# REALIMPACT: A Dataset of Impact Sound Fields for Real Objects

Samuel Clarke[1] Ruohan Gao[1] Mason Wang[1] Mark Rau[1]
Julia Xu[1] Jui-Hsien Wang[2] Doug L. James[1] Jiajun Wu[1]
[1]Stanford University [2]Adobe Research

## Abstract

*Objects make unique sounds under different perturbations, environment conditions, and poses relative to the listener. While prior works have modeled impact sounds and sound propagation in simulation, we lack a standard dataset of impact sound fields of real objects for audio-visual learning and calibration of the sim-to-real gap. We present* REALIMPACT*, a large-scale dataset of real object impact sounds recorded under controlled conditions.* REALIMPACT *contains 150,000 recordings of impact sounds of 50 everyday objects with detailed annotations, including their impact locations, microphone locations, contact force profiles, material labels, and RGBD images.*[*] *We make preliminary attempts to use our dataset as a reference to current simulation methods for estimating object impact sounds that match the real world. Moreover, we demonstrate the usefulness of our dataset as a testbed for acoustic and audio-visual learning via the evaluation of two benchmark tasks, including listener location classification and visual acoustic matching.*

## 1. Introduction

Object sounds permeate our everyday natural environments as we both actively interact with them and passively perceive events in our environment. The sound of a drinking glass bouncing on the floor assuages our fear that the glass would shatter. The click made by a knife making contact with a cutting board assures us that we have diced cleanly through a vegetable. And listening to the sound a painted mug makes when we tap it informs us of whether it is made of ceramic or metal. What we perceive from sound complements what we perceive from vision by reinforcing, disambiguating, or augmenting it.

Understanding the cause-and-effect relationships in these sounds at a fine-grained level can inform us about an object's material properties and geometry, as well as its contact and other environmental conditions. Capturing these relationships from real-world data can help us improve our models toward more realistic physical simulations, with applications in virtual reality, animation, and training learning-based frameworks in simulation.

The sounds we perceive from objects are the result of many intricate physical processes: they encode important properties about the object itself (e.g., geometry, material, mechanical properties), as well as the surrounding environment (e.g., room size, other passive objects present, materials of furniture in the room). More specifically, when a hard object is struck, it vibrates according to its mass and stiffness, and the shape of the object determines the *mode shapes* of the dominant vibration patterns (§3.1). Acoustic waves are then emitted into the medium, typically air, bouncing around in the room and interacting with surrounding objects and the room itself before reaching our ear or a microphone to be perceived as pressure fluctuations (§3.2).

Prior work has explored using physical simulation [26, 54] or learning-based methods [28, 29] to reconstruct the sound generation process virtually, as well as building 3D environments with simulated spatial audio for embodied audio-visual learning [7, 15, 18, 35, 42]. However, there has been little work on building physical apparatuses and feasible measurement process to quantify sounds made by the everyday objects, despite their importance and intimate relationship with our daily lives. As a result, the evaluations of the methods above are largely established on subjective metrics such as user studies.

To address this gap, we introduce REALIMPACT, a dataset containing 150k recordings of 50 everyday objects, each being struck from 5 distinct impact positions. For each impact point, we capture sounds at 600 field points to provide comprehensive coverage of the frequency components of the sounds and how they are distributed spatially. REALIMPACT thus provides all the inputs most current simulation frameworks needed to simulate each sound, while also providing the ground truth recording for comparison. We show that REALIMPACT can be used for various downstream auditory and audio-visual learning tasks, such as listener location classification (§5.2) and visual acoustic matching (§5.3). These results demonstrate that sound

[*] The project page and dataset are available at https://samuelpclarke.com/realimpact/

fields can help improve machine perception and understanding of the world, and motivate further studies of even more accurate simulation methodologies to reduce the sim-to-real gap for future applications.

We make three contributions. First, we design an automated setup for collecting high-fidelity, annotated recordings of sounds by controlled striking of everyday objects. Second, using this setup, we acquire a large dataset of spatialized object sounds, REALIMPACT. Third, we motivate the utility of REALIMPACT by (a) using it to perform comparisons to results generated by current state-of-the-art sound simulation frameworks and (b) evaluating two benchmark tasks for acoustic and audio-visual learning.

## 2. Related Work

**Datasets of Object Sounds.** Many datasets of object sounds have been introduced, each varying in the details of their collection, based on the applications they target. The *Greatest Hits Dataset* [40] includes audio-video recordings of thousands of impacts between real objects from the wild. The recordings were not taken in a controlled environment, and each impact is induced by a human with a drumstick, polluting each object's impact sound with the sound of the rather resonant drumstick. The *Sound-20K Dataset* [59] is a fully synthetic dataset of 20,000 simulated recordings of objects being dropped in virtual environments. More recently, ObjectFolder [17, 21] is introduced as a large dataset of trained implicit models for generating the sounds objects make when impacted at arbitrary locations. However, these are once again trained only on data from simulation, and they do not model the acoustic transfer properties of the objects, only their structural vibratory response.

**Physics-based Sound Rendering.** Realistic sound rendering has been a long-held goal in computer music, interactive virtual environments, and computer animation [27, 34, 51, 55]. By modeling the underlying physical processes of vibrations, the computer graphics community demonstrated convincing synthesized sounds for vibrating solids [26, 31, 38, 39], shells [5], rods [47], and even fluids [32, 55]. [26, 55] further showed that it is important for high-quality sound rendering to capture the amplitude and spatial structure of radiating sound fields. However, computing these acoustic transfer fields is time-consuming as they are typically solved in the frequency domain, one frequency at a time. In KLEINPAT [54], the authors showed that by conflating multiple vibrating modes into one time-domain solve, getting all-frequency transfer maps can be done much faster, usually on the order of minutes. These models require careful simulation and material parameters tuning for the best results. To alleviate this issue, several works proposed to sample audio clips [45] and impulse responses [53] to reconstruct the material definitions. Recently, a few works [28, 29] have explored using learning to approximate both the vibration and transfer computations using simulated training data. We provide timely real data that such simulations could use to validate their outputs and tune their performances.

**Recording Sounds Made by Real Objects.** Whereas many learning-based frameworks have traditionally used simulation results as their "ground truth" for learning acoustic models of object vibrations and their transfer, some works have proposed to fit acoustic object models directly from data using digital signal processing with more relaxed model assumptions about rigid body vibrations.

Pai *et al*. [41] describe a framework for scanning physical objects across multiple modalities, measuring visual, tactile, and audio properties of some everyday objects. They fit a data-driven acoustic model based on modal vibration for an object by striking it at different points and recording the ensuing sound from a single position per impact point, assuming constant acoustic transfer across the object. DiffImpact [12] similarly fits modal models to real recordings of objects, but assumes a constant modal response and transfer across the object since their data lacked annotations of the impact point and microphone location. Perhaps most similar to our work, Corbett *et al*. [13] collect recordings of striking an object at three different impact points, positioning a microphone at 19 different positions per impact point. We collect recordings from an order of magnitude more microphone positions to empirically demonstrate that acoustic transfer varies rather drastically over a much finer resolution than can be captured at 19 different locations.

Also, while these prior works have collected datasets from real audio, none have publicly released their datasets. Furthermore, since many simulation frameworks are designed to simulate audio of objects vibrating freely in an anechoic space, the recordings collected by these works are unsuitable for a fair comparison, since they are recorded with objects in contact conditions like resting on tables or grasped in hands, which greatly hinder free vibrations. In contrast, we propose a novel capture system where objects rest on a thread mesh in an acoustically treated room, which more closely approximates free vibration in an anechoic environment.

Finally, from outside the domain of object impact sounds, Bellows *et al*. [4] have an extensive project measuring the sound directivity of musical instruments while being played by musicians. The measurements take place in a large anechoic chamber and are recorded with a rotating semi-circular microphone array resulting in 2,522 unique microphone positions. The raw measurements are not provided, but the directivity patterns are available as spherical harmonic decompositions [3].

**Visual Learning of Sounds in Space.** Both audio and visual data convey crucial spatial information. Recent inspiring works have explored many interesting tasks connecting

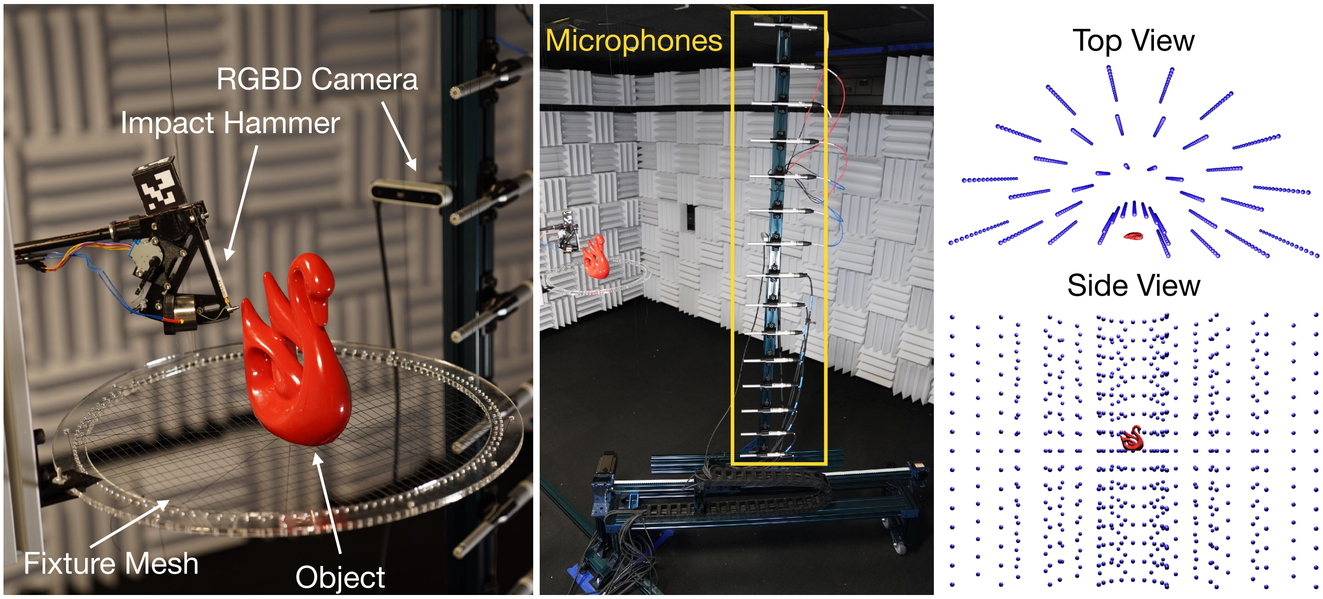

Figure 1. Pipeline for the acquisition of spatialized impact sounds: **(Left)** The object is placed at the center of the measurement platform and aligned with mesh threads. The impact hammer is positioned to strike a target vertex on the object. **(Center)** The gantry moves the microphones to 40 different positions within a semi-cylinder of the object, with the automated hammer mechanism striking the object to record the sound at each position. **(Right)** By the end of the recording process, for each of the 5 vertices of each object, recordings from 600 different microphone positions have been collected within the semi-cylinder to one side of the object, as shown.

visual learning and sound in 3D space, including visually-guided audio spatialization [20,22,37,57,61], sound source localization in video frames [1,6,48], learning audio-visual spatial correspondence [9, 58], and building audio-visual 3D environments [7, 15] for an array of embodied learning tasks [8, 14, 16, 18, 35, 42]. We show how our dataset can be used to evaluate real-world performance of auditory and audiovisual learning frameworks on two novel tasks.

## 3. Physics-Based Sound Synthesis

We begin with some background about physics-based sound simulation for rigid objects to motivate design choices for our dataset and provide context for our baseline simulation frameworks and their parameters. Here we briefly summarize a commonly used sound synthesis pipeline for rigid objects. For a more detailed introduction, we recommend the article from James *et al.* [27].

### 3.1. Modal Sound Synthesis

When a contact force is applied to an object (e.g., your dinner plate hits the dishwasher handle), depending on the location of contact, various vibration modes can get excited and eventually die down due to internal damping. Mathematically, the vibration's displacement vector $\boldsymbol{u}(t)$ can be low-rank approximated as

$$\boldsymbol{u}(t) = U\boldsymbol{q}(t) = [\hat{\boldsymbol{u}}_1 \cdots \hat{\boldsymbol{u}}_K]\boldsymbol{q}(t), \tag{1}$$

where $U$ is the modal matrix with mode shapes $\hat{\boldsymbol{u}}_i$ and $\boldsymbol{q}(t) \in \mathbb{R}^K$ the modal coordinates. The equations of motion are

$$M\ddot{\boldsymbol{u}} + C\dot{\boldsymbol{u}} + K\boldsymbol{u} = \boldsymbol{f}, \tag{2}$$

where $M$, $C$, and $K$ are the mass, damping, and stiffness matrix, respectively*, and $\boldsymbol{f}$ is the external force vector. It is typically assumed in the literature that the damping can be approximated by Rayleigh damping, $C = \alpha M + \beta K$; with this convenient assumption, Eq. (2) can be re-written in the subspace defined by $U$ as

$$\ddot{\boldsymbol{q}} + (\alpha\mathbf{I} + \beta\Lambda)\dot{\boldsymbol{q}} + \Lambda\boldsymbol{u} = U^\intercal\boldsymbol{f}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix and $\Lambda = \mathrm{diag}(\omega_1^2, ..., \omega_K^2)$ is a diagonal matrix of involving angular frequencies $\omega_i$. Since the damping can significantly affect material perception [30], the Rayleigh damping can potentially model real-world objects poorly. In addition, there are two scalar properties $\alpha$ and $\beta$ to fit, and in previous work, these are typically hand-picked to produce sounds that are closest to a given material. Also note that this formulation is based on linear modal analysis [49,50], which assumes the vibrations are infinitesimal, or, in other words, the object is approximately rigid.

### 3.2. Acoustic Transfer

Sound radiates from an object's surface into the surrounding medium as pressure waves. Since the modes de-

*Formulas of how to compute these matrices can be found in [39].

cay slowly over time, it is convenient to work in the frequency domain [26]. The distribution of the wave magnitudes in space, $\hat{p}(\boldsymbol{x};\omega)$, is referred to in the literature [26] as the *acoustic transfer* function. With a given set of vibrational boundary conditions, such as those given by Eq. (3), one can solve the frequency-domain Helmholtz equation [5, 26, 31] using boundary element methods [2, 11, 23] or a time-domain wave equation before reconstructing the acoustic transfer fields [54]. To display the acoustic transfer fields at runtime, we may compress the representation using multi-point dipole sources [26] or single-point multipole sources [60]. Another increasingly popular way of storing and displaying the field is to leverage the radial structure (or lack thereof) in the far-field radiation to store Far-field Acoustic Transfer (FFAT) maps, which are rectangular, image-like textures that capture the angular radiation pattern of a given object. FFAT maps can be quickly reparametrized (e.g., equalized, time-delayed) at runtime to display object impact sounds with user interactions [54].

## 4. The REALIMPACT Dataset

We introduce REALIMPACT, a dataset of 150,000 real object impact sounds. Along with these sounds, we record the force profiles from the impact hammer we used to strike the objects, as well as an RGBD image of the object from each azimuth angle and radial distance from which the audio recording is measured. Below, we introduce the hardware setup for collecting the data, the objects we use, and our data collection pipeline.

### 4.1. Hardware Setup

We collect all recordings in an acoustically treated room (see Appendix B for additional details). We designed a cylindrical gantry system for moving the microphones to precise positions in space, shown in Figure 1. The gantry system moves a 1.82-meter-tall vertical column of 15 Dayton Audio EMM6 calibrated measurement microphones which are evenly and precisely spaced along the column. It moves this column precisely in two degrees of freedom: azimuth and distance, with a precision of 1° and 1 mm, respectively. We suspended a mesh of polyester threads precisely at the axis of rotation of this gantry, centering it vertically along the column of 15 microphones. This mesh holds the objects in place while minimizing contact damping and maximizing the acoustic transparency of the surface holding the object. Furthermore, the layout of the mesh provides visual guidance for precisely positioning the objects in a repeatable manner.

To measure the acoustic transfer from the object to the microphones, the impact force needs to be recorded, allowing an input-output relation to be found. We used a PCB 086E80 impact hammer to strike each object. The impact hammer is incorporated into a custom automated striking mechanism, which strikes objects precisely and repeatedly while being as silent as possible. The mechanism uses a motor to wind the hammer back to contact an electromagnet; then, upon recording, the electromagnet releases the hammer. Actuating the electromagnet is completely silent, so the noise created by this mechanism is minimal during each strike. This mechanism is mounted on a microphone stand to be able to position it rigidly to strike objects at arbitrary locations. See Appendix C and D for additional details on the recording apparatus and hammer impacts, respectively.

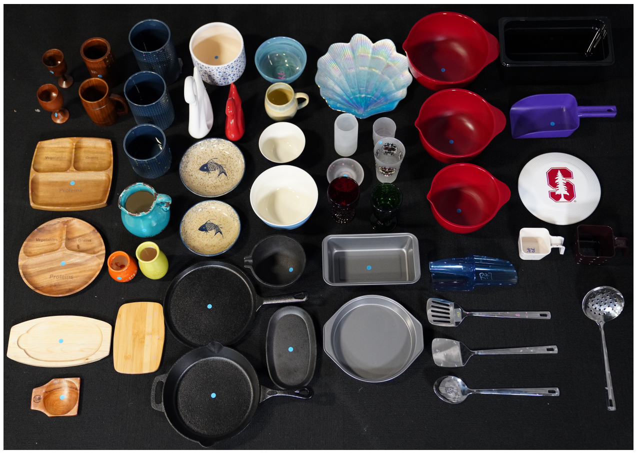

Figure 2. The real objects used in our dataset. Objects are clustered by material: (from top left) wood, ceramic, glass, plastic, (from bottom center) iron, steel, and polycarbonate.

The impact hammer has a calibrated force transducer in its tip, measuring contact forces at the same temporal resolution as our audio. The impact hammer and microphones are all read in a time-synchronized fashion by using two Motu 8M audio interfaces connected by an optical cable. Each audio interface has digitally-controlled amplifier gains, which must be tuned up or down for object sounds that are relatively quiet or loud, respectively, to boost the signal as much as possible while also preventing clipping. Because these gains are digitally controlled, we can record and adjust them in a precise and repeatable manner throughout our experiments. The recordings were made at a sample rate of 48000 Hz.

We also attach a RealSense D415 RGBD camera to the column, aligned with the first microphone above mesh-level, to take RGBD images with our audio measurements.

### 4.2. Objects

We purchase 50 objects from the ObjectFolder dataset [19], which is comprised of commonly used household objects like a ceramic mug, drinking cup, plastic bin, and wood vase. Each object in REALIMPACT has a high-resolution 3D mesh model generated from a scan of the real object [21], which can be used in simulation frameworks. We select objects which are rigid and consist of a single ho-

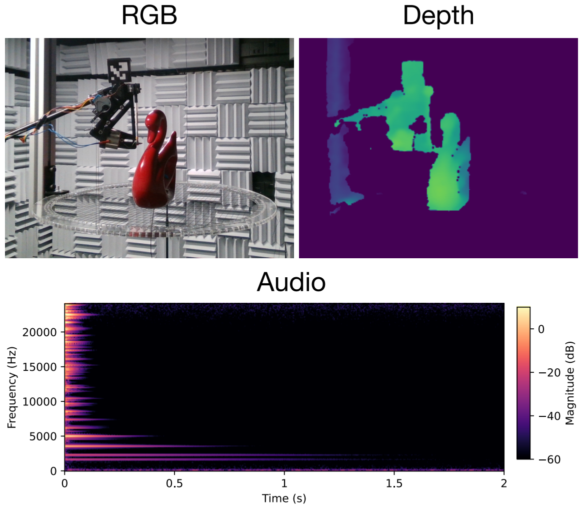

Figure 3. An RGBD image and audio recordings from all 15 microphones are collected at each gantry position for each vertex we impact on an object.

mogeneous material belonging to one of the following categories: ceramic, glass, wood, plastic, iron, polycarbonate, and steel—materials that have $\alpha$ and $\beta$ parameters available and widely used in the physics-based sound rendering literature [28, 29, 54]. Figure 2 shows all objects used for data collection. These objects all have a scale and mass suitable for data collection using our hardware setup.

### 4.3. Data Collection Pipeline

For each object, we first place the object on the supporting mesh, matching the features of the mesh to the distinctive geometric features of the object to position it in a repeatable manner. We then select 5 vertices from the virtual mesh at which to strike the object. For each vertex, we first position the hammer mechanism to strike the vertex. Since our gantry collects recordings on a semi-cylinder to one side of the object, we position the hammer mechanism to the opposite side of the semi-cylinder both to not impinge the motion of the gantry and to minimize blocking acoustic radiation from the surface of the object toward the microphones. For each vertex, we move the gantry to 40 positions: a grid of 10 angles in 20-degree increments from 0 through 180, at 4 distances of 0.23, 0.56, 0.90, and 1.23 meters from the center of the mesh. We take an RGBD image at each position in addition to the audio recordings. A diagram of the microphone positions relative to an example object is shown in Figure 1. An example of each of the modalities captured from one position is shown in Figure 3. See the Supplementary Materials for an example video of how an object is recorded.

### 4.4. Processing

The impact hammer strikes are not necessarily constant across measurements, but the discrepancy can be corrected since the force is measured. This is achieved by deconvolving the force signal from the microphone signal with frequency domain division as $m_c = \mathcal{F}^{-1}\left(\mathcal{F}\left(m\right)/\mathcal{F}\left(i\right)\right)$, where $i$ is the impact hammer signal, $m$ is the microphone signal, and $m_c$ is the corrected microphone signal. $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the forward and inverse discrete Fourier transforms, respectively. The hammer signal is windowed such that only the samples within 1% of the force peak are kept, and all other samples are deemed noise and set as zero, reducing noise in the corrected microphone signal.

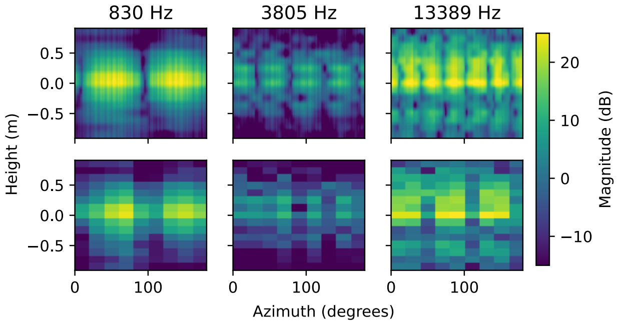

Figure 4. Comparing different azimuth resolutions for mode shape transfer maps of a ceramic bowl, measured at 23 cm from the center. The top row shows maps measured at a resolution of 1 degree, while the bottom row shows those measured at a resolution of 20 degrees.

To create transfer maps of the recordings, mode fitting is performed on each corrected microphone signal. The modes are fit using the method of [10]. First, the vibrational frequencies are fit with a simple peak-picking algorithm performed on $\mathcal{F}\left(m_c\right)$. Decay rates are fit by bandpassing $m_c$ at the mode frequencies, applying a Root-Mean-Squared level detector, and using linear regression to estimate the slope of the energy envelope. The amplitudes are set as the magnitude of the mode frequency peak in $\mathcal{F}\left(m_c\right)$. Transfer maps are then formed for each vibrational frequency by displaying the magnitude at each measurement location with respect to rotation and height, as shown in Figures 4 and 5 .

### 4.5. Validation

**Spatial Sampling.** We use a 20° resolution of azimuth angle for the spatial sampling as a compromise to reduce measurement time while still adding benefit for certain sound-related tasks. We take one set of measurements with 1° rotations on one of our objects (a ceramic bowl) as a comparison. Figure 4 shows measured acoustic transfer maps for sample vibrational frequencies with both 1° and 20° microphone rotations. The lowest-frequency mode shape varies gradually with the azimuth angle. But note that at the highest-frequency mode shape shown, the frequency of the repeating spatial pattern is beyond the Nyquist frequency of our azimuth sampling resolution. We show the implica-

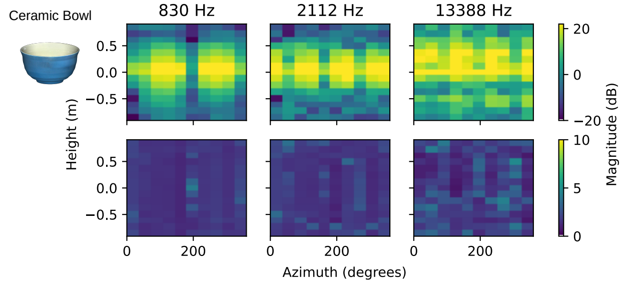

Figure 5. Measuring repeatability of our measurements by visualizing transfer maps of vibrational frequencies of the ceramic bowl, measured at 23 cm from the center. The top row shows the mean of 10 trials of measurements, while the bottom shows the relative standard deviation of the 10 trials.

tions of attempting to naïvely interpolate from these low-resolution transfer maps in Appendix E.

**Repeatability.** We verify two aspects of the repeatability of our design: the repeatability of the gantry's position and the repeatability of our resulting audio measurements. While our gantry is capable of achieving high angular precision while it is controlled, we completely power it off during each recording, in order to eliminate motor and power supply noises from our recordings. During these periods, the wheels may settle into the carpet at a slightly different angle than we have commanded. We perform four trials of moving to the commanded angles and find that the maximum angular error across all trials did not exceed 1°, with the mean error being 0.26°. This translates to a maximum error of 2 cm in Cartesian space, only reached when the gantry is at its farthest distance from the center.

For the repeatability of our measurements, we conduct 10 trials of our measurements on the same ceramic bowl, striking the same target vertex and using the sample positions we used throughout our dataset. We show the mean and standard deviations of the transfers we measured at some sample vibrational frequencies in Figure 5, with results of objects of additional materials in Appendix F. Our results suggest that variations may be highest at the boundaries of nodes in the transfer map. At these locations, minor errors in the azimuth angle could cause significant changes in transfer measurement. Furthermore, at these positions, the signal is lower at this frequency, so the effects of noise can be more pronounced.

# 5. Applications

In this section, we demonstrate some use cases of REALIMPACT with practical, multimodal applications.

## 5.1. Comparing Simulated and Real Impact Sounds

Our first task is to compare sounds synthesized by existing sound rendering methods to the recordings of REALIMPACT in order to demonstrate typical measurement and modeling discrepancies. For this purpose, we ran each baseline method *out-of-the-box* without any attempt to fine-tune its model and/or hyperparameters, including material parameters, such as elastic stiffness (Young's modulus) and damping (e.g., $\alpha$ and $\beta$ in the case of KLEINPAT). We also did not unify the finite element analysis representations across different methods, including finite element type (KLEINPAT uses first-order tetrahedral elements, whereas ObjectFolder 2.0 uses second-order ones), and tetrahedral meshes. These out-of-the-box comparisons simplify the analysis and highlight the ability to benchmark any existing or new simulation methods given a dataset such as REALIMPACT, but exhibit various modeling oversights. We leave the work to narrow the gap between each baseline to the dataset as future work. We provide more conjectures on why these discrepancies exist in the limitations section.

**Baselines.** We provide high-level descriptions of each baseline and refer the readers to Appendix G or directly to the linked work for more details:

- WHITE NOISE: Random noise which has been adjusted to the same loudness as the average loudness of the recordings on a per-object basis.
- RANDOM IMPACT SOUND: A random impact sound recording from our dataset.
- KLEINPAT [54]: The modal analysis is run using first-order tetrahedral elements, and the Far-field Acoustic Transfer (FFAT) maps are done using a one-term ($^1/_r$) scalar expansion.
- NEURALSOUND [29]: The modal analysis is run using an optimization which is warm-started with the outputs of a 3D sparse U-net on voxelized meshes. The FFAT maps are predicted directly by a ResNet-like encoder-decoder structure. For both steps, we use the pretrained weights.
- OBJECTFOLDER 2.0 [21]: The modal analysis is predicted by an implicit neural representation trained on simulation data using second-order tetrahedral elements. No acoustic transfer values exist in this baseline so we used $\hat{p}(\boldsymbol{x};\omega) = 1$ throughout.

Note that the final three baselines all require material properties of the object as input; we uniformly apply the same parameters from Table 4 of [54] for all objects with the same material label in our dataset.

**Metrics.** We evaluate using the following metrics: 1) L1 spectral loss, a loss based on taking the average L1 distance between log-magnitude spectrograms of different window sizes (used for impact sounds in [12]); 2) envelope distance, which measures the distance between two audio samples' envelopes over time (used for spatial audio in [37]); and

|  | RealImpact Deconvolved |  |  | RealImpact Deconvolved + Denoised |  |  |
|---|---|---|---|---|---|---|
|  | L1 Spectral | Envelope ($\times 10^{-3}$) | CDPAM | L1 Spectral | Envelope ($\times 10^{-3}$) | CDPAM |
| White Noise | 4.68 | 9.54 | 1.38 | 5.22 | 9.87 | 1.39 |
| Random Impact Sound | 0.728 | **4.17** | 0.121 | 0.150 | 4.97 | 0.0880 |
| KleinPat [54] | **0.632** | 4.63 | 0.117 | **0.0982** | **4.63** | 0.0975 |
| NeuralSound [29] | 0.673 | 23.0 | **0.102** | 0.133 | 22.8 | **0.0750** |
| ObjectFolder 2.0 [21] | 0.747 | 25.6 | 0.297 | 0.236 | 25.4 | 0.289 |

Table 1. Comparing with simulated object impact sounds. Lower is better for all metrics.

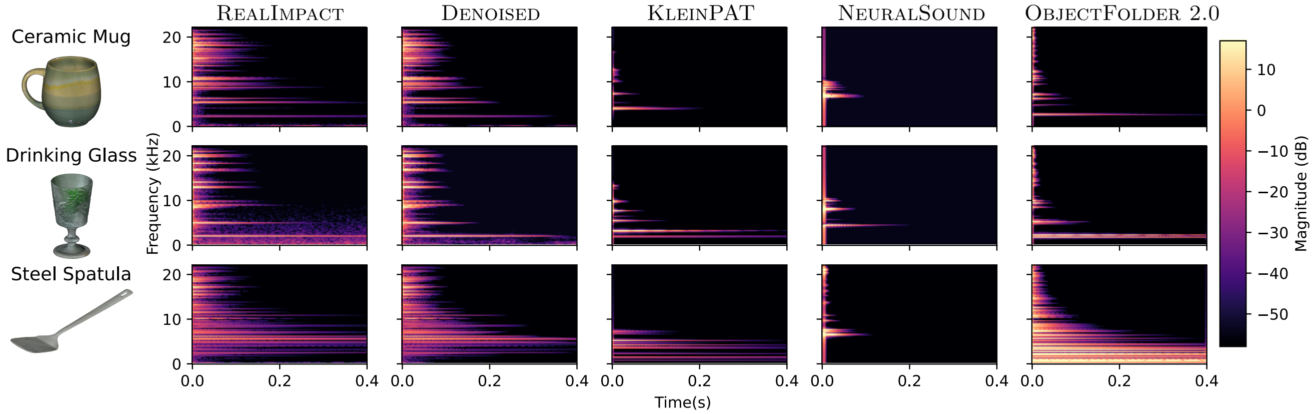

Figure 6. Comparison of spectrograms from our collected recordings versus simulation frameworks. Each spectrogram represents the sound recorded or simulated from a sample point at (8, 22, 13) cm in Cartesian space from the center of the base of the object, and each row corresponds to striking the same vertex on the object.

3) CDPAM [36], a learning-based perceptual distance metric trained from human judgments of detectable differences between clips.

To mitigate the effects of measurement noise in our evaluation, we compare each baseline both to our deconvolved recordings and to denoised versions of our recordings, which have been denoised with the algorithm of [46]. Whereas many denoising algorithms are optimized for human speech, this algorithm has been optimized and validated against broader categories of audio signals from nature. Comparisons of example spectrograms and their denoised counterparts are shown in Appendix H.

Quantitative results are shown in Table 1, and qualitative examples are shown in Fig. 6, comparing our recordings of real impact sounds with the simulated sounds using methods from [21, 29, 54]. The KleinPat baseline performs best according to a spectral loss, whereas NeuralSound performs best according to the perceptual CDPAM loss. Both of these baselines significantly outperform ObjectFolder, suggesting that explicitly modeling the acoustic transfer of objects rather than merely their structural vibrations is essential for achieving realism. A random impact sound only outperforms baselines in Envelope Loss when the recording has not been denoised. Each baseline other than white noise performs better on all metrics when compared against denoised versions of our recordings, suggesting that our raw recordings have non-negligible measurement noise, which must be accounted for in future comparisons.

## 5.2. Listener Location Classification

Identifying the location of the listener with respect to the sound source is of great practical interest to many applications in virtual reality and robotics [8, 43, 44]. In this task, we want to identify the microphone position (angle, height, or distance) from the impact sound recording.

For each impact in RealImpact, we have the recordings of the impact sound from 600 different listener locations collected from 10 different angles, 15 different heights, and 4 different distances, as illustrated in Fig. 1.

Particularly, we set up three separate classification subtasks: 1) angle classification, where the goal is to classify the sound into the 10 angle categories (0°, 20°, . . ., 180°); 2) height classification, where the goal is to classify the sound into the 15 height categories, each corresponding to the height of our 15 microphones; and 3) distance classification, where the goal is to classify the sound into the 4 distance categories (0.23 m, 0.56 m, 0.90 m, and 1.23 m).

For each subtask, we split 90/10 percent of impact sound recordings of an object into the train/test set, respectively. We train a ResNet-18 [24] network that takes the magnitude spectrogram of the impact sound as input to predict the an-

|  | Angle | Height | Distance |
|---|---|---|---|
| CHANCE | 10.0 | 6.7 | 25.0 |
| Ours | **57.9** | **60.7** | **67.4** |

Table 2. Listener location classification results. We report the accuracy (in %) for angle, height, and distance classification, respectively.

gle, height, or distance category. Table 2 shows the results averaged across all 50 objects. We observe that predicting height is comparatively easier. We suspect that differences in height strongly influence the spectral details for easier classification.

### 5.3. Visual Acoustic Matching

The ability to match a source sound with the correct corresponding visual input plays an important role in tasks such as speech and speaker recognition [33, 52] or object and event localization [25, 56]. This task aims to match a sound recording with the correct corresponding image. We set up this matching task as binary classification.

For 20 of our objects, we have a total of 200 RGBD images taken simultaneously with our audio recordings, which are collected at a fixed height from the 10 different angles and 4 different distances from which we took audio recordings for each of the 5 different vertices. We generate positive pairs by pairing each sound recording to an image taken at the corresponding angle and vertex. The height and distance of the image are fixed, so there are 50 possible images that the sound recordings correspond to. The distance of the paired RGBD image is selected such that the image captures the position of both the object and the impact hammer. Negative pairs are generated by pairing sound recordings with images that are not at the correct angle and vertex.

We randomly select two heights to be held out for validation and test sets, while the remaining 13 heights are used for the train set. We train an audio-visual network with a ResNet-18 backbone for both the image and audio streams. The network takes in an RGB image as visual input and an impact sound recording as audio input. A fusion layer combines the audio and visual information, and a final fully-connected layer is used to extract audio-visual features for binary classification. Table 3 shows the quantitative results of the visual acoustic matching task averaged across 20 objects, and we show example inputs and outputs in Appendix I.

## 6. Limitations and Conclusion

We presented REALIMPACT, a first-of-its-kind, large dataset of 150k real impact sounds systematically collected in an acoustically treated room, and demonstrated its several use cases on benchmarking existing simulation algorithms and applications on several auditory and audiovisual tasks.

|  | Accuracy ↑ | RMSE ↓ |
|---|---|---|
| CHANCE | 50.0 | 59.7 |
| Ours | **75.1** | **47.7** |

Table 3. Quantitative results of visual acoustic matching. We report the accuracy results (in %). RMSE angle error (in degrees) is the root-mean-square error in the difference between the angles of the image and sound recording. ↑ or ↓ signify higher or lower values are better, respectively.

The microphone array stack used in our measurement process is somewhat coarse to capture high-frequency details (e.g., the 12 cm microphone spacing in elevation roughly corresponds to the wavelength of 3 kHz). The angular resolution is chosen at only 20 degrees and can result in aliasing and distortion in the otherwise symmetric radiation fields, as shown in §4.5. The diversity of the recorded objects is restricted by the size and load capacity of our supporting thread mesh, and the microphone stack arm. The material descriptions of the objects are artificially lumped into the categories defined in previous work [54], but they may not describe the diversity of real-world materials (e.g., different kinds of steels will have different mechanical properties that might affect stiffness and thus the frequency distributions). Future work should look at more efficient ways of capture and sample a wider range of objects.

The comparison of real impact sounds to those generated by current simulation methods exhibit various discrepancies. Many things can lead to the gap between simulations and real recordings: object scanning resolution and reconstruction accuracy, material stiffness and damping parameters, finite-element analysis differences (e.g., element type), and insufficient meshing resolution. Also we did not explicitly model hollow objects in the comparisons despite some of our objects being hollow and/or thin, and contact damping models are missing, which can affect the perceived damping rates and thus the envelope accuracy. As a result, many of the out-of-the-box simulation models' vibrational frequencies do not agree, let alone their estimates of spatialized sound amplitudes, etc. Many of these oversights will affect the comparison "fairness" but it also demonstrates a significant benefit of the REALIMPACT dataset: for the first time, we can measure these models using the same yard stick. We hope that this dataset provides future incentives to improve not just the simulation methods and their usage, but also the capture process and datasets that can move the field forward in a significant way.

**Acknowledgments.** This work is in part supported by Adobe, NSF CCRI #2120095, NSF RI #2211258, ONR MURI N00014-22-1-2740, AFOSR YIP FA9550-23-1-0127, the Stanford Institute for Human-Centered AI (HAI), Amazon, and Meta.

## References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 3

[2] M. Bebendorf. Approximation of boundary element matrices. *Numerical Mathematics*, 86(4):565–589, Oct 2000. 4

[3] Samuel D. Bellows. Directivity. https://scholarsarchive.byu.edu/directivity/. Accessed: 2022-06-06. 2

[4] Samuel David Bellows and Timothy Ward Leishman. Spherical harmonic expansions of high-resolution musical instrument directivities. In *Proceedings of Meetings on Acoustics*, 2018. 2

[5] Jeffrey N Chadwick, Steven S An, and Doug L James. Harmonic shells: a practical nonlinear sound model for near-rigid thin shells. *ACM Transactions on Graphics (TOG)*, 28(5):1–119, 2009. 2, 4

[6] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022. 3

[7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3d environments. In *ECCV*, 2020. 1, 3

[8] Changan Chen, Sagnik Majumder, Al-Halah Ziad, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2021. 3, 7

[9] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. In *ECCV*, 2022. 3

[10] Jatin Chowdhury, Elliot K Canfield-Dafilou, and Mark Rau. Water bottle synthesis with modal signal processing. In *Int. Conf. Digital Audio Effects (DAFx)*, 2020. 5

[11] Robert D Ciskowski and Carlos Alberto Brebbia. *Boundary Element Methods in Acoustics*. Computational Mechanics Publications and Elsevier Applied Science, Southampton. UK, 1991. 4

[12] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. DiffImpact: Differentiable Rendering and Identification of Impact Sounds. In *CoRL*, 2021. 2, 6

[13] Richard Corbett, Kees Van Den Doel, John E Lloyd, and Wolfgang Heidrich. TimbreFields: 3D interactive sound models for real-time audio. *Presence*, 16(6):643–654, 2007. 2

[14] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. In *NeurIPS*, 2020. 3

[15] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. Threedworld: A platform for interactive multi-modal physical simulation. In *NeurIPS Datasets and Benchmarks Track*, 2021. 1, 3

[16] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 3

[17] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. ObjectFolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *CoRL*, 2021. 2

[18] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. VisualEchoes: Spatial Image Representation Learning through Echolocation. In *ECCV*, 2020. 1, 3

[19] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The ObjectFolder Benchmark: Multisensory Object-Centric Learning with Neural and Real Objects. In *CVPR*, 2023. 4

[20] Ruohan Gao and Kristen Grauman. 2.5D Visual Sound. In *CVPR*, 2019. 3

[21] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. ObjectFolder 2.0: A Multisensory Object Dataset for Sim2Real Transfer. In *CVPR*, 2022. 2, 4, 6, 7

[22] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. In *BMVC*, 2021. 3

[23] N. A. Gumerov and R. Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Elsevier Science, 2005. 4

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[25] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *NeurIPS*, 2020. 8

[26] Doug L James, Jernej Barbič, and Dinesh K Pai. Precomputed Acoustic Transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics (TOG)*, 25(3):987–995, 2006. 1, 2, 4

[27] Doug L. James, Timothy R. Langlois, Ravish Mehra, and Changxi Zheng. Physically based sound for computer animation and virtual environments. In *ACM SIGGRAPH 2016 Courses*, 2016. 2, 3

[28] Xutong Jin, Sheng Li, Tianshu Qu, Dinesh Manocha, and Guoping Wang. Deep-modal: real-time impact sound synthesis for arbitrary shapes. In *ACM MM*, 2020. 1, 2, 5

[29] Xutong Jin, Sheng Li, Guoping Wang, and Dinesh Manocha. Neuralsound: learning-based modal sound synthesis with acoustic transfer. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 1, 2, 5, 6, 7

[30] Roberta L Klatzky, Dinesh K Pai, and Eric P Krotkov. Perception of material from contact sounds. *Presence*, 9(4):399–410, 2000. 3

[31] Timothy R. Langlois, Steven S. An, Kelvin K. Jin, and Doug L. James. Eigenmode compression for modal sound models. *ACM Transactions on Graphics (TOG)*, 33(4), 2014. 2, 4

[32] Timothy R Langlois, Changxi Zheng, and Doug L James. Toward animating water with complex acoustic bubbles. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016. 2

[33] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *CVPR*, 2021. 8

[34] Shiguang Liu and Dinesh Manocha. Sound synthesis, propagation, and rendering: a survey. *arXiv preprint arXiv:2011.05538*, 2020. 2

[35] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In *ECCV*, 2022. 1, 3

[36] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In *ICASSP*, 2021. 7

[37] Pedro Morgado, Nono Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. In *NeurIPS*, 2018. 3, 6

[38] James F O'Brien, Perry R Cook, and Georg Essl. Synthesizing sounds from physically based motion. In *SIGGRAPH*, 2001. 2

[39] James F O'Brien, Chen Shen, and Christine M Gatchalian. Synthesizing sounds from rigid-body simulations. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2002. 2, 3

[40] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016. 2

[41] Dinesh K Pai, Kees van den Doel, Doug L James, Jochen Lang, John E Lloyd, Joshua L Richmond, and Som H Yau. Scanning physical interaction behavior of 3D objects. In *SIGGRAPH*, 2001. 2

[42] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman. Audio-visual floorplan reconstruction. In *ICCV*, 2021. 1, 3

[43] Chinmay Rajguru, Giada Brianza, and Gianluca Memoli. Sound localization in web-based 3D environments. *Scientific Reports*, 12(1):1–13, 2022. 7

[44] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017. 7

[45] Zhimin Ren, Hengchin Yeh, and Ming C Lin. Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1):1–16, 2013. 2

[46] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology*, 16(10):e1008228, 2020. 7

[47] Eston Schweickart, Doug L James, and Steve Marschner. Animating elastic rods with sound. *ACM Transactions on Graphics (TOG)*, 36(4):1–10, 2017. 2

[48] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 3

[49] Ahmed A Shabana. *Theory of Vibration: An Introduction*. Springer Science & Business Media, 2012. 3

[50] Ahmed A Shabana. *Dynamics of Multibody Systems*. Cambridge university press, 2013. 3

[51] Julius O Smith. *Physical Audio Signal Processing for virtual musical instruments and digital audio effects*. Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, 2010. 2

[52] Amirsina Torfi, Seyed Mehdi Iranmanesh, Nasser Nasrabadi, and Jeremy Dawson. 3D convolutional neural networks for cross audio-visual matching recognition. *IEEE Access*, 5:22081–22091, 2017. 8

[53] James Traer, Maddie Cusimano, and Josh H McDermott. A perceptually inspired generative model of rigid-body contact sounds. In *International Conference on Digital Audio Effects (DAFx)*, 2019. 2

[54] Jui-Hsien Wang and Doug L James. Kleinpat: Optimal mode conflation for time-domain precomputation of acoustic transfer. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 1, 2, 4, 5, 6, 7, 8

[55] Jui-Hsien Wang, Ante Qu, Timothy R. Langlois, and Doug L. James. Toward wave-based sound synthesis for computer animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018. 2

[56] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019. 8

[57] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *CVPR*, 2021. 3

[58] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *CVPR*, 2020. 3

[59] Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H McDermott, Joshua B Tenenbaum, and William T Freeman. Generative modeling of audible shapes for object perception. In *ICCV*, 2017. 2

[60] Changxi Zheng and Doug L. James. Rigid-body fracture sound with precomputed soundbanks. *ACM Transactions on Graphics (TOG)*, 29(4):1–13, 2010. 4

[61] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020. 3