "EchoCNN," can detect open or reflective surfaces, its depth, and material, thereby enhancing 3D object and scene reconstruction for robots and AR/VR systems. We give an overview of our system pipeline in Sec. III and highlight key results below:

- EchoCNN, a fused audio-visual CNN architecture for classifying open/closed surfaces, their depth, and material or sound source placement (Section IV);
- Automated data collection process and audio-visual ground truth data for real-world (and synthetic) scenes containing windows and mirrors (Section V);
- Application demonstration using a staged audio-visual 3D reconstruction pipeline that uses EchoCNN to enhance scene geometry containing windows, mirrors, and open surfaces with depth filtering and inpainting based on EchoCNN inferences (Section VI).

Using EchoCNN, we have been able to achieve consistently higher accuracy in classification of open/closed surfaces, depth estimation, and materials in both real-world scenes and controlled experiments, resulting in considerably improved 3D scene reconstruction with glass doors, windows and mirrors (see Fig. 1).

II. RELATED WORK

We discuss recent work in audio-based classifications, echolocation, and existing techniques for reconstructing open and reflective surfaces here.

[1]Department of Computer Science, University of North Carolina at Chapel Hill, United States wilson@cs.unc.edu

[2]Department of Computer Science, University of Maryland at College Park, United States nick1,lin@umd.edu

Fig. 1. *Left*: ground truth image. Before (*Middle*) and after (*Right*) audio-augmented rendering of an indoor scene with open and closed reflective surfaces. EchoCNN enhances scene reconstruction through more accurate surface detection, depth estimation, and material classification based on audio-visual reflecting sound and image inputs. Green arrows highlight areas enhanced by our method.

A. Acoustic Imaging and Audio-based Classifiers

We begin with an introduction into sound propagation, room acoustics, and audio-visual classifiers.

Acoustics: various models have been developed to simulate sound propagation in a 3D environment, such as wave-based [19], ray tracing based [20], sound source clustering [21], multipole equivalent source methods [22], and a single point multipole expansion method [23], representing outgoing pressure fields. [24] uses acoustics and a smartphone for an app to detect car location and distance from walking pedestrians using temporal dynamics. [25] further discusses theory and applications of machine learning in acoustics. Computational imaging approaches have also used acoustics for non-line-of-sight imaging [26], 3D room geometry reconstruction from audio-visual sensors [27], and acoustic imaging on a mobile device [28]. To reconstruct windows and mirrors, our work uses room acoustics given the surface materials of the room [13] and distance from sound source. However, prior work and downstream processes often require a watertight reconstruction which can be difficult to generate in the presence of glass. Our approach addresses these issues using an integrated audio-visual CNN to detect discontinuity, depth, and materials.

Audio-based classification and reconstruction: using principles from sound synthesis, propagation, and room acoustics, audio classifiers have been developed for environmental sound [29], [30], [31], material [3], and object shape [11] classification. For audio-based reconstruction, Bat-G net uses ultrasonic echoes to train an auditory encoder and 3D decoder for 3D image reconstruction [32]. Audio input can take the form of raw audio, spectral shape descriptors [33], [34], [35], or frequency spectral coefficients that we also adopt. Our method uses reflecting sound to perform surface detection, depth, and material estimation.

Audio-visual learning: similar to its applications in natural language processing (NLP) and visual questing & answering systems [36], [37], [38], multi-modal learning using both audio-visual sensory inputs has also been used for classification tasks [15], [16], [39], material estimation [40], audio-visual zooming [41], and sound source separation [42], [43]. The latter having also isolated waves for specific generation tasks. Although similar in spirit, our audio-visual method differs from the existing methods by learning absorption and reflectance properties to detect a reflective surface, its depth, and material.

B. Glass and Mirror Reconstruction

Reflective surfaces produce identifiable audio and visual artifacts that can be used to help their detection. For example, researchers have developed algorithms to detect reflections in images taken through glass using correlations of 8-by-8 pixel blocks [44], image gradients [45], mirror edges based on content differences inside and outside of mirror [46], [47], two layer renderings [8], polarization imaging reflectometry [48], and diffraction effects [49]. Adding hardware, [50] uses ultrasonic sensor logic to track continuous wave ultrasound, [51] to detect obstacles such as glass and mirrors by using frequencies outside of the human audible range, and Amazon Echo [52] and Google Nest [53] use *ultrasound sensing* for motion detection. More recently, reflective surfaces have been detected by utilizing a mirrored variation of an AprilTag [54], [55]. [9] uses the reflective surface to their advantage by recognizing the AprilTag attached to their Kinect scanning device when it appears in the scene. Depth jumps and incomplete reconstructions have also been used [56]. However, vision based approaches require the right illumination, non-blurred imagery, and limited clutter behind the surface that may limit the reflection. We show that sound creates a distinct audio signal, providing complementary data about the presence of windows and mirrors without additional sensors.

III. OVERVIEW

Echo is defined as *distinct* reflections of the original sound with a sufficient sound level to be clearly heard above the general reverberation [57]. Although perceptible echo is abated because of precedence (known as the Haas effect) [58], returning sound waves are received after reflecting off of a solid surface. We use these distinct, reflecting sounds to design a staged approach of audio and audio-visual convolutional neural networks. EchoCNN-A and EchoCNN-AV can be used to estimate depth based on reverberation times, recognize material based on frequency and amplitude, and

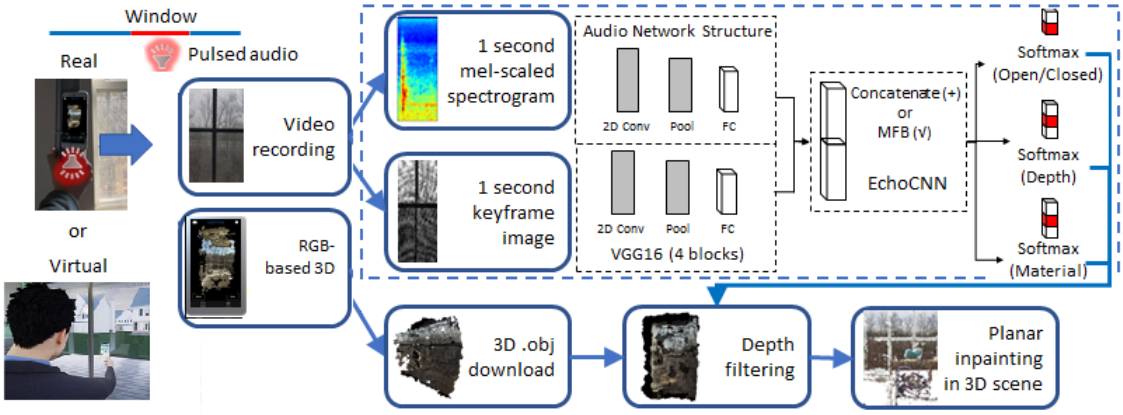

Fig. 2. *Staged approach* to estimate depth and materials for enhancing 3D scene and object reconstruction using audio-visual data. A smartphone emits and receives audio-visual signals for material/depth classification. It emits 100 ms pulsed audio and records video of the direct and reflecting sound. The receiving audio is split into 1.0 second intervals to allow for reverberation. Audio intervals are converted into mel-scaled spectrogram bins to reflect a logarithmic perception of frequency [15], [18]. They are passed through a multimodal convolutional neural network, EchoCNN, comprised of 2D convolutional, max pooling, fully connected, and softmax layers. EchoCNN informs hole filling steps to resolve planar discontinuities in scans caused by reflective surfaces, such as windows and mirrors. Binary classification is used for surface detection and multi-class classification is used for depth and material estimation.

handle both static and dynamic scenes with moving objects. All of which enhance scene and object reconstruction by detecting planar discontinuities from open or closed surfaces and then estimating depth and material.

A. Echolocation

Echolocation is the use of reflected sound to locate and identify objects, particularly used by animals like dolphins and bats [59]. This involves signal processing such as:

1) Doppler shift (the relative speed of the target),

$$\Delta f = f_D - f_0 = f_0 \frac{c_s}{c_0} cos(\theta) \tag{1}$$

2) time delay (distance to the target), and
3) frequency and amplitude in relation to distance (target object size and type recognition).

where the Doppler effect is the perceived change in frequency (Doppler frequency f_D minus transmitted frequency f_0) as a sound source with velocity c_s moves toward or away from the listener with velocity c_o and angle θ.

B. Staged Classification Pipeline

As depicted in Fig. 2, we take a staged approach to perform depth and material estimation for transparent surfaces, thereby enhancing scene and object reconstruction using audio-visual data. Our prototype system transmits and receives audio signals. Each audio emission is 100 ms of sound followed by 900 ms of silence to allow for the receiving microphone to capture reflections and reverberations (Subsection III-C). After the 3D scan is complete, an .obj file containing geometry and texture information is generated. 1 second frames are extracted from the recorded video to generate audio and visual input into the EchoCNN neural networks (Section IV). These networks are independently trained to detect whether a surface is open or closed, estimate depth to the surface from the sound source, and classify the material of the surface.

C. Sound Source

A smartphone emits recordings of human experimenter voice, whistle, hand clap, pure tones (ranging from 63 Hz to 16 kHz), chirps, and noise (white, pink, and brownian). All of which can be generated as either pulsed (PW) or continuous waves (CW). PW is preferred for theoretical and empirical reasons. First, the transmission frequency f_0 may experience considerable downshift as a result of absorption and diffraction effects [59]. Therefore, using pulsed waves independent for each emission is theoretically better than continuous waves compared to f_0. Section VI-B shows superior PW results over CW for given classification tasks.

Pure tones were generated with default 0.8 out of 1 amplitudes using the Audacity computer program and center frequencies of 63 Hz, 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, 8 kHz, and 16 kHz. These octave bands were used for training to model the experimenter whistle and voice in our simulation tests so live instead of recorded sound may be used. Human voice ranges from about 63 Hz to 1 kHz [58] (125 Hz to 8 kHz [57]) and an untrained whistler between 500 Hz to 5 kHz [60]. Chirps were linearly interpolated from 440 Hz to 1320 Hz in 100 ms. A hand clap is an impulsive sound that yields a flat spectrum [58]. All sound sources were recorded and played back with max volume. While playback of recorded sounds were used for consistency, live audio for augmentation and ease of use may also be used. Please see our supplementary materials for spectrograms across all sound sources.

Audio input: audio was generated in pulsed waves (PW). A smartphone emits sound and captures video which is used offline to perform a RGB-based reconstruction. 1 second audio frames is based on the Sabine Formula [61], [57] of reverberation time for a compact room calculated as:

$$T = 0.05 \frac{V}{a} = 0.05 \frac{V}{\sum S\alpha} = (0.05 \frac{sec}{ft}) \frac{1,296\ ft^3}{69.23\ ft^2} = 0.94\ sec \tag{2}$$

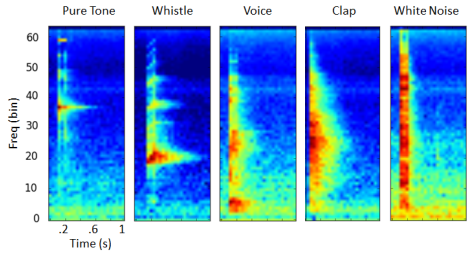

Fig. 3. Mel-scaled spectrograms of recorded impulses of different sound sources used. *From left to right*: narrow to disperse spectra. Not shown are other pure tone frequencies, chirp, pink noise, and brownian noise. Horizontal axis is time and vertical axis is frequency.

where T is the reverberation time (time required for sound to decay 60 dB after source has stopped), V is room volume (ft^3), and a is the total room absorption at a given frequency (e.g. 250 Hz). For the bathroom scene, $V = 9\ ft{*}16\ ft{*}9\ ft = 1,296\ ft^3$ and $a = 69.23\ ft^2$, which is the sum of sound absorption from the materials.

Visual input: images were captured from the same smartphone video as the audio recordings. Each corresponding image was cropped and grayscaled to reduce computational requirements. Image dimensions were 64 by 25 pixels. Visual data served as inputs for visual only and audio-visual model variation EchoCNN-AV.

IV. MODEL ARCHITECTURE

To augment visually based approaches, we use a multimodal CNN with mel-scaled spectrogram and image inputs. First, we perform surface detection to determine if a space with depth jumps and holes is in error or in fact open (i.e. open/closed classification). In the event of error, we estimate distance from recorder to surface using audio-visual data for depth filtering and inpainting. Finally, we determine the material. All of these classifications are performed using our audio and audio-visual convolutional neural networks, EchoCNN-A and EchoCNN-AV (Fig. 2). A CNN architecture was used for distinct audio-visual features.

Audio sub-network: our frame-based EchoCNN-A consists of a single convolutional layer with 262 filters and 3x5 kernel size followed by two dense layers with feature normalization. Sampled at $F_s = 44.1\ kHz$ to cover the full audible range, audio frames are 1 second mel-scaled spectrograms with Short-Time Fourier Transform (STFT) coefficients χ (Eq. IV). Mel-scale is a logarithmic transformation of a signal's frequency and decibel scale such that it is of equal distance to untrained humans ability to perceive and distinguish that frequency. Mel-scale spectrograms have been demonstrated to perform well as inputs into convolutional neural networks (CNNs) according to comparison [62].

Each audio example is classified independently and 1 second intervals to reflect an estimated reverberation time based on a compact room size (Eq. 2). With a 2048 sample Hann window (N), 25% overlap, and hop length ($H = 2048/4$) for spectrogram parameters, this results in a frequency dimension of 21.5 Hz (Eq. IV) and temporal dimension of 12 ms (Eq. IV) or 12% of each 100 ms pulsed audio. Each spectrogram is normalized and downsampled to a size of 62 frequency bins by 25 time bins.

STFT divides a time signal into segments of equal length and then compute the Fourier transform to determine frequency and phase over time. We define these frequency spectral coefficients [63] as:

$$\chi(m,k) = \sum_{n=0}^{N-1} x(n+mH)w(n)exp(-2\pi ikn/N) \quad (3)$$

for m^{th} time frame and k^{th} Fourier coefficient with real-valued DT signal $x : Z \rightarrow R$, sampled window function $w(n)$ *for* $n \in [0 : N-1] \rightarrow R$ of length N, and hop size $H \in N$ [63]. R denotes continuous time and Z denotes discrete time. Equal to $|\chi(m,k)|^2$, a spectrogram's horizontal axis is time and vertical axis is frequency.

$$F_{coef}(k) = \frac{k \cdot F_s}{N} = k\frac{44100}{2048} = k * 21.5\ Hz \quad (4)$$

$$T_{coef}(m) = \frac{m \cdot H}{F_s} = m\frac{2048 * 0.25}{44100} = m * 0.012\ seconds \quad (5)$$

A hop length of $H = N/2$ achieves a reasonable temporal resolution and data volume of generated spectral coefficients [63]. Temporal resolution is important in order to detect when a reflecting sound reaches the receiver. Therefore, we decided to use a shorter window length $N = 2048$ instead of $N = 4096$ for instance. This resulted in a shorter hop length and accepting the trade-off of a higher temporal dimension for increased data volume.

Visual sub-network: while audio information is generally useful for all three classifications tasks (Table I), visual information is particularly useful to aid material classification. We use AlexNet [64] as a visual-based baseline to compare to our audio and audio-visual methods. It also serves as a visual subnetwork and input into our audio-visual merge layer. AlexNet was chosen as the visual baseline because of its diverse set of classes and privacy-aware visual recognition, according to image-net.org.

Merge layer: we evaluated concatenation and multi-modal factorized bilinear (MFB) pooling [65] to fuse audio and visual fully connected layers. Concatenation of the two vectors serves as a straightforward baseline. MFB allows for additional learning in the form of a weighted projection matrix factorized into two low-rank matrices.

$$z_i = x^T W_i y = x^T U_i V_i^T y = 1^T (U_i^T x \ \circ \ V_i^T y) \quad (6)$$

where k is the factor or latent dimensionality with index i of the factorized matrices, $\circ$ is the Hadmard product or element-wise multiplication, and $1 \in R^k$ is an all-one vector.

A. Loss Function

For open/closed predictions, categorical cross entropy loss (Eq. 7) is used instead of binary to allow for estimating the extent of the surface opening (e.g. all the way open, halfway, or closed). A regression model is not used for depth

Fig. 4. Several different real-world scenes used in data collection and testing our audio-visual classification system. Listener at different distances from sound source (1 ft, 2 ft, and 3 ft). For more real-world scene figures and virtual scenes (from 0.5 m to 3 m), please see additional supplementary material at `https://cs.unc.edu/%7Ewilson/EchoCNN/`

estimation because ground truth data is collected in discrete 0.5 m or 1 ft increments within the free field for better noise reduction [57]. The Softmax function is used for output activations.

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{7}$$

where M is number of classes, y indicator for correct classification, and p for predicted probability that observation (o) is of class (c).

B. Depth filtering and planar inpainting

The outputs of our EchoCNN inform enhancements for 3D reconstruction. If depth jumps in the reconstruction are first classified as an open surface, then no change is required other than filtering loose geometry and small components. Else, there is a planar discontinuity (e.g. window or mirror) that needs to be filled. With depth estimated by EchoCNN, we filter the initial 3D mesh to within a user specified threshold of that depth. This gives us the plane needed to fill. Finally, EchoCNN classifies its surface material which can then be applied to the filled plane.

V. DATASETS

We implemented all EchoCNN and baseline models with Tensorflow [69] and Keras [70]. Training was performed using a TITAN X GPU running on Ubuntu 16.04.5 LTS. We used categorical cross entropy loss with Stochastic Gradient Descent optimized by ADAM [71]. Using a batch size of 32, remaining hyperparameters were tuned manually based on a separate validation set. We make our real-world and synthetic datasets available to aid future research in this area.

Our audio-based EchoCNN-A and audio-visual EchoCNN-AV CNNs are trained with center frequencies 63 Hz, 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, 8 kHz, and 16 kHz. Training uses these pulsed pure tone impulses and experimenter hand clap. Some time shift of sound start is allowed. The hold-out test data is comprised of sound sources excluded from training - white noise, experimenter whistle, and voice. While the test set contains sound sources not in training to evaluate generalization, their frequencies are covered given training across nine octave bands. Hold out scenes are also tested (see supplemental video).

A. Real-World Data

Training data is comprised of 1 second pulsed spectrograms (Fig. 3) from recorded pure tones, experimenter hand claps, brownian noise, and pink noise (N=857). Training and test examples were collected via video recordings and labeled for material, open/closed, and 1 ft depth increments based on surface distance. Nine octaves of pure tones, hand claps, and white noise cover a wide range of frequencies used to train our models.

The hold-out test dataset consists of 1-second pulsed spectrograms from recorded experimenter voice, whistle, chirp, and white noise (N=431). Voice and whistle recordings were chosen for the hold out test set to ease future transition to live and hands-free emitted sounds during reconstruction. Hold-out test data is excluded from training and only evaluated during testing. The same hold-out sets were used for visual and audio-visual evaluation. Note that the train-test split was based on audio since unheard sound from a different sound source in test may have the same visual appearance in training.

VI. IMPLEMENTATION & EVALUATION

A. Experimental setup

The listener and sound source device is a smartphone that emits pulsed waves at 3 feet, 2 feet, and 1 feet away from the reconstructing surface. Three feet was selected to remain in the free field. Beyond that, there will be less noise reduction due to reflecting sounds in the reverberant field [57]. Within a few feet of the reconstructing surface also create finer detail reconstructions.

We labeled our data based on scene, sound source, and surface properties - type of surface, material, and depth from sound source. The training set included pulsed sounds of pure tone frequencies, a hand clap, brownian noise, and pink noise. The hold-out test set consisted of voice, whistle, chirp, and white noise. For rooms with different sound-absorbing treatments, our real-world recordings include a bedroom (e.g. carpet and painted) and bathroom (e.g. tiled).

Real-World Test Scenes: We validated the results against the known ground truth through controlled experiments using synthetic scenes (see appendices posted at the project website). We also tested our system on several real-world scenes (see Fig. 1 and 4, and the supplementary video).

Accuracy of Reflecting Sounds used for Open/Closed, Depth Estimation, and Material Classification in Real Scenes

		Open/Closed		Depth Estimation				Sound Material		
Method	Input	Shower	Window	Overall	3 ft	2 ft	1 ft	Overall	Glass	Mirror
kNN [66]	A	56.5%	**100%**	21.3%	16%	21%	25%	44.0%	47.5%	52.4%
Linear SVM [67]	A	61.5%	91.7%	37.6%	38%	32%	41%	51.9%	46.0%	57.1%
SoundNet5 [68]	A	45.2%	46.6%	39.7%	40%	71%	8%	71.0%	98.4%	1.6%
SoundNet8 [68]	A	50.7%	46.6%	42.5%	92%	0%	33%	44.4%	16.4%	85.7%
EchoCNN-A (Ours)	A	**71.2%**	**100%**	**71.8%**	86%	54%	**76%**	**77.4%**	62.3%	**92.0%**
AlexNet [64]	V	78.1%	96.1%	45.2%	52%	83%	0%	80.6%	60.7%	100%
Acoustic Classification [13]	AV	N/A	N/A	N/A	N/A	N/A	N/A	———- 48% * ———-		
EchoCNN-AV Cat (Ours)	AV	**100%**	**100%**	**89.5%**	**95%**	**100%**	73%	**100%**	**100%**	**100%**
EchoCNN-AV MFB (Ours)	AV	**100%**	**100%**	84.9%	54%	**100%**	**100%**	80.6%	60.7%	**100%**

TABLE I

MULTIPLE MODELS (**OURS IS ECHOCNN**) AND BASELINES WERE EVALUATED FOR AUDIO AND AUDIO-VISUAL ANALYSIS. OVERALL, **71.2%** OF HELD OUT REFLECTING SOUNDS AND **100%** OF AUDIO-VISUAL FRAMES WERE CORRECTLY CLASSIFIED AS AN OPEN OR CLOSED INTERIOR SURFACE (I.E. GLASS WINDOW). OPEN/CLOSED CLASSIFICATION IS EVEN HIGHER FOR EXTERNAL FACING WINDOWS DUE TO OUTSIDE NOISE. **71.8%** OF 1-SECOND AUDIO FRAMES WERE CORRECTLY CLASSIFIED AS 1 FT, 2 FT, OR 3 FT AWAY FROM SURFACE BASED ON AUDIO ALONE; **89.5%** WHEN CONCATENATING WITH ITS CORRESPONDING IMAGE. FINALLY, **77.4%** AND **100%** OF AUDIO AND AUDIO-VISUAL INPUTS CORRECTLY LABELED THE SURFACE MATERIAL. * ACCORDING TO [13], 48% OF THE TRIANGLES IN ITS SCENES ARE CORRECTLY CLASSIFIED, WHERE ITS CLASSIFICATION IS MORE GRANULAR.

Activation Maximization Activation maximization generates an input that maximizes layer activations for a given class. This provides insights into the types of patterns the neural network is learning. Supplementary material shows the different inputs that would maximize EchoCNN activations for depth estimation. Notice lower frequencies tend to occur at 3 ft (longer reverberation times) than at 1 and 2 ft (high frequencies) due to the typical high frequency damping and absorption.

B. Results

Overall, **71.2%** of hold out reflecting sounds and **100%** of audio-visual frames were correctly classified as an open or closed boundary in the home (Table I). **71.8%** of 1 second audio frames were correctly classified as 1 ft, 2 ft, or 3 ft away from the surface based on audio alone; **89.5%** when concatenating with its corresponding image. Finally, **77.4%** of audio and **100%** of audio-visual inputs correctly labeled the surface material (glass, mirror, wall).

AlexNet, a visual only baseline, is higher at 78.1% than audio-only EchoCNN-A for open/closed classification. This is partly due to the fact that the hold out set was to test audio generalization (i.e. unheard sound sources). But unheard sound sources does not guarantee unseen visual data. In other words, different sound does not mean different appearance. Therefore, images similar to those found in training are also present in test which may help explain this observation.

C. Analysis

According to [58], 10 dB of exterior to interior noise reduction can be attributed to closed compared to open windows. Using audio, we also noticed noise reduction between winter and spring due to more foliage on the trees. We also observed flutter echoes, which can be heard as a "rattle" or "clicking" from a hand clap and have been simulated in spatial audio [72]. They became more pronounced the closer to the wall surface in the bathroom scene. Audio is unable to augment failure cases of the shower from initial RGB-based reconstructions using either [73] or [74]. (Background UV textures are placed at a fixed 1 ft (0.3 m) behind estimated surface depth.) We compare our 3D reconstructions to depth estimates based on related work (see Table I).

VII. CONCLUSION AND FUTURE WORK

This work introduces the first audio and audio-visual techniques for enhancing robot localization, mapping and 3D scene reconstruction with windows and mirrors. Our staged pipeline emits and receives pulsed audio from a variety of sound sources for surface detection, depth estimation, and material classification. These classifications can assist in resolving planar discontinuities caused by open spaces and reflective surfaces using depth filtering and planar filling. Our system performs well compared to baseline methods given experiment results for real-world and virtual scenes containing windows, mirrors, and open surfaces. We intend to release our audio-visual data, in addition to reflection separation data for future research.

Limitations: Instead of a staged pineline, an integrated, end-to-end pipeline may further improve 3D scene reconstruction and simultaneous localization and mapping. With a defined set of output classes for EchoCNN, alternative baselines such as Non-Negative Matrix Factorization (NMF), source separation techniques, and the pYIN algorithm [75] may be able to extract the fundamental frequency f_0, i.e. the frequency of the lowest partial of the sound, to further improve the robustness of the results. The effects of ambient and directed sound sources have not been understood and may also enhance results for sound source separation. Finally, our current implementation holds out voice and whistle data, which is different from the audio used during training. Some insights may be gained by experimenting with a different training dataset for testing audio-only, visual-only, and audio-visual methods.

REFERENCES

[1] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," *CoRR*, vol. abs/1611.05267, 2016. [Online]. Available: http://arxiv.org/abs/1611.05267

[2] S. Golodetz*, M. Sapienza*, J. P. C. Valentin, V. Vineet, M.-M. Cheng, A. Arnab, V. A. Prisacariu, O. Kähler, C. Y. Ren, D. W. Murray, S. Izadi, and P. H. S. Torr, "SemanticPaint: A Framework for the Interactive Segmentation of 3D Scenes," Department of Engineering Science, University of Oxford, Tech. Rep. TVG-2015-1, October 2015, released as arXiv e-print 1510.03727.

[3] A. Arnab, M. Sapienza, S. Golodetz, J. Valentin, O. Miksik, S. Izadi, and P. Torr, "Joint object-material category segmentation from audio-visual cues," in *Proceedings of the British Machine Vision Conference (BMVC)*, 09 2015.

[4] S. Song, F. Yu, A. Zeng, A. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 07 2017, pp. 190–198.

[5] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," pp. 1912–1920, 06 2015.

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[7] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. Engel, R. Mur-Artal, C. Y. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. A. Newcombe, "The replica dataset: A digital replica of indoor spaces," *ArXiv*, vol. abs/1906.05797, 2019.

[8] S. Sinha, J. Kopf, M. Goesele, D. Scharstein, and R. Szeliski, "Image-based rendering for scenes with reflections," *ACM Transactions on Graphics - TOG*, vol. 31, pp. 1–10, 07 2012.

[9] T. Whelan, M. Goesele, S. J. Lovegrove, J. Straub, S. Green, R. Szeliski, S. Butterfield, S. Verma, and R. Newcombe, "Reconstructing scenes with mirror and glass surfaces," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 102:1–102:11, July 2018. [Online]. Available: http://doi.acm.org/10.1145/3197517.3201319

[10] R. Chabra, J. Straub, C. Sweeney, R. A. Newcombe, and H. Fuchs, "Stereodrnet: Dilated residual stereo net," *CoRR*, vol. abs/1904.02251, 2019. [Online]. Available: http://arxiv.org/abs/1904.02251

[11] Z. Zhang, J. Wu, Q. Li, Z. Huang, J. Traer, J. H. McDermott, J. B. Tenenbaum, and W. T. Freeman, "Generative modeling of audible shapes for object perception," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1260–1269, 2017.

[12] Z. Ren, H. Yeh, and M. C. Lin, "Example-guided physically based modal sound synthesis," *ACM Trans. Graph.*, vol. 32, no. 1, pp. 1:1–1:16, Feb. 2013. [Online]. Available: http://doi.acm.org/10.1145/2421636.2421637

[13] C. Schissler, C. Loftin, and D. Manocha, "Acoustic classification and optimization for multi-modal rendering of real-world scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 1246–1259, 2018.

[14] Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha, "Scene-aware audio rendering via deep acoustic analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1991–2001, 2020.

[15] A. Sterling, J. Wilson, S. Lowe, and M. C. Lin, "Isnn: Impact sound neural network for audio-visual object classification," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 578–595.

[16] J. Wilson, A. Sterling, and M. Lin, "Analyzing liquid pouring sequences via audio-visual neural networks," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11 2019, pp. 7702–7709.

[17] J. Wilson and M. C. Lin, "Avot: Audio-visual object tracking of multiple objects for robotics," in *ICRA 2020*, 2020.

[18] K. Prahallad, "Spectrogram, cepstrum and melfrequency analysis," 2011.

[19] R. Mehra, A. Rungta, A. Golas, M. Lin, and D. Manocha, "Wave: Interactive wave-based sound propagation for virtual environments," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 21, pp. 434–442, 04 2015.

[20] A. Rungta, C. Schissler, R. Mehra, C. Malloy, M. Lin, and D. Manocha, "Syncopation: Interactive synthesis-coupled sound propagation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 4, pp. 1346–1355, Apr. 2016. [Online]. Available: https://doi.org/10.1109/TVCG.2016.2518421

[21] N. Tsingos, E. Gallo, and G. Drettakis, "Perceptual audio rendering of complex virtual environments," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 249–258, Aug. 2004. [Online]. Available: http://doi.acm.org/10.1145/1015706.1015710

[22] D. L. James, J. Barbič, and D. K. Pai, "Precomputed acoustic transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources," in *ACM SIGGRAPH 2006 Papers*, ser. SIGGRAPH '06. New York, NY, USA: ACM, 2006, pp. 987–995. [Online]. Available: http://doi.acm.org/10.1145/1179352.1141983

[23] C. Zheng and D. L. James, "Toward high-quality modal contact sound," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 38:1–38:12, July 2011. [Online]. Available: http://doi.acm.org/10.1145/2010324.1964933

[24] D. Godoy, B. Islam, S. Xia, M. T. Islam, R. Chandrasekaran, Y.-C. Chen, S. Nirjon, P. Kinget, and X. Jiang, "Paws: A wearable acoustic system for pedestrian safety," in *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 04 2018, pp. 237–248.

[25] M. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. Roch, S. Gannot, and C. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, pp. 3590–3628, 11 2019.

[26] D. B. Lindell, G. Wetzstein, and V. Koltun, "Acoustic non-line-of-sight imaging," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2019, pp. 6773–6782.

[27] H. Kim, L. Remaggi, P. J. B. Jackson, F. M. Fazi, and A. Hilton, "3d room geometry reconstruction using audio-visual sensors," in *2017 International Conference on 3D Vision (3DV)*, 10 2017, pp. 621–629.

[28] W. Mao, M. Wang, and L. Qiu, "Aim: Acoustic imaging on a mobile," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 468–481. [Online]. Available: https://doi.org/10.1145/3210240.3210325

[29] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 03 2017, pp. 776–780.

[30] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 1015–1018. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806390

[31] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 1041–1044. [Online]. Available: http://doi.acm.org/10.1145/2647868.2655045

[32] G. Hwang, S. Kim, and H.-M. Bae, "Bat-g net: Bat-inspired high-resolution 3d image reconstruction using ultrasonic echoes," in *NeurIPS*, 2019.

[33] B. Michael, A. Silvia, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 18, 01 2005.

[34] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895 – 2907, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865503001478

[35] J. O. Smith III, "Physical audio signal processing," https://ccrma.stanford.edu/~jos/pasp/, 2020.

[36] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," 2016.

[37] H. Kim, H. Tan, and M. Bansal, "Modality-balanced models for visual dialogue," 2020.

[38] D. Hannan, A. Jain, and M. Bansal, "Manymodalqa: Modality disambiguation and qa over diverse inputs," 2020.

[39] A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Ambient*

Sound Provides Supervision for Visual Learning, vol. 9905, 10 2016, pp. 801–816.

[40] A. Davis*, K. L. Bouman*, J. G. Chen, M. Rubinstein, O. Büyüköztürk, F. Durand, and W. T. Freeman, "Visual vibrometry: Estimating material properties from small motions in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 732–745, 2017.

[41] A. A. Nair, A. Reiter, C. Zheng, and S. Nayar, "Audiovisual zooming: What you see is what you hear," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019, pp. 1107–1118. [Online]. Available: http://doi.acm.org/10.1145/3343031.3351010

[42] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *CoRR*, vol. abs/1804.03619, 2018. [Online]. Available: http://arxiv.org/abs/1804.03619

[43] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, ser. NIPS'00. Cambridge, MA, USA: MIT Press, 2000, pp. 535–541. [Online]. Available: http://dl.acm.org/citation.cfm?id=3008751.3008829

[44] Y. Shih, D. Krishnan, F. Durand, and W. Freeman, "Reflection removal using ghosting cues," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2015, pp. 3193–3201.

[45] J. Kopf, F. Langguth, D. Scharstein, R. Szeliski, and M. Goesele, "Image-based rendering in the gradient domain," *ACM Transactions on Graphics (TOG)*, vol. 32, pp. 199:1–199:9, 11 2013.

[46] J. Lin, G. Wang, and R. W. Lau, "Progressive mirror detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[47] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin, and R. W. Lau, "Where is my mirror?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[48] J. Riviere, I. Reshetouski, L. Filipi, and A. Ghosh, "Polarization imaging reflectometry in the wild," *ACM Transactions on Graphics (TOG)*, vol. 36, pp. 1 – 14, 2017.

[49] A. Toisoul and A. Ghosh, "Practical acquisition and rendering of diffraction effects in surface reflectance," *ACM Transactions on Graphics*, vol. 36, p. 1, 07 2017.

[50] I. E. Sutherland, "A head-mounted three dimensional display," in *AFIPS '68 (Fall, part I)*, 1968.

[51] Y. Zhang, M. Ye, D. Manocha, and R. Yang, "3d reconstruction in the presence of glass and mirrors by acoustic and visual fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 07 2017.

[52] Amazon, "What is ultrasound motion detection?" [Online]. Available: https://www.amazon.com/gp/help/customer/display.html?nodeId=GSR22RYDWS3KBUYW

[53] A. Udall, "How ultrasound sensing makes nest displays more accessible," 12 2019. [Online]. Available: https://blog.google/products/google-nest/ultrasound-sensing/

[54] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *2011 IEEE International Conference on Robotics and Automation*, 06 2011, pp. 3400 – 3407.

[55] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10 2016, pp. 4193–4198.

[56] I. Lysenkov, V. Eruhimov, and G. R. Bradski, "Recognition and pose estimation of rigid transparent objects with a kinect sensor," in *Robotics: Science and Systems*, 2012.

[57] M. D. Egan, *Architectural Acoustics*. McGraw-Hill Custom Publishing, 1988.

[58] M. Long, *Architectural Acoustics*, 2nd ed. Academic Press, 2014.

[59] T. L. Szabo, "Chapter 12 - nonlinear acoustics and imaging," in *Diagnostic Ultrasound Imaging: Inside Out (Second Edition)*, T. L. Szabo, Ed. Boston: Academic Press, 2014, pp. 501 – 563. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123964878000124

[60] M. Nilsson, J. Bartunek, J. Nordberg, and I. Claesson, "Human whistle detection and frequency estimation," *Image and Signal Processing, Congress on*, vol. 5, pp. 737–741, 05 2008.

[61] R. W. Young, "Sabine reverberation equation and sound power calculations," 1959.

[62] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *CoRR*, vol. abs/1706.07156, 2017. [Online]. Available: http://arxiv.org/abs/1706.07156

[63] M. Mller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, 1st ed. Springer Publishing Company, Incorporated, 2015.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[65] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," *CoRR*, vol. abs/1708.01471, 2017. [Online]. Available: http://arxiv.org/abs/1708.01471

[66] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 01 1967.

[67] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, Y. Lechevallier and G. Saporta, Eds. Paris, France: Springer, 01 2010, pp. 177–187. [Online]. Available: http://leon.bottou.org/papers/bottou-2010

[68] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016.

[69] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[70] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[72] T. Halmrast, "A very simple way to simulate the timbre of flutter echoes in spatial audio," in *EAA Spatial Audio Signal Processing Symposium*, 2019.

[73] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys, "Live metric 3d reconstruction on mobile phones," in *2013 IEEE International Conference on Computer Vision*, 12 2013, pp. 65–72.

[74] Metashape, "Agisoft metashape standard," 2020. [Online]. Available: https://www.agisoft.com/downloads/installer/

[75] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 05 2014, pp. 659–663.