Acoustic Non-Line-of-Sight Imaging

David B. Lindell*
Stanford University

Gordon Wetzstein
Stanford University

Vladlen Koltun
Intel Labs

Abstract

Non-line-of-sight (NLOS) imaging enables unprecedented capabilities in a wide range of applications, including robotic and machine vision, remote sensing, autonomous vehicle navigation, and medical imaging. Recent approaches to solving this challenging problem employ optical time-of-flight imaging systems with highly sensitive time-resolved photodetectors and ultra-fast pulsed lasers. However, despite recent successes in NLOS imaging using these systems, widespread implementation and adoption of the technology remains a challenge because of the requirement for specialized, expensive hardware. We introduce acoustic NLOS imaging, which is orders of magnitude less expensive than most optical systems and captures hidden 3D geometry at longer ranges with shorter acquisition times compared to state-of-the-art optical methods. Inspired by hardware setups used in radar and algorithmic approaches to model and invert wave-based image formation models developed in the seismic imaging community, we demonstrate a new approach to seeing around corners.

1. Introduction

Non-line-of-sight (NLOS) imaging techniques aim to recover 3D shape and reflectance information of objects hidden from sight by analyzing multiply scattered light, i.e. light that bounces off of visible parts of a scene, interacts with hidden scene parts, and then reflects back into the line of sight of the detector. With important applications in remote sensing, robotic vision, autonomous vehicle navigation, and medical imaging, NLOS imaging has the potential to unlock unprecedented imaging modalities in a wide range of application scenarios.

Some of the most promising approaches to NLOS imaging use ultra-fast light sources and single-photon-sensitive detectors [9, 10, 15, 17, 21, 23, 27, 33, 34, 36, 38, 39]. Unfortunately, the specialized hardware required for these setups is extremely expensive. Moreover, acquisition times for hidden diffuse objects are very long due to the rapid signal falloff at increasing distances. Alternatively, inexpensive continuous wave (CW) time-of-flight systems have been used for NLOS imaging, but require strong priors and significant compute time to reconstruct the hidden scene [16, 18]. Intensity-only information has been used for tracking NLOS objects [22] or estimating limited scene information [7]; however, high-quality 3D NLOS scene reconstruction remains challenging due to the limited amount of information in the measurements.

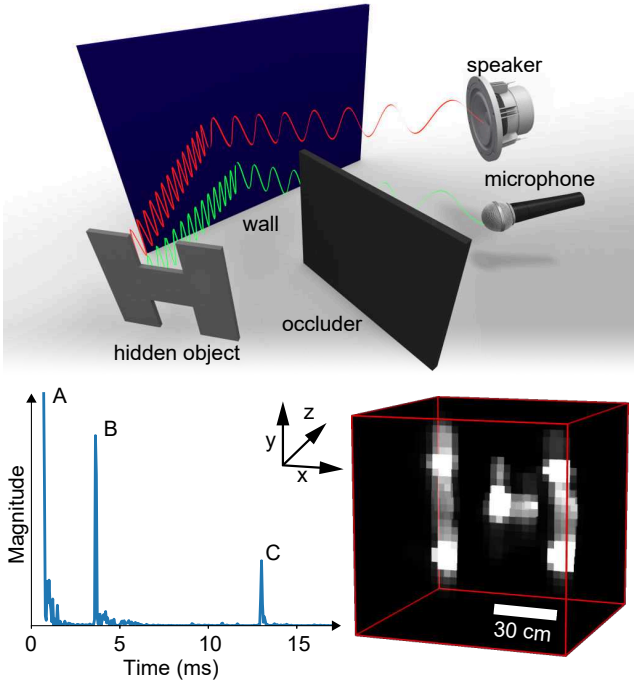

Figure 1. Overview of acoustic NLOS imaging. Modulated sound waves are emitted from a speaker, travel around the corner to a hidden object, and are then recorded by a microphone as they reflect back. The processed measurements (bottom left) contain peaks indicating the path lengths of sound which travels directly from the speaker to the microphone (A, peak is clipped), to the wall and back (B), and also to the hidden object and back (C). Such measurements are captured for a range of speaker and microphone positions to reconstruct the 3D geometry of the hidden object (bottom right).

We demonstrate acoustic non-line-of-sight imaging, which uses readily available, low-cost microphones and speakers to image and resolve 3D shapes hidden around corners. The key motivation of our work is that the reflection

*Work performed during an internship at Intel Labs.

properties of walls are usually specular, i.e. mirror-like, for acoustic waves, so they should reveal hidden scene details more easily than setups relying on visible or near-infrared light. However, there are no focusing optics for acoustics, so we cannot directly measure an "image" of the hidden scene. Moreover, we need to take wave effects into account when modeling sound propagation. The algorithmic framework we develop is inspired by seismic imaging [40], where shock waves are created by explosive charges on the surface to probe hidden underground structures, and returning wavefronts are analyzed to estimate the shape of these structures. While image formation models and inverse methods in seismic imaging share certain properties with acoustic NLOS imaging, the hardware setups and also the applications are very different. Our acoustic imaging setup more closely resembles that of synthetic aperture radar (SAR) [5]; we emit sound chirps from an emitter array and measure the returning wavefront with an array of microphones. In contrast to existing SAR techniques, however, we use off-the-shelf audio hardware and tackle the problem of imaging around corners by analyzing multi-bounce sound effects.

2. Related Work

Optical NLOS Most non-line-of-sight imaging techniques discussed in the literature operate in the optical domain. These approaches can be broadly classified as being passive [7, 22], i.e. not requiring structured illumination, or active. Active systems usually either rely on ultra-fast illumination and detection [9, 10, 15, 17, 21, 23, 27, 33, 34, 36, 38, 39] or on coherence properties of light [6, 19, 20]. To date, passive NLOS systems have only demonstrated scene reconstructions with limited quality, systems that use coherent light are typically limited to microscopic scenes, and time-resolved systems require expensive equipment, such as streak cameras or single-photon avalanche diodes and ultra-fast lasers. Our approach extends time-of-flight techniques to the acoustic domain, leveraging acoustic scattering properties to more efficiently image diffuse objects with comparatively lower acquisition times, longer ranges, and with less expensive equipment than optical techniques.

Radar NLOS The capability of imaging or tracking objects through walls has also been demonstrated using other parts of the electromagnetic spectrum, such as wifi and radar [1, 2, 3, 42]. These approaches are successful for through-wall imaging because the wavelengths they operate at physically propagate through walls without much scattering. Thus, this inverse problem is significantly easier than that of optical approaches, which more closely resembles diffuse optical tomography. Seeing *around* corners with wifi or radar is substantially more difficult than seeing *through* walls because the energy would have to be scattered off of the wall and not through it. Further challenges

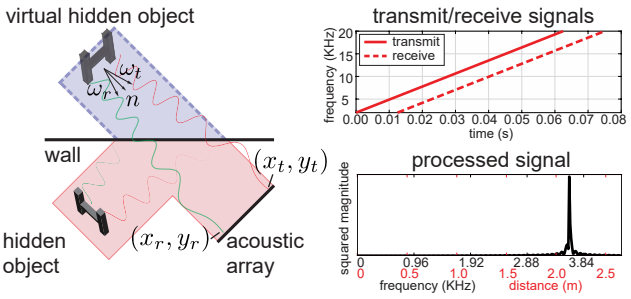

Figure 2. Illustration of the scene geometry and measurement capture. The acoustic array emits an acoustic signal which reflects specularly off of the wall, to the hidden object, and back. Due to the mirror-like scattering of the wall at acoustic wavelengths, the measurements appear to be captured from a mirrored volume located behind the wall, as if the wall were transparent. The transmit signal is a linear ramp in frequency over time. For a single reflector, the return signal is a delayed version of the transmit signal (top right). The receive and transmit signals are mixed together and Fourier transformed, producing a sharp peak at a frequency proportional to the distance of the reflector (bottom right).

of some of these methods include strict government regulations on through-wall imaging systems [35], which make it difficult to release data and fully disclose algorithmic methods. Our approach focuses on seeing around corners with readily available, low-cost acoustic systems.

Acoustic Imaging Imaging simple shapes with sound has been proposed in the past [12, 13]. Moreover, visual-acoustic imaging techniques have been successful for generating sound from video [11, 29, 44], for localizing sound sources or speech signals from video [14, 28, 31, 41], or for imaging with microphone arrays [25]. Acoustic imaging techniques are also common in seismic applications [4, 26, 40], for through-tissue imaging with ultrasound [37], and for line-of-sight imaging, for example with sonar [24]. To the best of our knowledge, this is the first approach to non-line-of-sight 3D scene reconstruction with acoustics.

3. Acoustic NLOS Imaging

3.1. Observation Model

We parameterize the acoustic wavefield such that the transmitting speakers and receiving microphones are located on the plane $\{(x, y, z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} \,|\, z = 0\}$. The wavefield is a 5D function given by $\tau(x_t, y_t, x_r, y_r, t)$, where x_t, y_t indicate the spatial positions of the speakers, x_r, y_r indicate the microphone positions, and t indicates time (see Figs. 1, 2).

We model the measurements as a function of the spatially-varying albedo, $\rho(x, y, z)$, and an acoustic bidirectional reflectance distribution function (BRDF), $f(\boldsymbol{\omega_t}, \boldsymbol{\omega_r})$ [32], which depends on the normalized vector $\boldsymbol{\omega_t}$ pointing from a point (x, y, z) to the transmitting speaker and

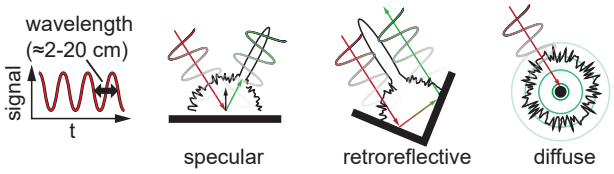

Figure 3. Illustration of acoustic scattering BRDFs. Surfaces that are flat on a scale larger than the wavelength exhibit specular scattering (center left). Corner geometries on the scale of the wavelength exhibit retroreflective scattering (center right). For surfaces smaller than the wavelength, diffraction around the object causes a diffuse scattering event (right).

the normalized direction $\boldsymbol{\omega_r}$ pointing to the receive location with $\boldsymbol{\omega_t}, \boldsymbol{\omega_r} \in \mathbb{R}^3$. The measurements then capture the response of a volume to an acoustic signal where the volume occupies the half-space $\Omega = \{(x, y, z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} \,|\, z > 0\}$. The acoustic signal is transmitted from $(x_t, y_t, z = 0)$, and the response recorded at $(x_r, y_r, z = 0)$:

$$\tau(x_t, y_t, x_r, y_r, t) = \iiint_{\Omega} \frac{1}{(r_t + r_r)^2} \rho(x, y, z) \\ f\left(\boldsymbol{\omega_t}, \boldsymbol{\omega_r}\right) g\left(t - (r_t + r_r)/c\right) dx\, dy\, dz. \tag{1}$$

Here, g is the acoustic signal (described below), c is the speed of sound (≈340 m/s in air), and the distance variables r_t and r_r are given by

$$r_{t/r} = \sqrt{(x_{t/r} - x)^2 + (y_{t/r} - y)^2 + z^2}. \tag{2}$$

Like other NLOS image formation models, we assume the volume to be free from self-occlusions and do not explicitly model visibility terms [27].

The measurement geometry is further illustrated in Figs. 1 and 2. At acoustic wavelengths, the wall acts as a mirror-like reflector, scattering the transmit signal g specularly around the corner, to the hidden object, and back to the acoustic array. Due to the specular scattering of the wall, in the measurements the hidden object appears to be located at a position beyond the wall. For this cause we ignore the wall, such that the image formation models the capture of measurements from a virtual object located behind a transparent wall.

For smooth hidden objects which also exhibit specular scattering, we assume that the surface normals of the virtual object are oriented towards the acoustic array so that the signal can be observed. This assumption is also made *e.g.* by radar systems which image through walls and capture specular scattering [1, 3, 42].

Acoustic Scattering The magnitude of the reflected acoustic wave, or the observed acoustic albedo, ρ, depends on the difference in material density and speed of sound at the interface between air and the scattering object. As the density of the object material increases, more of the sound is reflected rather than transmitted through the object.

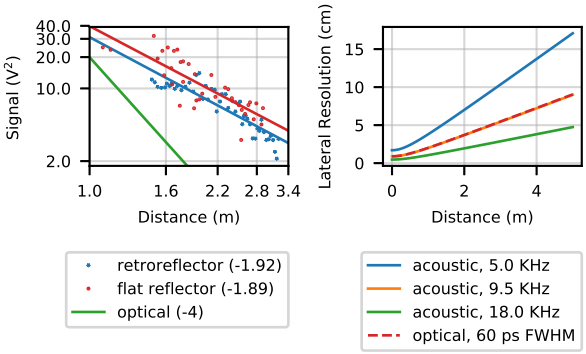

Figure 4. Signal falloff (left) and resolution analysis (right). Measurements captured for a corner reflector and a flat, specularly scattering target are plotted along with a linear regression on a log-log scale. The signal decay is approximately $d^{-1.92}$ for the corner reflector and $d^{-1.89}$ for the flat target which roughly matches the expected d^{-2} falloff. The d^{-4} falloff for optical NLOS imaging with a diffuse reflector is also shown. The lateral resolution over distance is shown for a range of acoustic signal bandwidths compared to a typical optical setup.

The scattering response also depends on the acoustic BRDF, f. For different size objects with different surface geometries, the observed BRDF varies as illustrated in Fig. 3. Specular scattering dominates from surfaces which are on the order of the wavelength in size and flat relative to the wavelength. In our implementation, the wavelength varies from roughly 2 to 20 cm (i.e., $2 - 20$ kHz). For this specularly scattering case, f can be modeled as a delta function as given by Snell's law:

$$f_{\text{specular}}\left(\boldsymbol{\omega_t}, \boldsymbol{\omega_r}\right) = \delta\left(\boldsymbol{\omega_r} - \left(2\langle \boldsymbol{n}, \boldsymbol{\omega_t} \rangle \boldsymbol{n} - \boldsymbol{\omega_t}\right)\right), \tag{3}$$

where $\boldsymbol{n}$ is the surface normal. For objects with sharp angular geometries and corners, which are larger than the wavelength, a retroreflective effect can be observed. That is, the sound is directed back in the direction from which it originated. We can again model the BRDF as a delta function with a non-zero value where this criterion is satisfied:

$$f_{\text{retroreflective}}\left(\boldsymbol{\omega_t}, \boldsymbol{\omega_r}\right) = \delta\left(\boldsymbol{\omega_t} - \boldsymbol{\omega_r}\right). \tag{4}$$

For objects which are smaller than the wavelength, or at edges, the acoustic wave diffracts around the object. This diffractive scattering event can send energy in nearly all directions, which can be modeled as diffuse reflection. In this case the BRDF is Lambertian. Note that such diffuse scattering events create a much weaker signal than strong specular reflections or reflections from corners. Therefore, we rely primarily on specular and corner reflections to reconstruct the hidden object in this work.

Distance Falloff The magnitude of the measured reflection relative to the emitted signal also depends on the distance to the scatterer. The emitted signal propagates along

a spherical wavefront from the emitter to the scatterer and back. As specular and corner reflections redirect the wavefront of sound rather than causing an additional diffuse scattering event, the energy that finally arrives back to the acoustic array is proportional to the total area of the spherical wavefront over a distance of $r_t + r_r$. The signal falloff is therefore proportional to $1/(r_t+r_r)^2$ compared to $1/(r_t^2 r_r^2)$ for diffuse reflections which are common in optical NLOS imaging. We experimentally verify this falloff in Fig. 4.

Transmit Signal While we wish to capture the response of the scene to an acoustic impulse, producing such an impulse at high volume from conventional speakers is impractical. Instead, we transmit a modulated acoustic signal and pre-process the receive signal to emulate the response of the scene to a short pulse.

The modulation and preprocessing is adopted from frequency-modulated continuous wave (FMCW) radar, which provides a good tradeoff between hardware complexity and range resolution compared to other CW or pulsed modulation schemes [5]. The transmit signal, $g(t)$, is a linear sweep from an initial frequency f_0 to a frequency f_1 over a time T (see Fig. 2):

$$g(t) = \sin\left[2\pi\left(f_0 t + \frac{f_1 - f_0}{2T}t^2\right)\right], 0 \le t \le T. \quad (5)$$

The captured measurements $\tau(x_t, y_t, x_r, y_r, t)$ thus contain attenuated and delayed copies of $g(t)$ backscattered by each reflector in the scene. To emulate the response of a scene to a short pulse we mix the received signal, τ, with the original transmit signal, $g(t)$, along the time dimension and then take the squared magnitude of the Fourier transform. For a fixed transmit and receive position, each reflector produces a peak at frequency f_b given by

$$f_b = \frac{r_t + r_r}{Tc}B, \quad (6)$$

where $B = f_1 - f_0$ is the bandwidth of $g(t)$ [5]. We therefore approximate the FMCW pre-processed measurements $\tilde{\tau}$ as the scene response to an impulse, δ, such that

$$\tilde{\tau}(x_t, y_t, x_r, y_r, t) = \iiint_{\Omega} \frac{1}{(r_t + r_r)^2}\rho(x,y,z) \quad (7)$$
$$f\left(\boldsymbol{\omega_t}, \boldsymbol{\omega_r}\right)\, \delta\left((r_t + r_r) - tc\right) dx\, dy\, dz,$$

where t is the time dimension after scaling the frequency axis of FMCW pre-processing by T/B.

3.2. Reconstruction from Confocal Measurements

In the case where the transmit and receive locations are at the same spatial position ($x_t = x_r$ and $y_t = y_r$), we can develop closed-form solutions for the reconstruction procedure. This is referred to as a "confocal" scanning arrangement in optical non-line-of-sight imaging where the laser and sensor illuminate and image the same position on the wall [27]. We proceed with this assumption of confocal acoustic measurements for closely spaced speakers and microphones and then show how to incorporate non-confocal measurements into this framework.

When the hidden object is specular, we assume that the hidden object has surface normals that direct sound back to the acoustic array, or that $n(x,y,z) \approx \frac{\boldsymbol{\omega_t} + \boldsymbol{\omega_r}}{2}$, which also implies that $\boldsymbol{\omega_t} \approx \boldsymbol{\omega_r}$ and $f(\boldsymbol{\omega_t}, \boldsymbol{\omega_r}) = \delta\left(\sum_i \boldsymbol{\omega_t^i} - \boldsymbol{\omega_r^i}\right)$, where i indexes an element of the vector. Then the confocal measurements are given by

$$\tilde{\tau}_c(x_t, y_t, t) = \frac{1}{(tc)^2}\iiint_{\Omega}\rho(x,y,z)\,\delta\left(r_t - \frac{tc}{2}\right) \cdot\delta\left(\sum_i \boldsymbol{\omega_t^i} - \boldsymbol{\omega_r^i}\right) dx\, dy\, dz, \quad (8)$$

where we use the relationship $r_t = \frac{tc}{2}$ to pull the attenuation factor $\frac{1}{(2r_t)^2}$ from Eq. 1 out of the integral. Note that in the confocal case, we only have $\boldsymbol{\omega_t} = \boldsymbol{\omega_r}$ when $x_t = x_r \approx x$, $y_t = y_r \approx y$, and $\frac{tc}{2} \approx z$. Therefore the captured measurements approximate the reconstructed volume, or $\tilde{\tau}_c(x_t, x_r, \frac{2t}{c}) \approx \rho(x,y,z)$. In other words, if the hidden object contains surface normals that return sound to the acoustic array, we can directly capture the specular image of the object. Of course, in the confocal case, the acoustic reflection is only captured if the normal is oriented directly towards the transmit and receive location. Being able to capture and efficiently reconstruct the hidden volume from specular reflections which return to other (non-confocal) receive locations can provide a large increase in the captured signal and the reconstruction quality.

For diffuse or retroreflective hidden objects, the captured measurement can be approximated as

$$\tilde{\tau}_c(x_t, x_r, t) = \frac{1}{(tc)^2}\iiint_{\Omega}\rho(x,y,z)\,\delta\left(r_t - \frac{tc}{2}\right) dx\, dy\, dz, \quad (9)$$

which ignores the Lambertian terms and assumes isotropic scattering. Given this image formation model, the Light-Cone Transform (LCT) can be used as a closed-form solution for $\rho(x,y,z)$. The transform consists of a re-interpolation of the measurements along the t dimension and deconvolution with a pre-calibrated kernel [27].

3.3. Non-Confocal Reconstructions

Confocal measurements enable efficient reconstruction of the 3D geometry of the hidden volume using the methods outlined in the previous section; however, we also wish to derive efficient reconstruction routines for the more general

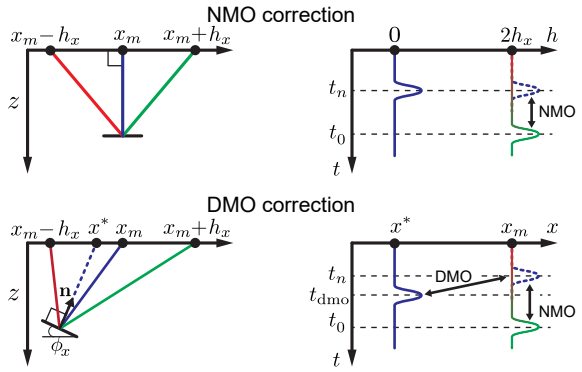

Figure 5. Illustration of normal moveout (NMO) correction and dip moveout (DMO) correction. Given a known offset between the transmit and receive positions with respect to a midpoint position x_m and a scatter oriented with surface normal perpendicular to the measurement plane, NMO correction adjusts the offset measurements to emulate a confocal measurement taken at x_m. If the scatterer is not oriented perpendicular to the measurement plane, an additional DMO correction shifts the measurements in time and in space to confocal position x^*.

case of non-confocal measurements captured by the acoustic array which may contain additional specular reflections returning outside the confocal receiver positions.

We achieve efficient processing of the non-confocal measurements by computationally adjusting them such that they emulate measurements captured over a confocal sampling grid. This computational adjustment consists of three steps. First, we reparameterize the measurements by their midpoint and offset locations (x_m, y_m) and (h_x, h_y) instead of transmit and receive positions (x_t, y_t) and (x_r, y_r). The midpoint and offset parameters are given by $x_m = (x_t + x_r)/2$ and $h_x = |x_r - x_t|/2$. Second, we resample along the time dimension to remove the additional roundtrip propagation time of measurements with non-zero offset $(h_x, h_y > 0)$ relative to confocal measurements with zero-offset $(h_x, h_y = 0)$. Third, we apply an additional corrective factor that adjusts the midpoint and time of captured measurements to account for surface orientation. The second and third processing steps are common to seismic imaging and are known as normal moveout (NMO) correction and dip moveout (DMO) correction [40]. A pseudocode description of this adjustment procedure is included in the supplementary material.

The emulated confocal measurements are then obtained by integrating along the offset dimension of the NMO and DMO corrected measurements $\tilde{\tau}^*$ given as

$$\tilde{\tau}_c^*(x_m, y_m, t_n) = \iint_{\Omega_{h_x h_y}} \tilde{\tau}^*(x_m, y_m, h_x, h_y, t_n)\, dh_x\, dh_y. \tag{10}$$

Here $\Omega_{h_x h_y}$ is the region of support of the offsets, τ_c^* represents the emulated confocal measurements, and t_n is the normal-moveout-corrected time dimension. For clarity we describe NMO and DMO correction in detail for two dimensions (used in our linear acoustic array) in the following sections and include the three-dimensional equations in the supplementary information.

Normal Moveout Correction Normal moveout correction takes measurements whose transmit and receive locations have a midpoint location x_m and offset h_x, and applies an offset-dependent shift in time so that the resulting measurements approximate confocal, or zero-offset, measurements taken at the midpoint x_m. Given the measurement geometry of Fig. 5, the time difference between measurements taken with offset h_x and zero-offset is

$$t_n = \sqrt{t^2 - \frac{4h_x^2}{c^2}}. \tag{11}$$

This formulation assumes that the measurements are captured from scatterers with location $x = x_m$, or that the normal points in a direction perpendicular to the acoustic array (see Fig. 5).

Dip Moveout Correction For scatterers whose surface normals do not point perpendicular to the acoustic array, an additional dip moveout (DMO) correction adjusts the time of arrival and midpoint location to align with those of the confocal measurement. Let ϕ_x be the angle orientation of a scatterer and $n_x = \sin\phi_x$ and $n_z = -\cos\phi_x$ be the associated normal vectors (shown in Fig. 5), then the corrected time [40] is

$$t_{\text{dmo}} = \sqrt{t_n^2 + \frac{4h_x^2 \sin^2\phi_x}{c^2}}. \tag{12}$$

Generally, though, the angle ϕ_x is unknown. Furthermore, the correction should adjust not only the measurement time, but also the midpoint location of the measurement.

To apply these corrections without knowledge of ϕ_x, we use a Fourier domain approach from seismology called log-stretch DMO correction. This approach provides a closed-form solution to DMO correction by re-interpolating or stretching the NMO-corrected measurements along the time domain such that $t'_n = \ln t_n$ and applying a phase shift in the Fourier domain. Let T_{nmo} be the Fourier transform of the NMO-corrected measurements along the x_m and t'_n dimensions, and let k_x and W be their Fourier duals. Then the Fourier transform of the DMO corrected measurements, T_{dmo}, is derived by Zhou et al. [43] and given as

$$T_{\text{dmo}}(k_x, W; h_x) = e^{j\Phi}\, T_{\text{nmo}}(k_x, W; h_x), \tag{13}$$

$$\Phi = \begin{cases} 0, & k_x h_x = 0 \\ k_x h_x, & W = 0 \\ \frac{W}{2}\left\{\sqrt{1 + \left(\frac{2k_x h_x}{W}\right)^2} - 1 - \ln\left[\frac{\sqrt{1 + \left(\frac{2k_x h_x}{W}\right)} + 1}{2}\right]\right\}, & W \neq 0 \end{cases} \tag{14}$$

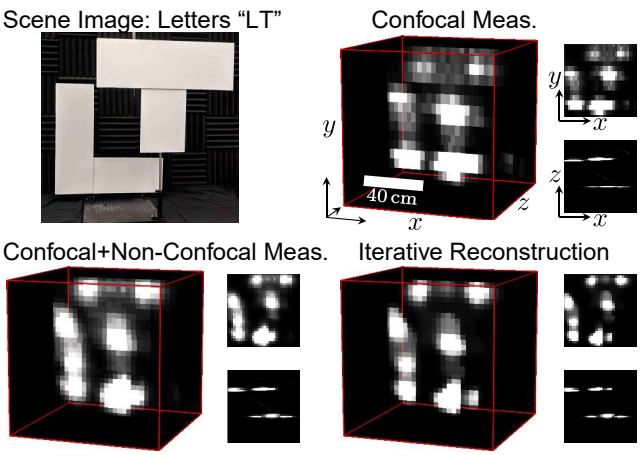

Figure 6. Reconstruction pipeline shown for a scene with two cutout shapes (top left). The subset of confocal measurements (top right) is augmented by processing the set of non-confocal measurements to emulate the confocal geometry (bottom left). An iterative reconstruction procedure applies spatial deconvolution with a measured spatial PSF and sparsity and sparse gradient priors to produce the final reconstruction (bottom right).

The correction is performed for each offset h_x, the result is inverse Fourier transformed and then un-stretched along the time dimension to yield the output of the correction, $\tilde{\tau}^*$. Additional details on DMO correction can be found in the supplementary material.

Reconstruction The reconstruction procedure consists of applying NMO and DMO correction to the captured measurements and using the LCT if the hidden object exhibits diffuse or retroreflective scattering. We use the LCT selectively, as applying it to specular measurements reduces reconstruction quality (see supplementary material). To further mitigate spatial blur and improve reconstruction results, we apply an iterative reconstruction based on the alternating direction method of multipliers [8]. The spatial blur is measured by fitting a Gaussian to the initial reconstruction of a small (5 cm) corner reflector at a distance of approximately 1 m from the acoustic array. Along with the deconvolution, we apply sparsity and total variation priors on the reconstructed volume. Intermediate results of the reconstruction (visualized using the Chimera volume renderer [30]) are shown in Fig. 6, which demonstrates how incorporating the non-confocal measurements improves signal quality and increases spatial sampling by reconstructing on the midpoint grid. Further details on the iterative reconstruction can be found in the supplementary material.

4. Implementation

Our experimental setup is shown in Fig. 7. A linear array of 16 pairs of collocated speakers and microphones is mounted vertically on a horizontally scanning translation stage. The speakers and microphones are evenly spaced along 1 m in the vertical dimension, and the translation stage provides 1 m of travel distance. An acoustic foam barrier is placed between the hidden object and the array, leaving an indirect path for sound to scatter off of a visible wall, to the hidden object, and back to the array.

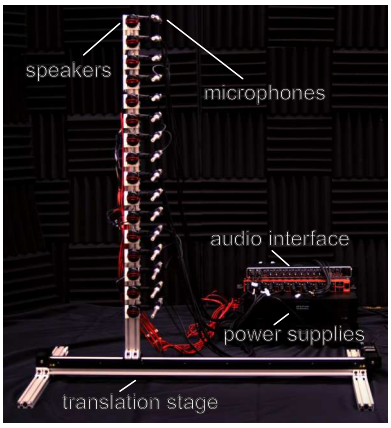

Figure 7. Photograph of the prototype system. The prototype comprises a linear array of 16 speakers and microphones mounted vertically on a 1 m translation stage. Power amplifiers and a set of audio interfaces drive the speakers and record from the microphones.

Hardware The hardware comprises a set of off-the-shelf omnidirectional measurement microphones (Dayton Audio EMM-6), 1-inch automobile speakers (DS18 TWC), and two 8-channel acoustic interfaces (Behringer ADA8200, UMC1820) that are synchronized by fiber optic cable to provide 16 channels of input and output at a sampling rate of 48 kHz. We use two sets of 8-channel amplifiers (Emotiva A-800) to drive the speakers with our transmit signal. The translation stage (Zaber X-BLQ-E) is scanned to take measurements at 32 positions along a 1 m interval.

The transmit signal is a linear frequency chirp from 2 to 20 kHz with a duration of 0.0625 s. Here, the chirp bandwidth is limited by the frequency response and sampling constraints of the hardware. We measure the chirp volume to be approximately 80 dB_{SPL} at 1 m. Given the constraint that we have only a single chirp in flight at a given time, the maximum range for the system is less than $0.0625\text{s} \times \frac{340}{2}\frac{\text{m}}{\text{s}} = 10.6$ m. The total chirp time at each scan position is therefore 16×0.0625 s = 1 s and the total scan time, including mechanical scanning, is approximately 4.5 min.

Software All procedures are implemented in Python. An initial reconstruction including NMO and DMO correction with the LCT over a gated 2 m range ($32 \times 30 \times 250$ resolution) requires 4 s on an Intel 2.50 GHz Core i7-4870-HQ. The iterative reconstruction requires 0.1 s per iteration without the LCT operator, and 9 s per iteration with the LCT operator and typically converges in several hundred iterations. All datasets and software are available online[1].

[1]https://github.com/computational-imaging/AcousticNLOS

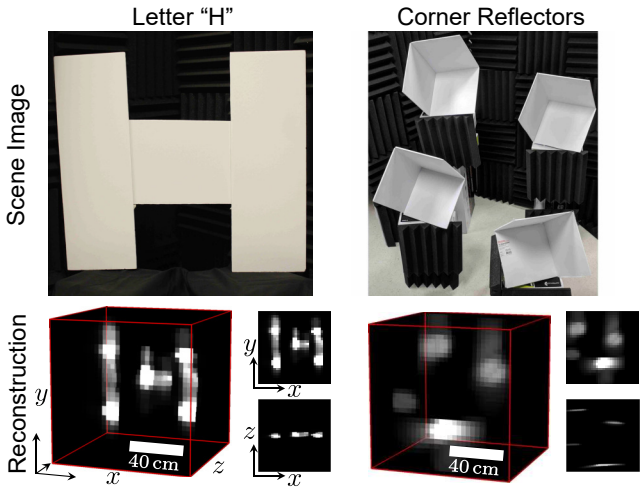

Figure 8. Results captured with the hardware prototype. Photographs of each scene are shown in the top row, and maximum projection visualizations are shown in the bottom row.

Calibration We calibrate the microphone gain on the acoustic interfaces to be approximately equal across channels by facing the acoustic array towards a flat target and tuning the analog controls to equalize the received signal. The microphone frequency response is also calibrated to be approximately flat from 2 to 20 kHz using frequency-dependent scaling factors provided for each microphone from a factory calibration procedure. The experiment room is isolated using acoustic foam paneling and we also subtract a measurement taken without the hidden object to further mitigate any signal from irrelevant room geometries. Before the reconstruction, we scale the measurements by $(tc)^2$ to compensate for squared distance falloff. The reconstructed volumes enclosing the hidden object are visualized by digitally gating out the measurements from the direct path between the speaker and microphone and also diffuse reflections from the surface of the wall.

5. Results

Experimental Results We capture experimental results with the prototype hardware system as shown in Figs. 6 and 8. The results include the following: *Letter “H”*, *Corner Reflectors*, and *Letters “LT”*.

Letter “H”: This scene consists of a letter cut out from posterboard which measures 76 cm by 86 cm. We place the letter around the corner at a distance of 2.2 m from the acoustic array along the indirect path of propagation and angle it towards the direction of sound incident from the wall. The reconstructed result captures the clear shape of the letter as shown in Fig. 8. The dark gaps in the reconstruction correspond to seams where posterboard panels are joined together. At these locations, the acoustic waves appear to be refracted around the letter or diffracted rather than strongly reflected as at other locations.

Corner Reflectors: The four corner reflectors are placed at different distances and heights in the scene to demonstrate how the reconstructions resolve the relative position of each reflector. The reflectors have a side length of 25 cm and are centered at a distance of approximately 2.8 m from the acoustic array along the indirect path. We place acoustic foam in front of the stands which hold the reflectors to lessen their contribution to the measurements. Since these objects are retroreflectively scattering, we use the LCT in the initial and iterative reconstructions and show their relative 3D position in the reconstructed result of Fig. 8.

Letters “LT”: Two letters cut out of posterboard are placed approximately 2.6 m from the acoustic array along the indirect path. The “L” cutout is placed roughly 40 cm behind the “T” cutout, and the letters are approximately 25 cm in width. The reconstruction recovers the shape of both letters as shown in Fig. 6.

Signal Falloff We measure the signal falloff for acoustic NLOS by placing a corner reflector and flat wall (made of posterboard) at increasing distances around the corner from the acoustic array. A single speaker emits the FMCW waveform and we measure the peak squared voltage of the backscattered signal after FMCW processing. The squared voltage (proportional to receive power) falls off roughly as the square of the distance, as expected for a specularly reflecting wavefront. We measure the falloff using a linear regression fit to the signal and distance values on a log-log scale as shown in Fig. 4. The slope of the line indicates the falloff; we find the slope to be -1.91 for a retroreflector and -1.89 for a wall, where the expected value is -2.

Resolution We also derive resolution bounds on the lateral resolution of our system which incorporates the FMCW modulation scheme (see supplementary material for extended derivation) . For a temporal resolution of γ, the lateral resolution Δx is given as

$$\Delta x = \frac{\gamma c}{(x_t - x)/r_t + (x_r - x)/r_r}, \tag{15}$$

which provides the resolution given the locations of the scatterer, source, and receiver positions. For FMCW modulation, $\gamma = \frac{1}{2B}$ where B is the bandwidth of $g(t)$ (see Eq. 6). The lowest resolution is achieved with a confocal measurement at the position of the acoustic array which maximizes the lateral distance from the scatterer. Fig. 4 shows the theoretical lateral resolution for a scatterer with 0.5 m lateral distance over a range of axial distances from a confocal measurement position. We also plot resolution curves for various bandwidth values and for a confocal optical NLOS setup with a temporal resolution of 60 ps [27]. Due to the relatively slow speed of sound through air compared to the speed of light, a relatively small acoustic bandwidth of 9.5 kHz achieves roughly the same lateral resolution as the optical setup.

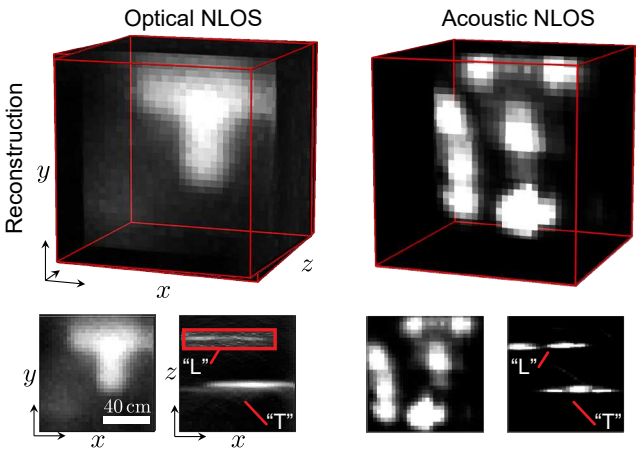

Figure 9. Comparison between optical and acoustic NLOS reconstructions. The acoustic reconstruction (right) recovers the "L" letter while the optical reconstruction (left) fails to capture it due to the more rapid signal falloff over distance (scaled "L" signal shown in red box of x-z max projection). Reconstructed volumes are centered in z around the scene for visualization.

While these results show a theoretical lateral resolution, the achieved resolution also depends on the diffraction limited bandwidth of the transmit signal. For scatterers which are smaller than the wavelength, less signal scatters back, effectively reducing the bandwidth of the received signal. This effect could be partially mitigated in an acoustic system by using shorter wavelengths, *e.g.* by using ultrasonic transducers. We show additional resolution experiments in the supplementary material.

Comparison to Optical NLOS In order to provide a qualitative comparison of the acoustic and optical NLOS reconstruction quality, we also capture the *Letters "LT"* scene in a dark room using the optical setup and LCT reconstruction method of O'Toole et al. [27] and compare the results in Fig. 9. For the optical scan, we place the "T" of the scene 50 cm away from the wall and scan a confocal grid of 32×32 points with an exposure of 6 s per scan point. This exposure time is roughly two orders of magnitude greater than the acoustic chirp duration of 0.0625 s per speaker. While we capture the acoustic result at a distance of roughly 1.6 m from the wall, the signal decay of the optical setup requires a closer distance in order to reconstruct the closest letter, "T". The position of the more distant letter, "L", is only barely visible above the noise floor (see the x-z max projection of the reconstructed volume in Fig. 9). Due to the lower rate of signal falloff for acoustic NLOS, the recovered shape of both letters is distinctly visible.

6. Discussion

In summary, we demonstrate an alternate modality for NLOS imaging using sound. Inspired by inverse methods from seismology and synthetic aperture radar, we develop computational methods for acoustic NLOS imaging and demonstrate the approach using a hardware prototype built with inexpensive, off-the-shelf components. We also evaluate the resolution limits and signal decay of this modality and provide comparisons to optical techniques.

Limitations and Future Work Our current hardware setup simulates a 2D array by scanning a linear array; though other hardware configurations are possible. For example, a single scanned speaker and microphone could be used to capture measurements from a compact device, a 1D array could be used without scanning to capture 2D measurements, perhaps for NLOS object detection, or a full 2D array could capture the 3D volume without scanning, enabling faster acquisition speeds. In this work, we find that the scanned 1D array allows a convenient tradeoff between system complexity, measurement quality, and acquisition speed.

For hidden objects with a weak or non-existent diffuse component, which are not retroreflective, or which have surface normals that reflect sound away from the acoustic array, the reconstruction may fail in the absence of backscattered signal. Moreover, features much smaller than the emitted wavelengths can be difficult to resolve. In such cases optical systems may yield better results, but at shorter distances and with much longer exposure times due to a more rapid falloff in signal intensity with distance. While this shortcoming also applies to other wifi or radar-based systems, acoustic imaging at shorter wavelengths, e.g. with ultrasound, can potentially increase the amount of signal returning by causing smaller surface features to act as diffuse reflectors or retroreflectors.

We currently evaluate optical and acoustic NLOS separately; however, both methods could be combined in a system which leverages their unique benefits. A relevant application could be for autonomous vehicle navigation where optical systems have difficulty imaging reflections from dark regions such as roads, tires, or buildings, but an acoustic signal would be strongly reflected. Many vehicles already deploy small arrays of ultrasonic transducers on their bumpers, and so acoustic NLOS imaging in this scenario could be practicable with existing hardware.

Acknowledgements This project was supported by a Terman Faculty Fellowship, a Sloan Fellowship, by the National Science Foundation (CAREER Award IIS 1553333), the DARPA REVEAL program, the ARO (Grant W911NF-19-1-0120), and by the KAUST Office of Sponsored Research through the Visual Computing Center CCF grant. We also thank Ioannis Gkioulekas for an inspiring discussion.

References

[1] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand. Capturing the human figure through a wall. *ACM Trans. Graph.*, 34(6), 2015. 2, 3

[2] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller. 3D tracking via body radio reflections. In *Proc. NSDI*, 2014. 2

[3] F. Adib and D. Katabi. See through walls with Wi-Fi! In *ACM SIGCOMM*, 2013. 2, 3

[4] F. Admasu and K. Toennies. Automatic method for correlating horizons across faults in 3D seismic data. In *Proc. CVPR*, 2004. 2

[5] C. Baker and S. Piper. Continuous wave radar. In *Principles of Modern Radar: Volume 3: Radar Applications*, pages 17–85. Institution of Engineering and Technology, 2013. 2, 4

[6] M. Batarseh, S. Sukhov, Z. Shen, H. Gemar, R. Rezvani, and A. Dogariu. Passive sensing around the corner using spatial coherence. *Nature Communications*, 9(1):3629, 2018. 2

[7] K. L. Bouman, V. Ye, A. B. Yedidia, F. Durand, G. W. Wornell, A. Torralba, and W. T. Freeman. Turning corners into cameras: Principles and methods. In *Proc. ICCV*, 2017. 1, 2

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 6

[9] M. Buttafava, J. Zeman, A. Tosi, K. Eliceiri, and A. Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Opt. Express*, 23(16):20997–21011, 2015. 1, 2

[10] S. Chan, R. E. Warburton, G. Gariepy, J. Leach, and D. Faccio. Non-line-of-sight tracking of people at long range. *Opt. Express*, 25(9):10109–10117, 2017. 1, 2

[11] A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman. The visual microphone: Passive recovery of sound from video. *ACM Trans. Graph.*, 33(4), 2014. 2

[12] I. Dokmanić, Y. M. Lu, and M. Vetterli. Can one hear the shape of a room: The 2-D polygonal case. In *Proc. ICASSP*, 2011. 2

[13] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013. 2

[14] 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. *ACM Trans. Graph.*, 37(4), 2018. 2

[15] G. Gariepy, F. Tonolini, R. Henderson, J. Leach, and D. Faccio. Detection and tracking of moving objects hidden from view. *Nature Photonics*, 10(1):23–26, 2016. 1, 2

[16] M. Gupta, S. K. Nayar, M. B. Hullin, and J. Martin. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Trans. Graph.*, 34(5), 2015. 1

[17] O. Gupta, T. Willwacher, A. Velten, A. Veeraraghavan, and R. Raskar. Reconstruction of hidden 3D shapes using diffuse reflections. *Opt. Express*, 20(17):19096–19108, 2012. 1, 2

[18] F. Heide, L. Xiao, W. Heidrich, and M. B. Hullin. Diffuse mirrors: 3D reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors. In *Proc. CVPR*, 2014. 1

[19] O. Katz, P. Heidmann, M. Fink, and S. Gigan. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature Photonics*, 8(10):784–790, 2014. 2

[20] O. Katz, E. Small, and Y. Silberberg. Looking around corners and through thin turbid layers in real time with scattered incoherent light. *Nature Photonics*, 6(8):549–553, 2012. 2

[21] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar. Looking around the corner using transient imaging. In *Proc. ICCV*, 2009. 1, 2

[22] J. Klein, C. Peters, J. Martin, M. Laurenzis, and M. B. Hullin. Tracking objects outside the line of sight using 2D intensity images. *Scientific Reports*, 6:32491, 2016. 1, 2

[23] X. Liu, S. Bauer, and A. Velten. Analysis of feature visibility in non-line-of-sight measurements. In *Proc. CVPR*, 2019. 1, 2

[24] X. Lurton. *An Introduction to Underwater Acoustics: Principles and Applications*. Springer Science & Business Media, 2002. 2

[25] A. O'Donovan, R. Duraiswami, and J. Neumann. Microphone arrays as generalized cameras for integrated audio visual processing. In *Proc. CVPR*, 2007. 2

[26] S. M. O'Malley and I. A. Kakadiaris. Towards robust structure-based enhancement and horizon picking in 3-D seismic data. In *Proc. CVPR*, 2004. 2

[27] M. O'Toole, D. Lindell, and G. Wetzstein. Confocal non-line-of-sight imaging based on the light cone transform. *Nature*, 555(7696):338–341, 2018. 1, 2, 3, 4, 7, 8

[28] A. Owens and A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proc. ECCV*, 2018. 2

[29] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proc. CVPR*, 2016. 2

[30] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13):1605–1612, 2004. 6

[31] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon. Learning to localize sound source in visual scenes. In *Proc. CVPR*, 2018. 2

[32] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja. The room acoustic rendering equation. *The Journal of the Acoustical Society of America*, 122(3):1624–1635, 2007. 2

[33] C.-Y. Tsai, K. Kutulakos, S. G. Narasimhan, and A. C. Sankaranarayanan. The geometry of first-returning photons for non-line-of-sight imaging. In *Proc. CVPR*, 2017. 1, 2

[34] C.-Y. Tsai, A. Sankaranarayanan, and I. Gkioulekas. Beyond volumetric albedo—A surface optimization framework for non-line-of-sight imaging. In *Proc. CVPR*, 2019. 1, 2

[35] United States Federal Government. Electronic code of federal regulations title 22, part 121, 2018. Accessed 2018-09-12. 2

[36] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Communications*, 3:745, 2012. 1, 2

[37] A. Webb and G. C. Kagadis. *Introduction to Biomedical Imaging*. John Wiley and Sons Inc., 2003. 2

[38] D. Wu, G. Wetzstein, C. Barsi, T. Willwacher, M. O'Toole, N. Naik, Q. Dai, K. Kutulakos, and R. Raskar. Frequency analysis of transient light transport with applications in bare sensor imaging. In *Proc. ECCV*, 2012. 1, 2

[39] S. Xin, S. Nousias, K. Kutulakos, A. Sankaranarayanan, S. Narasimhan, and I. Gkioulekas. A theory of Fermat paths for non-line-of-sight shape reconstruction. In *Proc. CVPR*, 2019. 1, 2

[40] Ö. Yilmaz. *Seismic Data Analysis: Processing, Inversion, and Interpretation of Seismic Data*. Society of Exploration Geophysicists, 2001. 2, 5

[41] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proc. ECCV*, 2018. 2

[42] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi. Through-wall human pose estimation using radio signals. In *Proc. CVPR*, 2018. 2, 3

[43] B. Zhou, I. M. Mason, and S. A. Greenhalgh. An accurate formulation of log-stretch dip moveout in the frequency-wavenumber domain. *Geophysics*, 61(3):815–820, 1996. 5

[44] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proc. CVPR*, 2017. 2