CHAPTER 6: AUDIO-AUGMENTED SCENE RECONSTRUCTION ON MOBILE DEVICES[1]

This chapter describes echoreconstruction, an audio-visual method that uses the reflections of sound to aid in geometry and audio reconstruction. The mobile phone prototype emits pulsed audio while recording video for RGB-based 3D reconstruction and audio-visual classification. 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 scene 3D reconstructions containing open spaces and reflective surfaces by depth filtering, inpainting, and placement of unmixed sound sources in the scene.

6.1 Introduction

Reconstruction techniques have enabled significant contributions in detection (117), segmentation (67; 9), and semantic understanding (217). They have also been used to generate large-scale, labeled datasets of object (252) and scene (48) geometric models to further aid training and sensing in a 3D environment. However, scenes containing open and reflective surfaces, such as windows and mirrors, can present a unique set of challenges. First, they are difficult to detect and reconstruct due to their transparency and

[1] This chapter is currently under review.

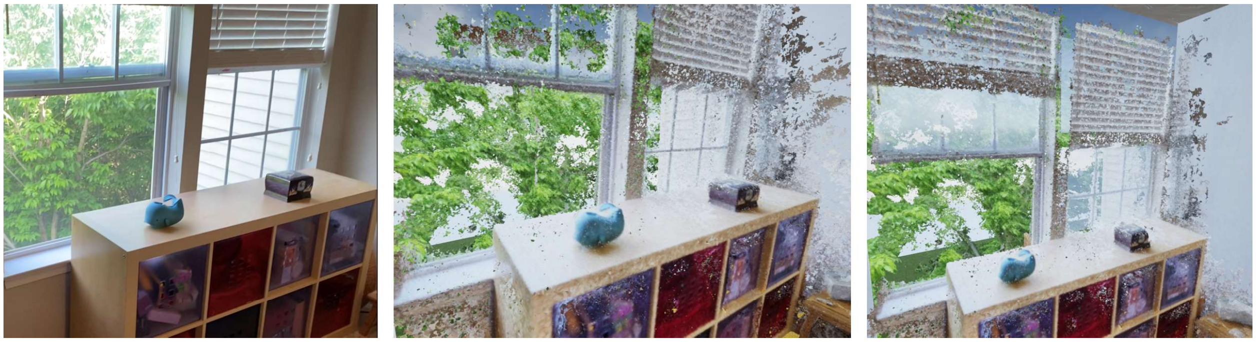

Figure 6.1: *Left*: ground truth image. Before (*Middle*) and after (*Right*) audio-augmented rendering of an indoor scene with open and closed reflective surfaces. The reconstruction is enhanced by EchoCNN inferences of surface detection, depth estimation, and material classification based on audio-visual reflecting sound and image inputs.

high reflectivity. Distinguishing between glass (e.g. window) and an opening in the space is an important part of the audio-visual experience. Finally, illumination, background objects, and min/max depth ranges can be confounding factors. While advances have been made to account for these challenging surfaces (213; 242; 33), our work augments these state-of-the-art visual methods by adding an audio context of surface detection, depth, and material estimation.

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 (9) and neural networks to emulate the sensory interactions of human information processing (268). Audio has also been used to automatically compute material (191), object (268), scene (202), and acoustical (224) properties. Better still, using both audio and visual sensory inputs has been shown to be even more effective; for example, multi-modal learning for object classification (220; 246) and object tracking (244).

Fusing multiple modalities, such as vision and sound, provide a wider range of possibilities than either single modality alone. In this work, we demonstrate that augmenting vision-based techniques with audio, referred to as "EchoCNN," can detect open and reflective surfaces, its depth, and material, thereby enhancing 3D object and scene reconstruction. We highlight some key results below:

- EchoCNN, a fused audio-visual CNN architecture for classifying open/closed surfaces, their depth, and material;
- EchoReconstruction, a staged audio-visual 3D reconstruction pipeline that uses mobile phones to enhance scene geometry containing windows, mirrors, and open surfaces with depth filtering and inpainting based on EchoCNN inferences;
- Semantic rendering of window and mirror in audio-augmented reconstructions based on point of view (e.g. environment outside of the window or reflected view of a TV);
- Real and synthetic audio-visual ground truth data for multiple scenes containing windows and mirrors in addition to reflection separation data (direct, early, or late reverberations).

6.2 Related Work

Previous research in 3D reconstruction, audio-based classifications, and echolocation are discussed in this section in addition to existing techniques for reconstructing open and reflective surfaces.

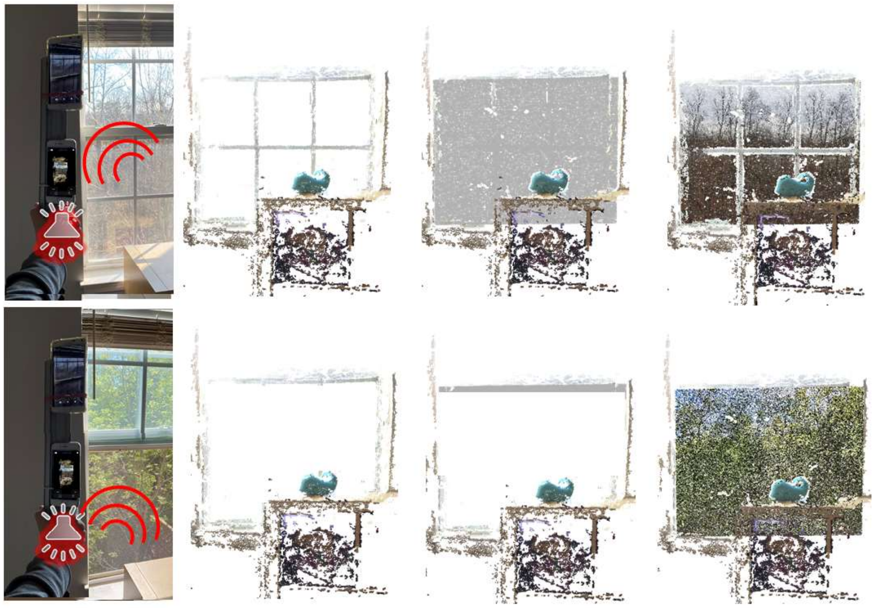

Figure 6.2: *Top row*: closed window in winter. *Bottom row*: opened in spring. *Column 1*: mobile echoreconstruction prototype; the bottom phone emits pulsed audio and performs a RGB-based 3D reconstruction (live (225) or photogrammetric (144)); the top phone records video. *Column 2*: initial reconstruction based on state-of-the-art commercially available Astrivis app. *Column 3*: our audio-visual EchoCNN convolutional neural network classifies open or closed surface, depth, and material. *Column 4*: semantic reconstruction of the window accounting for EchoCNN inferences.

6.2.1 3D reconstruction

Object and scene reconstruction methods generate 3D scans using RGB and RGB-D data. For example, Structure from Motion (SFM) (241), Multi-View Stereo (MVS) (205), and Shape from Shading (264) are all techniques to scan a scene and its objects. Static (154; 67) and dynamic (155; 49) scenes can also be scanned in real-time using commodity sensors such as the Microsoft Kinect and GPU hardware. 3D scene reconstructions have also been performed with sound based on time of flight sensing (46). Not only has this previous research generated large amounts of 3D scene (209; 217) and object (212; 115; 252) data, they also benefit from these datasets by using them for training vision-based neural networks for classification, segmentation, and other downstream tasks. Depth estimation algorithms (57; 4; 33) also create 3D reconstructions by fusing depth maps using ICP and volumetric fusion (93).

6.2.1.1 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

Example 3D Reconstruction Methods

Type	Methods
Active (RGB-D)	KinectFusion, DynamicFusion, BundleFusion
Passive (RGB)	SLAM, SFM, (225), ScanNet, (242)
Stereo	MVS, StereoDRNet
Lidar	(100)
Ultrasonic	(265)
Time of flight	(46)

Table 6.1: 3D reconstruction methods by type such as passive (RGB), active (RGB-D), or other sensor (e.g. ultrasound, lidar, etc.); single or multiple views; and static or dynamic scenes.

glass using correlations of 8-by-8 pixel blocks (208), image gradients (108), and two layer renderings (213). (222) used ultrasonic sensor logic to track continuous wave ultrasound and (265) to detect obstacles such as glass and mirrors by using frequencies outside of the human audible range. More recently, reflective surfaces have been detected by utilizing a mirrored variation of an AprilTag (163; 237). (242) use 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 (136). 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 reconstruction methods complementary data about the presence of windows and mirrors without additional sensors.

6.2.2 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 (142), ray tracing based (195), sound source clustering (228), multipole equivalent source methods (98), and single point multipole expansion (272), representing outgoing pressure fields. (66) uses acoustics and a smartphone for an app to detect car location and distance from walking pedestrians using temporal dynamics. (20) further discusses theory and applications of machine learning in acoustics. Computational imaging approaches have also used acoustics for non-line-of-sight imaging (127), 3D room geometry reconstruction from audio-visual sensors (101), and acoustic imaging on a mobile device (137). To reconstruct windows and mirrors, our work uses room acoustics given the surface materials of the room (202) and distance from sound source. However, prior work and downstream pro-

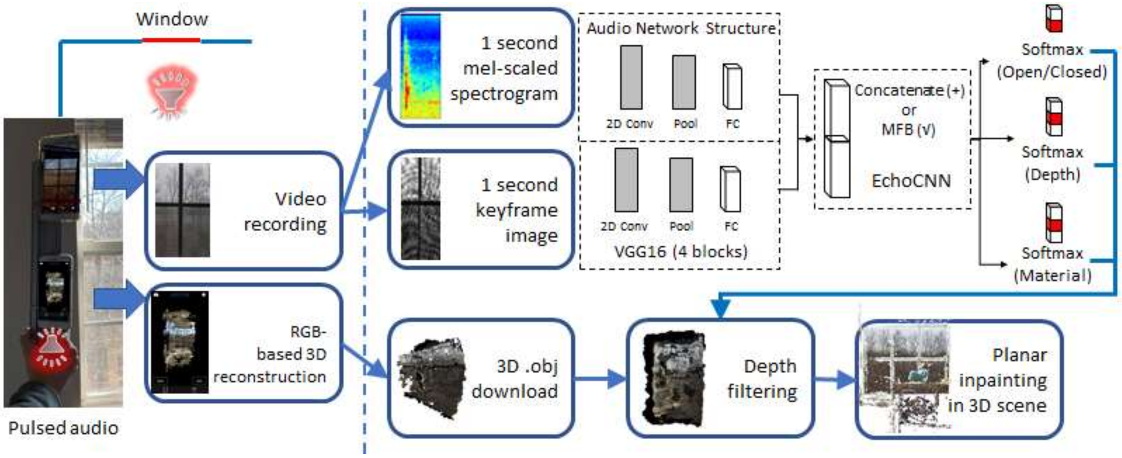

Figure 6.3: *Staged approach* to enhance scene and object reconstruction using audio-visual data. Our echoreconstruction prototype consists of two smartphones - one recording (top) and one emitting/reconstructing (bottom). As the bottom smartphone moves to reconstruct the scene and emits 100 ms pulsed audio (Section 6.3.3), the top smartphone is used to record video of the direct and reflecting sound. The receiving audio is split into 1.0 second intervals to allow for reverberation. These audio intervals are converted into mel-scaled spectrograms and passed through a multimodal echoreconstruction convolutional neural network (we refer to as EchoCNN) comprised of 2D convolutional, max pooling, fully connected, and softmax layers. EchoCNN classifications inform depth filtering and hole filling steps to resolve planar discontinuities in scans caused by reflective surfaces, such as windows and mirrors. Binary classification is used to predict if a window is open or closed. Multi-class classification is used for depth and material estimation.

cesses 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 that can detect discontinuity, depth, and materials.

Audio-based classification: using principles from sound synthesis, propagation, and room acoustics, audio classifiers have been developed for environmental sound (64; 170; 198), material (9), and object shape (268) classification. Audio input can take the form of raw audio, spectral shape descriptors (145; 45; 215), or frequency spectral coefficients that we also adopt in our method.

Audio-visual learning: similar to its applications in natural language processing (NLP) and visual questing & answering systems (103; 102; 74), multi-modal learning using both audio-visual sensory inputs has also been used for classification tasks (220; 246), audio-visual zooming (152), and sound source separation (59; 120) which have also isolated waves for specific generation tasks. Although similar in spirit, our audio-visual method, "Echoreconstruction" differs from the existing methods by learning absorption and reflectance properties to detect a reflective surface, its depth, and material.

6.3 Technical Approach

In this work, we adopt "echolocation" as an analog for our echoreconstruction method. According to (56), echo is defined as *distinct* reflections of the original sound with a sufficient sound level to be clearly heard above the general reverberation. Although perceptible echo is abated because of precedence (known as the Haas effect) (131), 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 (Fig. 6.9), recognize material based on frequency and amplitude, and handle both static and dynamic scenes with moving objects based on Doppler shift. All of which enhance scene and object reconstruction by detecting planar discontinuities from open or closed surfaces and then estimating depth and material.

6.3.1 Echolocation

Echolocation is the use of reflected sound to locate and identify objects, particularly used by animals like dolphins and bats. According to (223), bats emit ultrasound pulses, ranging between 20-150 kHz, to catch an insect prey with a resolution of 2-15 mm. This involves signal processing such as:

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

$$\Delta f = f_D - f_0 = f_0 \frac{c_s}{c_0} cos(\theta) \tag{6.1}$$

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 shift (or 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/observer with velocity c_o and angle θ.

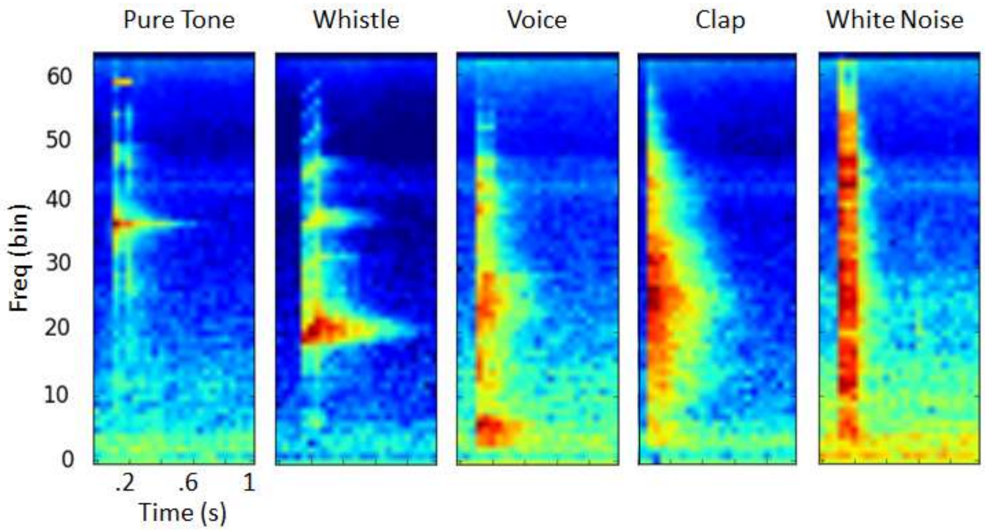

Figure 6.4: Mel-scaled spectrograms of recorded impulses of different sound sources used. *From left to right*: narrow to disperse spectra. Not shown are other pure tone frequencies, chirp, pink noise, and brownian noise. Horizontal axis is time and vertical axis is frequency.

6.3.2 Staged classification and reconstruction pipeline

As depicted in Fig. 6.3, we take a staged approach to enhance scene and object reconstruction using audio-visual data. Our echoreconstruction prototype consists of two smartphones - one recording (top) and one emitting/reconstructing (bottom). 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 (Section 6.3.3). 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 6.3.4). 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. Using mobile accelerometer data and coarse audio with fine visual data to augment depth estimation will be explored as future work.

6.3.3 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 (223). Therefore, using pulsed waves independent for each emission is theoretically better than continuous waves compared to f_0. Furthermore, Section 7.5 shows superior PW results over CW for the 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. Human voice ranges from about 63 Hz to 1 kHz (131) (125 Hz to 8 kHz (56)) and an untrained whistler between 500 Hz to 5 kHz (159). 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 (131). All sound sources were recorded and played back with max volume (Fig. 6.4). While recorded sounds were used for consistency, we plan to add live audio for augmentation and future ease of use during reconstruction. Please see our supplementary materials for spectrograms across all sound sources.

Audio input: audio was generated in pulsed waves (PW). One smartphone to emit the sound while performing a RGB-based reconstruction and the second smartphone to capture video. As future work, a single mobile device or Microsoft Kinect paired with audio chirps could be used for audio-visual capture and reconstruction instead of two separate devices. Each pulsed wave emitted into the scene was a total of 1 second consisting of an 100 ms impulse followed by silence. 1 second audio frames is based on the Sabine Formula of reverberation time for a compact room of like dimensions calculated as:

$$T = 0.05\frac{V}{a} = 0.05\frac{V}{\sum S\alpha} = (0.05\frac{\text{sec}}{\text{ft}})\frac{1,296\ \text{ft}^3}{69.23\ \text{ft}^2} = 0.94\ \text{sec} \tag{6.2}$$

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 in Table 6.2.

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

Total room absorption a using $a = \sum S\alpha$ at 250 Hz

Real bathroom scene	S	α	a (sabins)
Painted walls	432 x	0.10 =	43.20
Tile floor	175 x	0.01 =	1.75
Glass	60 x	0.25 =	15.00
Ceramic	39 x	0.02 =	0.78
Mirror	34 x	0.25 =	8.50
		Total a =	69.23 sabins

Table 6.2: According to the Sabine Formula (Eq. 6.3.3), reverberation time can be calculated as room volume V divided by total room absorption a. For an indoor sound source in a reverberant field, a is the total room absorption at a given frequency (sabins), S is the surface area (ft^2), and α is the sound absorption coefficient at a given frequency (decimal percent). At 250 Hz, the total room absorption a for our real-world bathroom scene is 69.23 sabins.

6.3.4 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, referred to as EchoCNN-A and EchoCNN-AV (Fig. 6.3).

TO BE UPDATE, uncomment out this table

Audio sub-network: our frame-based EchoCNN-A consists of a single convolutional layer 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 STFT coefficients χ (Eq. 7.3.2). Each audio example is classified independently and 1 second intervals to reflect an estimated reverberation time based on a compact room size (Eq. 6.3.3). With a 2048 sample Hann window (N), 25% overlap, and hop length ($H = 2048/4$), this results in a frequency dimension of 21.5 Hz (Eqn. 6.3.4) and temporal dimension of 12 ms (Eqn. 6.3.4) or 12% of each 100 ms pulsed audio. Each spectrogram is individually normalized and downsampled to a size of 62 frequency bins by 25 time bins.

We define the frequency spectral coefficients (148) as:

$$\chi(m,k) = \sum_{n=0}^{N-1} x(n + mH)w(n)exp(-2\pi ikn/N) \tag{6.3}$$

for m^{th} time frame and k^{th} Fourier coefficient with real-valued DT signal $x : \mathbb{Z} \rightarrow \mathbb{R}$, sampled window function $w(n)$ for n $\in [0 : N-1] \rightarrow \mathbb{R}$ of length $N \in \mathbb{N}$, and hop size $H \in \mathbb{N}$ (148). $\mathbb{R}$ denotes continuous time and $\mathbb{Z}$ denotes discrete time. Equal to $|\chi(m,k)|^2$, spectrograms have been demonstrated to perform well as inputs into convolutional neural networks (CNNs) (92). Their horizontal axis is time and vertical axis is frequency.

$$F_{coef}(k) = \frac{k\dot{F}_s}{N} = k\frac{44100}{2048} = k * 21.5 \text{ Hz} \tag{6.4}$$

$$T_{coef}(m) = \frac{m\dot{H}}{F_s} = m\frac{2048 * 0.25}{44100} = m * 0.012 \text{ seconds} \tag{6.5}$$

A hop length of $H = N/2$ achieves a reasonable temporal resolution and data volume of generated spectral coefficients (148). 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 4.2) visual information is particularly useful to aid material classification. We use ImageNet (113) as a visual-based baseline to compare to our audio and audio-visual methods. It also serves as an input into our audio-visual merge layer. Future work will explore whether or not another image classification method is better suited as a baseline and to fuse with audio.

Merge layer: we evaluated concatenation and multi-modal factorized bilinear (MFB) pooling (262) 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.

$$z_i = x^T W_i y = x^T U_i V_i^T y = 1^T(U_i^T x \circ V_i^T y) \tag{6.6}$$

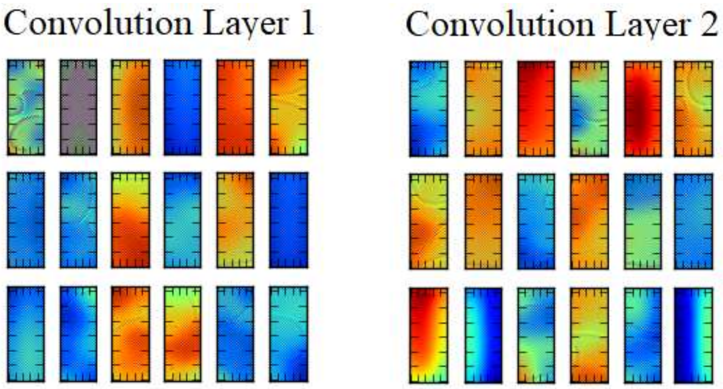

Figure 6.5: Sample visualizations of the filters for the two convolutional layers in the audio-based EchoCNN-A neural network. The model learns filters for octave bands, frequencies, reflections, reverberations, and damping.

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 $1 \in \mathbb{R}^k$ is an all-one vector.

6.3.5 Loss Function

Categorical cross entropy loss is used for EchoCNN inferences. For open/closed predictions, categorical cross entropy loss is used instead of binary if estimating the extent of the surface opening (e.g. all the way open, halfway open, or closed). A regression model is not used for depth estimation because ground truth data is collected in 1 foot increments within the free field for better noise reduction (56). The Softmax function is used for output activations.

6.3.6 Depth filtering and planar inpainting

The outputs of our EchoCNN inform enhancements for 3D reconstruction (Algorithm **??**). 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. Otherwise, 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 threshold of that depth. This gives us the plane size needed to fill. Finally, EchoCNN classifies its surface material.

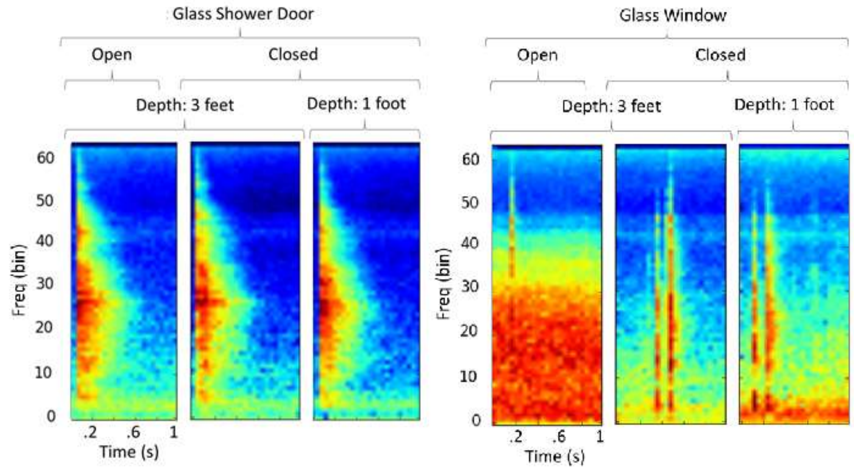

Figure 6.6: Spectrograms from a recorded hand clap in front of an interior glass shower door and exterior glass window. For the interior door, reflected sounds experience intensified damping as we go from opened (*left*) to closed (*middle*) and then from 3 feet to 1 foot depth (*right*). Damping increases with fewer late reverberations and intensity increases with more early reflections. For the exterior window, closing it decreases outside noise up to a distance.

6.4 Datasets

Our audio-based EchoCNN-A and audio-visual EchoCNN-AV convolutional neural networks are trained across nine octave bands with center frequencies 63 Hz, 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, 8 kHz, and 16 kHz. Training is done using these pulsed pure tone impulses along with experimenter hand clap. The hold out test data is comprised of sound sources excluded from training - white noise, experimenter whistle, and voice. The test set contains sound sources not in the training set to evaluate generalization.

6.4.1 Real and synthetic datasets

Real: training data is comprised of 1 second pulsed spectrograms (Fig. 6.6) 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 in 1 ft depth increments based on the distance from the surface. Nine octaves of pure tones, hand claps, and white noise cover a disperse range of frequencies and were thus 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=227). Voice and whistle recordings were chosen for the hold out test

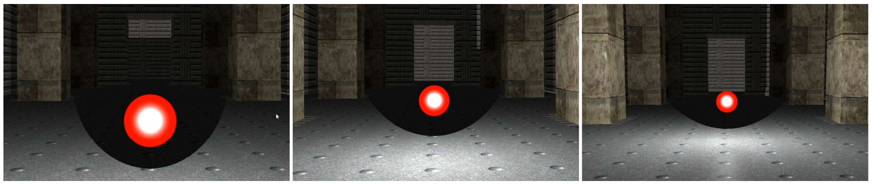

Figure 6.7: Listener at different distances of 1, 2, 3 ft from sound source (red dot) in a virtual environment used to generate synthetic audio-visual data. In addition to open/closed, depth, and material, we make synthetic, unmixed reflection separation data (direct, early, or late) available for future research.

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. While the same hold out sets were used for visual and audio-visual evaluation, unheard is not the same as unseen. Unheard audio can have the same visual appearance between training and test. Other new training and test datasets for visual and audio-visual methods will be future work.

Synthetic: we employ a ray-based geometric sound propagation approach (203). Given scene materials (e.g. carpet, glass, painted, tile, etc.), a sound source (e.g. voice), and listener position, we generate impulse responses for a given scene of varying sizes. From each listener, specular and diffuse rays are randomly generated and traced into the scene. The energy-time curve for simulated impulse response $S_f(t)$ is the sum of these rays:

$$S_f(t) = \sum \delta(t - t_j) I_{j,f} \tag{6.7}$$

where $I_{j,f}$ is the sound intensity for path j and frequency band f, t_j is the propagation delay time for path j, and $\delta(t - t_j)$ is the Dirac delta function or impulse function. As these sound rays collide in the scene, their paths change based on absorption and scattering coefficients of the colliding objects. Common acoustic material properties can be referenced in (56). We assume a sound absorption coefficient, $\alpha = 1.0$ for open windows.

Along with sound intensity $S_f(t)$, a weight matrix W_f is computed corresponding to materials within the scene. Each entry $w_{f,m}$ is the average number of reflections from material m for all paths that arrived at the listener. It is defined as:

$$w_{f,m} = \frac{\sum I_{j,f} d_{j,m}}{\sum I_{j,f}} \tag{6.8}$$

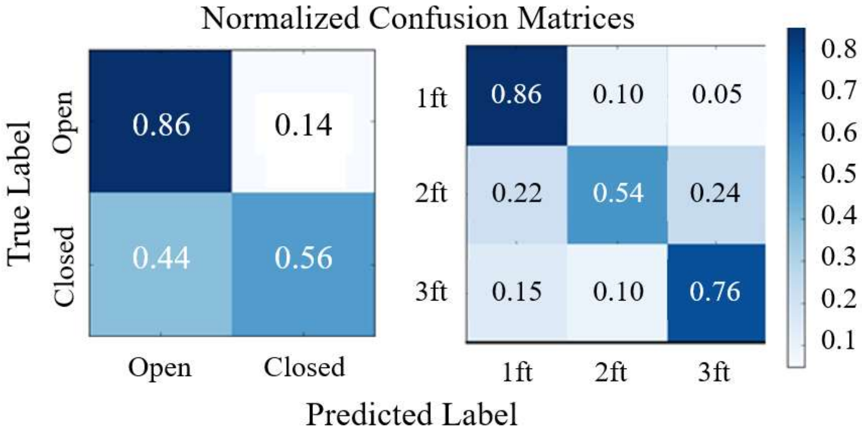

Figure 6.8: EchoCNN-A (*Left*) Confusion matrix to classify open/closed for an interior glass shower door. Open predictions (86%) were more accurate than closed (56%). (*Right*) Confusion matrix to classify depth from same interior glass door. Notice that our EchoCNN is learning to differentiate distance based on reflecting sounds from pulsed ambient waves of a smartphone.

where $d_{j,m}$ is the number of times rays on path j collide with material m, weighted according to the sound intensity $I_{j,f}$ of the path j. To mirror our real-world data, sound source directivity was disabled. Future work is needed to compare ambient and directed sound sources. This data may also be used for material sound separation.

6.5 Experiments and 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 4.2). 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.

ImageNet, 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. Images similar to those found in training are present in test. A proper hold out set based on image (e.g. different depths) should be evaluated as future work.

6.5.1 Experimental setup

Listener (top smartphone, e.g. Galaxy Note 4) and sound source (bottom smartphone, e.g. iPhone 6) are separated vertically by 7 cm. Pulsed sounds are emitted 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 (56). 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 single 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).

6.5.2 Implementation details

We implemented all EchoCNN and baseline models with Tensorflow (1) and Keras (39). 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 (104). 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.

6.5.2.1 Initial 3D Reconstruction

We evaluated the following smartphone-based reconstruction applications to obtain an initial 3D geometry for which our method would enhance. The Astrivis application, based on (225), generates better live 3D geometries for closed object rather than scene reconstructions since it limits feature points per scan. On the other hand, Agisoft Metashape produces scene reconstructions offline from smartphone video. Enabling the software's depth point and guided camera matching features further improved reconstructed geometries.

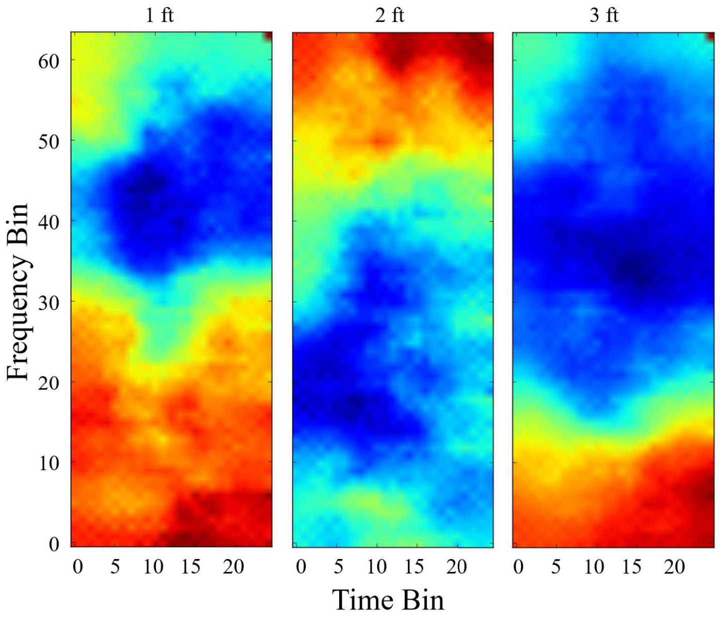

Figure 6.9: *From left to right*: audio input (i.e. mel-scaled spectrogram) which would produce the highest activation for a given depth class from 1 ft, 2 ft, and 3 ft away from an object. Longer reverberation times tend to occur at lower frequencies (3 ft) than at high frequencies (1 and 2 ft) due to typical high frequency damping and absorption.

6.5.3 Results by source frequency and object size

We will evaluate dynamic source frequencies based on the physical size of the objects, since sound wave behavior relates to wavelength. For example, if an object is much smaller than the wavelength, the sound flows around it rather than scattering (131).

$$\lambda = \frac{c}{f} \tag{6.9}$$

where λ is wavelength (ft) of sound in air at a specific frequency, f is frequency (1 Hz), and c is speed of sound in air (ft/s).

6.5.4 Activation Maximization

The objective of activation maximization is to generate an input that maximizes layer activations for a given class. This provides insights into the types of patterns the neural network is learning. Fig. 6.9 shows the different inputs that would maximize EchoCNN activations for depth estimation. Notice lower

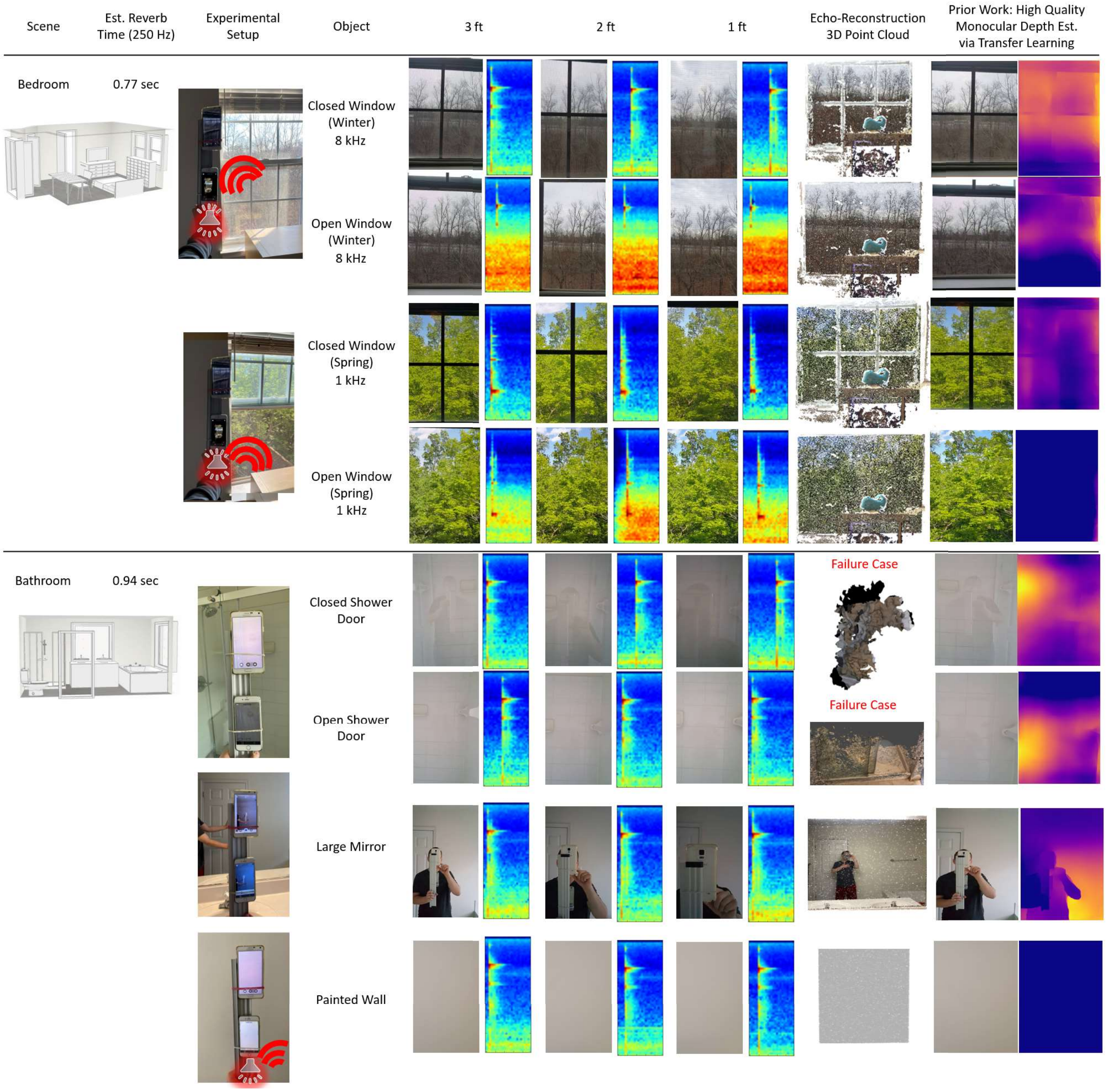
Scene
Est. Reverb Time (250 Hz)
Experimental Setup
Object
3 ft
2 ft
1 ft
Echo-Reconstruction 3D Point Cloud
Prior Work: High Quality Monocular Depth Est. via Transfer Learning
Bedroom
0.77 sec
Closed Window (Winter) 8 kHz
Open Window (Winter) 8 kHz
Closed Window (Spring) 1 kHz
Open Window (Spring) 1 kHz
Bathroom
0.94 sec
Closed Shower Door
Failure Case
Open Shower Door
Failure Case
Large Mirror
Painted Wall

Figure 6.10: We evaluated our method on real and virtual scenes. *Column 1*: we used the off-the-shelf MagicPlan app to obtain 3D models and dimensions to calculate estimated reverberation time based on room size and materials. Our experimental setups consists of a two smartphone prototype. One phone performs an initial reconstruction using state-of-the-art commercial Astrivis application and also emits pulsed audio. The second phone captures video for audio-visual input data into our EchoCNN. We tested glass, mirror, and other objects and surfaces within each scene at different depths, materials, and open/-closed. Using audio, we 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. Background UV textures are placed at a fixed 1 ft (0.3 m) behind estimated surface depth. Audio unable to augment failure cases of the shower from initial RGB-based reconstructions using either (225) or (144). We leave calculating the background depth as future work. We compare our 3D reconstructions to depth estimates based on related work.

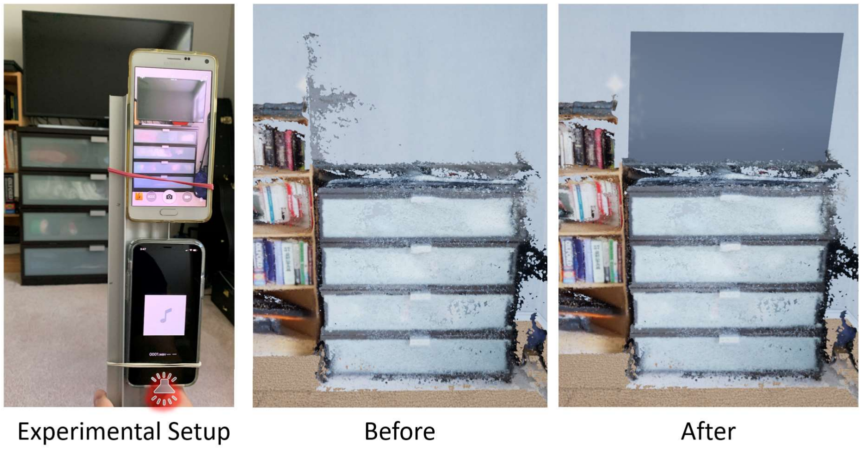

Figure 6.11: Echoreconstruction of a TV on a dresser. (*From left to right*) Photo of prototype system in action, initial 3D reconstruction, depth filtering applied, and resulting echoreconstruction. Semantic rendering is applied post-processing during the render stage of the pipeline.

frequencies tend to occur at 3 ft (longer reverberation times) than at 1 and 2 ft (high frequencies) due to typical high frequency damping and absorption.

6.5.5 Applications

When using a head mounted display (HMD) users are alerted within the virtual environment, when they approach the physical space boundaries established during room setup. However, if room setup does not accurately reflect these boundaries or changes occur after setup, a user risks walking into unseen real-world objects such as glass and walls. Using our method, transmitted sound from the HMD could be used to locate physical objects and appropriately notify the user as an added safety measure. Depth estimation from audio can also be used to unmix and place unseen but heard sound sources from video into a virtual environment. In addition to scene reconstruction, echoreconstruction also reconstructs audio (Fig. 6.12).

6.6 Conclusion and Future Work

To the best of our knowledge, these are the first audio and audio-visual techniques introduced for enhancing scene reconstructions that contain windows and mirrors. Our multi-smartphone prototype and staged echoreconstruction pipeline emits and receives pulsed audio from a variety of sound sources for surface detection, depth estimation, and material classification. These classifications enhance scene and object 3D reconstruction by resolving planar discontinuities caused by open spaces and reflective surfaces

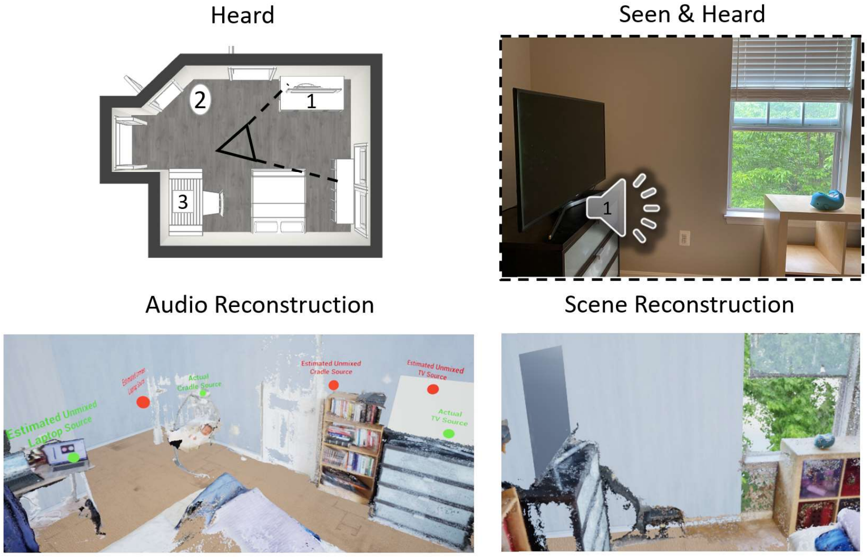
Heard
2
1
3
Seen & Heard
1
Audio Reconstruction
Estimated Unmixed Laptop Source
Estimated Unmixed Laptop Source
Actual Cradle Source
Estimated Unmixed Cradle Source
Estimated Unmixed TV Source
Actual TV Source
Scene Reconstruction

Figure 6.12: EchoCNN may also be used to reconstruct the audio of a scene from video. Instead of depth estimation, our method can be trained to approximate sound source position, which is especially useful for objects that are outside of the camera field of view. Ground truth (green dots) and estimated (red dots) sound source placements. Seen and heard sound source (TV) from the video capture placed more accurately than unseen but heard sound sources (cradle and laptop). Audio-visual compared to audio only. Please see our supplementary video for a VR demo and improved sound source placement as future work.

using depth filtering and planar filling. Our prototype performs well compared to baseline methods given our experiment results for multiple real-world and virtual scenes containing windows, mirrors, and open surfaces. We make publicly available our real and synthetic audio-visual ground truth data in addition to reflection separation data (direct, early, or late reverberations) for future research.

Future Work: To further extend this research, performing audio emission, reception, and 3D reconstruction simultaneously and in real-time instead of having a staged approach would be one possible alternative to explore. This approach could possibly enable mapping classifications to 3D geometry more densely than fusing RGB-D, tracking, or Iterative Closest Point (ICP) (93). An integrated approach may not only be more efficient but also more effective by using audio feedback as part of the reconstruction code. Another possible avenue of exploration is to investigate the impact of live audio for training and/or testing our neural network variations. 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 (141) to extract the fundamental frequency f_0, i.e. the frequency of the lowest partial of the sound, are suggested as future work. Finally, our current implementation holds out voice and whistle data, which is different from the audio used during training. However, unheard sounds does not equate to unseen images. Therefore, some insights can be possibly gained by experimenting with a different training dataset for testing audio-only, visual-only, and audio-visual methods.