Hearing Anything Anywhere

Mason Long Wang[1*] Ryosuke Sawata[1,2*] Samuel Clarke[1]
Ruohan Gao[1,3] Shangzhe Wu[1] Jiajun Wu[1]
[1]Stanford University [2]Sony AI [3]University of Maryland, College Park
masonlwang.com/hearinganythinganywhere

Abstract

Recent years have seen immense progress in 3D computer vision and computer graphics, with emerging tools that can virtualize real-world 3D environments for numerous Mixed Reality (XR) applications. However, alongside immersive visual experiences, immersive auditory experiences are equally vital to our holistic perception of an environment. In this paper, we aim to reconstruct the spatial acoustic characteristics of an arbitrary environment given only a sparse set of (roughly 12) room impulse response (RIR) recordings and a planar reconstruction of the scene, a setup that is easily achievable by ordinary users. To this end, we introduce DIFFRIR*, a differentiable RIR rendering framework with interpretable parametric models of salient acoustic features of the scene, including sound source directivity and surface reflectivity. This allows us to synthesize novel auditory experiences through the space with any source audio. To evaluate our method, we collect a dataset of RIR recordings and music in four diverse, real environments. We show that our model outperforms state-of-the-art baselines on rendering monaural and binaural RIRs and music at unseen locations, and learns physically interpretable parameters characterizing acoustic properties of the sound source and surfaces in the scene.*

1. Introduction

Much of the impetus to realize immersive virtual reality (VR) stems from the desire to recreate and share *real* scenes and experiences. Motivated by this goal, recent progress in 3D computer vision and computer graphics has led to tools that can virtualize real-world 3D environments using simple consumer devices (e.g., cellphone cameras) for numerous Mixed Reality (XR) applications. Alongside immersive visual experiences, immersive auditory experiences are equally vital to our holistic perception of an environment. For instance, while the interior of Carnegie Hall in New York City is visually beautiful, one cannot fully appreciate the majesty of its design without experiencing a musical performance in-person and hearing its unique acoustics.

In this paper, our goal is to capture the acoustic intrinsics of a real-world scene using a sparse set of measurements, in order to render arbitrary source audio at any location, hence the name, "Hearing Anything Anywhere". This is analogous to the task of sparse-view novel view synthesis (NVS) in computer vision and graphics [5, 34, 50].

However, there are two key differences between light and sound that make common approaches to visual NVS inapplicable to audio. First, light is typically emitted from continuous sources and travels steadily and almost instantly through space, resulting in a largely stationary visual scene. In contrast, sound signals are usually time-varying and travel through space at a much slower pace, resulting in a constantly changing 4D acoustic field with both numerous early reflections and late reverberations. Second, a single camera captures *millions* of pixels in a split second, each recording a distinct light ray from a *particular* direction. In contrast, a typical microphone only records an amalgamation of sound waves arriving to a *single* location from *all* directions, with different times-of-arrival. Therefore, while it is possible to capture the appearance of a 3D scene by simply walking through it with a camera, the same approach falls short to record the entire 4D acoustic field.

Thus, capturing a fully immersive acoustic field often necessitates setting up hundreds of microphones densely across the space [30, 38, 40, 45], which is impractical for many consumer use cases. In this work, we attempt to capture real-world acoustic spaces with a *basic* hardware setup, e.g., 12 microphones, which can be easily scaled to arbitrary environments.

To capture the acoustic properties of the scene, we measure a room impulse response (RIR) between the sound source and each microphone location. An RIR is a time-series signal that estimates how a perfect impulse emitted from the source, traveling and bouncing in the room, would be perceived at the listener location. RIRs effectively capture a room's intrinsic acoustic properties between source

*Equal contribution.

and listener points, and are thus widely used in acoustic simulation [3]. In order to simulate the sound of an arbitrary source for a particular listener location in a room, the RIR associated with the source-listener pair is simply convolved with the source audio [28].

We thus formulate our *Hearing Anything Anywhere* task as inferring RIRs and music at novel listener locations from a sparse set of RIRs measured between a single source and a small set of microphone locations spatially distributed within the scene. Towards this goal, we introduce a fully differentiable impulse response rendering framework DIFFRIR that reasons about the individual contributions of each acoustic reflection path between the source and the receiver, including the time delay and magnitude of the sound on each path, as well as the influence of reflections from each surface in the scene.

By explicitly modeling the sound source location, the directivity map of the source, and the reflection properties of the surfaces in the scene in a fully differentiable audio rendering framework, we can characterize the parameters of each model through an analysis-by-synthesis paradigm by optimizing the output of DIFFRIR against the known subset of measured RIRs. After optimizing the interpretable parameters of our model, we can estimate the RIR from any unseen location in the scene.

To validate our method, we collect a dataset that contains RIR measurements from four real-world environments that represent a diverse range of room materials, shape, and complexity. Through experiments comparing our framework with current state-of-the-art methods, DIFFRIR shows greater robustness in real, data-limited scenarios. Moreover, with the explicit and interpretable models of source and surface reflection properties, we can easily synthesize novel auditory experiences with different speaker orientations and locations, which can be useful in applications such as virtual reality and acoustics-aware interior design. In addition, the differentiable and interpretable models of our framework allow us to estimate acoustic parameters of the sound source and surfaces in the room, which can be useful in applications like robotics and architectural design for acoustics.

Our contributions are threefold. First, we contribute DIFFRIR, a differentiable acoustic inverse rendering framework that can recover the fully immersive acoustic field of a room from a set of 12 sparsely located RIR measurements. Second, we contribute a new dataset of real-world RIRs measured from hundreds of locations in four different real environments. Third, we compare our method to existing methods across various settings, demonstrating that our method is more effective than existing methods on real data in our data-limited scenarios, predicting more accurate RIRs and music at unseen locations. Code and data are available at the project website.

2. Related Work

Learning-Based Room Acoustics Prediction. While many acoustical learning frameworks model room acoustics implicitly, others explicitly interpolate and predict RIRs at novel points. Frameworks that predict RIRs at novel points in a room vary not only in their underlying techniques, but also in their inputs. Some methods do not use vision or geometry to make their estimates, but instead learn to directly approximate a function mapping spatial coordinates to RIRs [38, 40]. These methods can require large training set sizes on the order of 1,000 RIRs from a room to effectively interpolate RIRs to novel points within the same room. Alternatively, some methods use geometric features of the scene [30], such as [45], which learns a diffuse reflection model from a small subset of points in the mesh of the environment, to achieve a performance improvement over pure audio-based methods. Our method uses environment geometry to explicitly model specular reflections on each surface. To validate our approach, we compare against three baselines, including one audio-only method [40] and two methods that use scene geometry [30, 45].

Audio-Visual (AV) Room Acoustics Prediction. Other methods learn relationships between visual inputs and room acoustics to perform tasks such as predicting the dereverberated signal from an audio recording and a panoramic image of the recording environment [12], or predicting how an input audio signal would sound in a target space based on an image of the space [9]. Many works use visual inputs to explicitly perform the novel view acoustic synthesis (NVAS) task. For instance, Chen et al. [11] proposed the Visually-Guided Acoustic Synthesis (ViGAS) network, which outputs the spatial audio of the speech of a human in corresponding visual frames. Furthermore, by using audio-visual features as well as geometric ones, Ahn et al. [1] show that the important sub-tasks of NVAS, e.g., sound source localization, separation, and dereverberation, can be jointly solved. AV-NeRF [29] improved the performance of both NVS and NVAS tasks via multi-task training by using an audio-based Neural Radiance Field (NeRF). Their audio NeRF estimates variations in the magnitudes of audio perceived from varying locations, whereas we explicitly estimate the RIR, a much more holistic characterization of the environment acoustic properties.

Similar to our binaural prediction task, Garg et al. [19] predict binaural audio from an AV scene's monaural audio and visual features extracted from the scene's video frames. Although AV approaches can sometimes outperform uni-modal audio-only models at estimating environment acoustics, collecting large enough datasets of synchronized audio-visual pairs for these models can be laborious. Perhaps for this reason, many such models, even one boasting few-shot generalization [31], present results from eval-

uating exclusively on simulated data.

Geometry-Based RIR Simulation. Many of the aforementioned works use datasets of simulated RIRs generated by the SoundSpaces framework [10], a fast acoustic simulator based on geometric acoustic methods. They simulate the acoustics of virtualized versions of real rooms from datasets of meshes reconstructed from RGBD scans of real rooms in home and workplace environments, such as the Matterport3D dataset [8] or the Replica dataset [44]. The Geometric-Wave Acoustic (GWA) dataset uses a hybrid propagation algorithm combining wave-based methods [22] with geometric acoustic methods, intending to model low-frequency wave effects more accurately, albeit at the cost of longer run-time. The input meshes are from a dataset of professionally designed virtual home layouts [18]. The Mesh2IR framework uses the GWA dataset to learn a conditional generative adversarial network (cGAN) to more quickly predict RIRs from meshes of rooms [39]. The authors do not show how their cGAN's estimates of RIRs compare to measured RIRs from real rooms.

Differentiable Acoustics. The previously mentioned simulators are not differentiable, which precludes gradient-based optimization techniques which can be used in solving inverse problems. Differentiable audio rendering techniques have been used to solve such inverse problems estimating acoustic properties of musical instruments [17] and everyday objects [14], as well as the reverberation properties of the environments they are in. The authors of [13] implemented a differentiable acoustic ray tracer for inverse tasks in underwater acoustics, such as estimating the absorption of the seabed on simulated 2D data. We use similar principles for estimating absorption parameters of surfaces in 3D environments from our real, airborne sound data.

3. Method

We first lay out the definition of our task, and then introduce our proposed DIFFRIR framework to approach it.

3.1. Task Formulation

To achieve our goal of virtualizing real acoustic spaces, our method should require information about the room that is as easy as possible to obtain. With this objective in mind, we show that our method produces accurate results, while only requiring the following:

1. A small set of omnidirectional RIR recordings captured at sparse locations (e.g., 12), with the xyz coordinates at which they were captured.
2. The room's rough geometry, expressed as a small number of planes.

RIRs can be easily captured by playing a sine sweep from the source location and recording it from a microphone at the listener location. In our setup, we assume a stationary audio source whose orientation and position are unknown. With this information, our goals are to simulate monoaural and binaural RIRs and music at arbitrary listener locations and orientations in the room.

3.2. The DIFFRIR Framework

To achieve this task, we design a differentiable RIR rendering framework, dubbed DIFFRIR. As an overview of the DIFFRIR framework, we use the sound source and microphone location, along with the planar decomposition of the environment, to trace all specular reflection paths between the source and a listener location, up to a certain number of reflections. We estimate the sound arriving to the listener from each path using a series of parametric models for the sound source directivity and impulse response, as well as the acoustic reflection of each surface. Each model is fully differentiable, with interpretable parameters. We compute each RIR as the sum of contributions of the sound arriving from each path, combined with a learned residual. We use these models in a differentiable audio renderer to optimize parameters according to a loss function comparing our estimates to the known subset of ground-truth RIRs. We describe each model in detail below.

3.2.1 Characterizing the Sound Source

Source Localization. We first estimate the location of the sound source for all subsequent steps. Based on the known subset of RIRs we use their locations and the timing of the first peak to localize the source using a traditional time-of-arrival method. More details are provided in Appendix E.

Source Directivity. Most real sound sources do not radiate sound uniformly in all directions. For instance, a loudspeaker will usually be much louder from the front, and human speakers also have distinct directivity patterns [37]. The source's *directivity* describes the way in which the source radiates sound differently in different directions and is generally frequency dependent. For example, a loudspeaker will overall sound much louder from the front, with the higher-frequency components radiating in especially narrow beams and lower-frequency components more omnidirectionally. The sound source's directivity has a significant impact on the acoustic field of the room and is therefore important to model.

We model the filtering effect of exiting the sound source in any particular direction with the *directivity response*. Let $\vec{d_p}$ be the absolute direction (given as a unit vector) in which the sound path exits the speaker. Our goal is to fit $D(\vec{d_p})$, a function mapping $\vec{d_p}$ to a magnitude frequency response that accounts for the effect of exiting the speaker in the direction of $\vec{d_p}$. When a sound exits the speaker in the direction of $\vec{d_p}$, the frequency content of the sound wave is *multiplied* by $D(\vec{d_p})$.

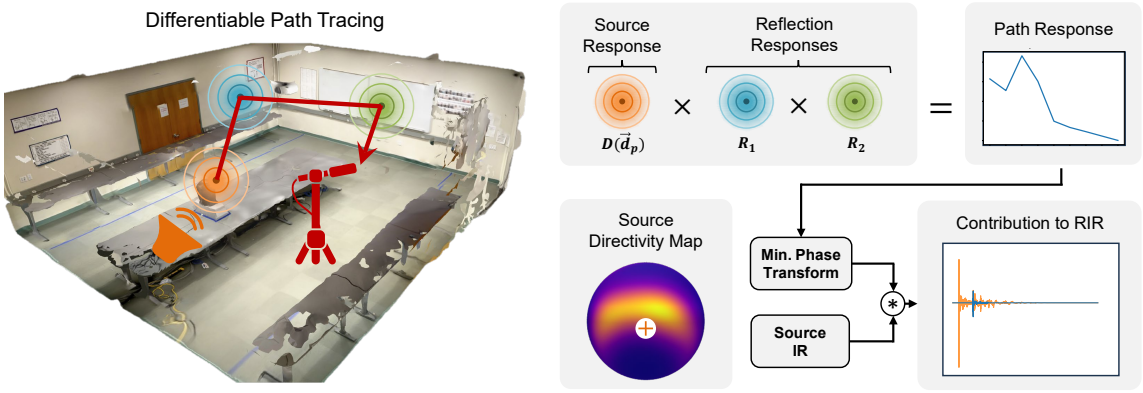


Figure 1. **Differentiable Room Impulse Response Rendering Framework (DIFFRIR).** Our model renders the contribution to the RIR of a single traced reflection path. After computing a reflection path, we characterize it by the direction at which it exits the speaker, its length, and the surfaces on which it reflects. The sound source has a learned frequency response that depends on the outgoing direction, and each surface has a different learned frequency response. We multiply each of these responses to estimate the overall path response. To determine the reflection path's time-domain contribution to the final RIR, we apply a minimum-phase inverse-Fourier transform to the path response, convolve it with the source impulse response, and then shift the result in time based on the path length and the speed of sound.

To model the direction-dependent frequency response, we fit F different heatmaps on unit spheres centered on the speaker, one heatmap for each of F octave-spaced center frequencies comprising vector $\mathbf{f}$. To do this, we distribute 128 points evenly along the surface of the unit sphere, using a Fibonacci lattice [23]. We denote this set of points L. Let $A_{\vec{x},f_o}$ be the log-amplitude gain for sound traveling out of the speaker in the direction of $\vec{x}$ at frequency f_o. To determine the log-amplitude gain at f_o in direction $\vec{d}_p$, we interpolate between the points on the heatmap using a spherical Gaussian weighting function, inspired by [49]:

$$A_{\vec{d}_p,f_o} = \frac{\sum_{x\in L} A_{x,f_o} e^{-\lambda(1-\vec{d}_p\cdot x)}}{\sum_{x\in L} e^{-\lambda(1-\vec{d}_p\cdot x)}}, \tag{1}$$

where λ is a fixed sharpness value shared across all heatmaps. In order to obtain the full frequency response for the direction d, we linearly interpolate between the log-amplitude gains as in [24], and then exponentiate them to convert them to linear amplitude values:

$$D(\vec{d}_p, f_o) = \mathrm{e}^{\ell(\mathbf{A_d},\mathbf{f},\mathrm{f_o})}, \tag{2}$$

where ℓ represents linear interpolation on the vector of decibel values $\mathbf{A_d}$ indexed by center frequencies $\mathbf{f}$, based on query frequency f_o.

Source Impulse Response. Since the room impulse response relates the source signal fed to the speaker to the sound heard in the room, we must also account for the way that the source modifies the source signal being fed to it. For instance, if the source is a loudspeaker, it may attenuate or boost certain frequencies. We model these effects by learning a source impulse response IR_s in the time domain, thus approximating the source's response as a linear system [6] and convolving it with our RIR.

3.2.2 Modeling and Characterizing Reflections

We trace each specular reflection path and model the acoustic effects of each reflection along the path, with unique reflection parameters for each surface in the environment.

Reflectivity. When a sound wave encounters a surface, a fraction of the sound wave's energy will be specularly reflected, while the remaining energy will be absorbed, transmitted, diffusely reflected, or diffracted. These effects vary by frequency, depending on the texture and material properties of each surface.

For each surface s, we fit a vector $\mathbf{V_s}$ of F different values representing the magnitude of sound specularly reflected by the surface at each of F octave-spaced centered frequencies in vector $\mathbf{f}$. We apply the sigmoid function to these values to determine the *energy* reflection coefficients (the proportion of specularly reflected sound energy) at each frequency. Next, we determine the *amplitude* reflection coefficients (the amount that the surface attenuates the incoming sound at each frequency in terms of linear amplitude gain) by taking the square root of the energy reflection coefficients [26]. Using the amplitude reflection coefficients at the F center frequencies, we obtain the amplitude gains for arbitrary frequencies through linear interpolation. This gives us the *reflection response* R_s, a magnitude frequency response representing the surface's effect on incoming audio of different frequencies. Thus, the formula for R_s is:

$$R_s(f_r) = \ell\left(\sqrt{\sigma(\mathbf{V_s})}, \mathbf{f}, f_r\right). \tag{3}$$

Here, σ denotes the sigmoid function, and ℓ is a linear interpolation from the coefficients $\mathbf{V_s}$ based on the relation of the query frequency f_r to the center frequencies $\mathbf{f}$.

Reflection Paths. Given the estimated source location S_{xyz}, a listener location L_{xyz}, and a planar representation of the room's geometry, we use the image-source method [2] to efficiently compute all of the specular reflection paths between the source and listener in the room, up to a particular order N (e.g., 5). The method considers all permutations from 1 to N of these surfaces with repetition and, for each permutation, determines if there is a valid reflection path that travels from the source to the listener after reflecting specularly off of each of the surfaces in order. For each valid reflection path p from source to listener, we track the length of the reflection path l_p, the ordered list S_p of reflection surfaces along the path, and the direction from which the path exits the source $\vec{d_p}$.

Rooms often contain parallel surfaces, which lead to prominent higher-order reflections. These reflections result in "axial modes," which are powerful room resonances with especially long reverberation times [41]. Thus, in addition to computing all N^{th}-order reflection paths for all possible orderings of surfaces, our image-source algorithm also computes all valid reflection paths for pairs of parallel walls, up to a much higher order, e.g., 50. This modification, which we call *axial boosting*, improves the model's performance (see Appendix D.4) in adversarial cases like the Hallway, with a computational overhead that scales linearly rather than exponentially with reflection order. We discuss additional surface interactions, such as diffuse reflection, in Section 3.2.3.

3.2.3 Combining Models

We combine these reflection and sound source models to estimate the contribution of each reflection path. We then sum the contributions across all paths and add a residual to estimate the RIR for a given source and listener location.

Contribution of a Single Reflection Path. In summary, for each individual reflection path p, the outgoing direction $\vec{d_p}$ from the source, the ordered list S_p of reflected surfaces, and the total path length l_p each have distinct effects on rendering the path's contribution. $D(\vec{d_p})$ characterizes the frequency response of the source from the path's outgoing direction. The reflection of each surface $s \in S_p$ attenuates the amplitude of the sound in a frequency-dependent fashion parameterized by R_s. The total reflection-based attenuation is the product of the frequency response across all $s \in S_p$. Finally, we use the path length l_p to compute the time of arrival t_p by dividing the path length l_p by the speed of sound. We also use l_p to estimate the attenuation of the amplitude due to spherical propagation, where the amplitude is inversely proportional to l_p, as well as air absorption, which we characterize by air absorption coefficient α [43].

Thus, the function K that computes the time-domain contribution of each individual path is:

$$K(d_p, S_p, t_p) = \frac{\alpha^{t_p}}{\rho}\tau \left[\mathcal{M}\left(D(d_p) \odot \prod_{s \in S_p} R_s \right), t_p \right], \tag{4}$$

where $\odot$ is the element-wise product, ρ is the length of the reflection path in meters, and τ_t is the time-shift operator, which delays its input signal by t_p seconds. $\mathcal{M}$ is a minimum-phase inverse Fourier transform, which computes a time-domain filter from a magnitude frequency response, assuming minimum phase. The minimum phase assumption can be used to approximate the phase of an acoustic reflection given a desired magnitude frequency response [32]. More details are in Appendix E.

Modeling Residual Effects. For the purposes of gradient-based optimization, we require a model that is fast, simple, and differentiable. Consequently, we do not explicitly model many physical phenomena, including diffuse reflection, diffraction, transmission, refraction, and higher-order specular reflections. Modeling all of these effects would increase our model's computational footprint, impeding the iterative process of fitting to a real scene. Instead, we approximate these effects as spatially uniform, with some theoretical justifications. As the reflection order increases, the number of reflection paths grows exponentially, making individual reflections less distinguishable. This comprises a sound field that, in real rooms, is approximately uniform and isotropic [25, 35, 36]. Diffuse reflections in particular can contribute to the uniformity of the sound field [46]. We approximate the total effect of high-order specular reflections, diffuse reflections and other effects as uniform, modeling them with a spatially-invariant residual signal r.

Overall Formula. Given respective source and listener locations S_{xyz} and L_{xyz}, we render the early-stage RIR by summing the contributions from all reflection paths, then convolve the result with the source's impulse response IR_s.

$$\text{RIR}(S_{xyz}, L_{xyz}) = \gamma \left[IR_s \circledast \sum_{p \in P} K(d_p, S_p, t_p) \right] + (1-\gamma) r \tag{5}$$

In this formula, $\circledast$ denotes convolution, and P is the set of all paths between the source and listener locations. As r is intended to capture higher-order reflections, its effects are likely to become more dominant later in the impulse response, whereas the traced paths are intended to characterize the early-stage reflections. For this reason, we fit 16 points on a temporal spline γ that interpolates a relative weighting between the contributions of the late-stage residual and those of explicitly computed reflection paths.

(a) Classroom (b) Dampened Room (c) Hallway (d) Complex Room

Figure 2. Photos of each room used for the DIFFRIR Dataset, each shown in its base configuration.

3.2.4 Fitting and Inference

We estimate the parameters of each acoustic model in the environment in an iterative analysis-by-synthesis process. Inspired by [14] and [17], we optimize according to a multi-scale log-spectral loss comparing rendered RIR $\hat{W}$ with the ground-truth RIR W measured at the same location. The specific loss formulation is in Eq. 6 in Appendix E.

For inference, we simply compute Equation 5 for a point at a novel location, computing all the specular paths below the maximum order between the source and the novel location, etc., and using the parameters we determined from the analysis-by-synthesis process.

Binauralization. We train our model on single-channel RIRs recorded using omnidirectional microphones. However, immersive spatial audio requires binauralization - the process of converting single-channel audio into left and right channels, in a way that mimics human perception. The shape of the head, the acoustic shadow it casts, and the differences in time-of-arrival between the left and right ears all result in distinct perceptual cues that help place the listener in the scene [20, 48]. These effects are typically modeled by head-related impulse responses (HRIRs). There is a different HRIR for each incoming audio direction. To render binaural audio, the incoming audio from each reflection path is convolved with an HRIR sampled from the SADIE II dataset [4] corresponding to its incoming direction. This allows our model to approximate perceptually accurate binaural audio, which captures the effects of the human head, with merely monaural supervision.

4. The DIFFRIR Dataset

To evaluate methods of rendering and interpolating RIRs, we collect a novel dataset of real monoaural and binaural RIRs and music data in four different rooms, as illustrated in Figure 2. Table 1 further summarizes the dimensions and reverberation time measurements of each room. In particular, we choose the following rooms to represent a wide range of room layouts, sizes, geometric complexities, and reverberation effects:

1. **Classroom.** A standard classroom with 13 rectangular tables combined into three groups, a chalkboard, two whiteboards, drywall walls, a carpeted floor, office tile ceiling, and three doors. There is ventilation noise.
2. **Dampened Room.** A semi-anechoic chamber with a carpeted floor, all four walls covered with jagged acoustic foam wedges, and specialty acoustic tile ceiling.
3. **Hallway.** A narrow, highly reverberant hallway, with two wooden doors, a tile floor, and drywall ceiling and walls.
4. **Complex Room.** A room with an irregular shape that resembles a pentagonal prism. Portions of the side wall and ceiling are covered with acoustic panels. There are three pillars in the middle of the room, one slanted diagonally. A portion of the rear wall is glass which is internally covered with paper posters. There are 7 tables, one of which is in a figure-eight shape. There are exposed air ducts, six hanging lights, water pipes, monitors, and chairs, as well as various large objects, such as a shelf. There is significant ventilation noise.

To collect audio recordings, we place a QSC K8.2 Loudspeaker in a particular location and orientation in the room and play sine sweeps to measure real RIRs in several hundred precisely-measured listener locations using a custom-built microphone array. In addition, we play and record several 10-second music clips selected from the Free Music Archive dataset [16] from the same listener and speaker locations. The music and RIRs are recorded using multiple time-synchronized Dayton Audio EMM6 omnidirectional microphones, as well as a 3Dio FS XLR microphone, which features ear-shaped silicone microphones to model human hearing and captures binaural audio.

Additional Configurations. We also collect additional subdatasets in some rooms where we slightly modify each room configuration. In each such subdataset, we vary the location and/or orientation of the speaker, or the presence and location of standalone whiteboard panels in the room. We use these additional configurations to evaluate zero-shot virtual speaker rotation and translation, and panel insertion and relocation. We include these evaluations and details on these configurations in Appendix C. While previous RIR datasets include varying room configurations [21, 33, 47]

Room	Size (m)	RT60 (s)	# of Points
Classroom	$7.1 \times 7.9 \times 2.7$	0.69	630
Dampened	$4.9 \times 5.2 \times 2.7$	0.14	768
Hallway	$1.5 \times 18.1 \times 2.8$	1.41	936
Complex	$8.4 \times 13.0 \times 6.1$	0.78	672

Table 1. Characteristics of each room and corresponding subdataset. The last column is the number of distinct microphone-speaker location pairs for which both RIRs and music are recorded, across all configurations. RT60 reverberation times are each room's average across frequencies and sub-configurations. For the Complex room, the size of its bounding box is reported.

the DIFFRIR Dataset is the first to our knowledge that also includes monoaural and binaural music recordings.

5. Experiments

For each room in our collected dataset, we evaluate our performance on the tasks of rendering both omnidirectional RIRs and music at unseen listener locations. In each room configuration, we select 12 omnidirectional RIRs to train our model. We then use our model to render RIRs at unseen locations in the test set, and compare our rendered RIRs to the ground-truth RIRs using metrics we detail in Section 5.1. To simulate music playing in the room, we convolve our rendered RIRs with five different source music files, and compare the result to real recordings of the same music files being played in the room, across the same metrics.

Baselines. We compare our method with nearest neighbor (NN) and linear interpolation baselines, which are widely used to interpolate RIRs [11, 30, 40]. We also compare with Deep Impulse Response (DeepIR) [40] and Neural Acoustic Fields (NAF) [30], which are both deep-neural-network-based (DNN-based) frameworks. DeepIR predicts the monaural RIR at novel locations based only on the location's coordinates, while NAF uses the location combined with local geometric features to estimate the RIR. In addition, NAF was originally designed for binaural rendering. Thus, we modify NAF to output monaural audio for the monaural RIR estimation task. We also compare our method with Implicit Neural Representation for Audio Scenes (INRAS) [45], which uses a combination of DNNs to more explicitly model specular and diffuse reflections at a subset of points in a scene's 3D mesh.

Additional details on baselines and any necessary adjustments we made to them are included in Appendix F.

5.1. Results

Metrics. We compare rendered audio to ground-truth audio using two metrics:

1. **Multiscale Log-Spectral L1 (Mag)**. A comparison of rendered and GT waveforms in time-frequency domain at multiple temporal and frequency resolutions [14, 17].
2. **Envelope Distance (ENV)**. The L1 distance between the log-energy envelopes of the ground-truth and rendered waveforms. Energy decay envelopes are used to extract the decay curve of the RIR, which characterizes the room's reverberant qualities [15]. We compute the signal's energy envelope by taking the envelope of the squared signal [7]. Satoh et al. [42] directly use this log-energy (squared) envelope of an RIR to measure the room's RT60 reverberation time, which is a common way of characterizing the room's acoustics [27].

Analysis. Our results for the base monaural prediction task are shown in Table 2. For the monaural prediction task, our model significantly outperforms all baselines on our metrics, across all rooms. Results for the binaural prediction task are shown in Appendix D.1.

5.2. Interpretability

We show the physically interpretable parameters our model learns for the source's directivity and reflection coefficients.

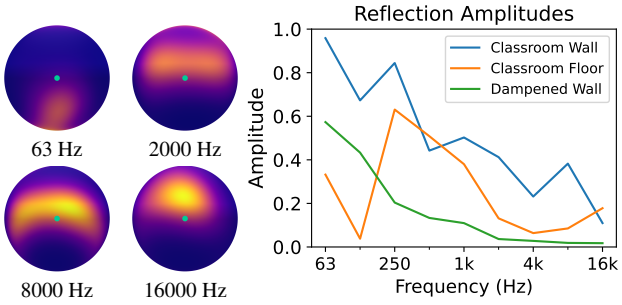


Figure 3. Visualization of our model's learned parameters. The left images show sample spherical heatmaps that our model fits to the speaker's directivity pattern when trained on 12 points from the Classroom subdataset. The green dot indicates the direction the speaker is facing, and the yellow regions indicate higher volume. The right image shows reflection amplitude responses that our model learns for various surfaces.

Directivity Maps. The left side of Figure 3 shows the source directivity heatmaps at various frequencies, learned from 12 training points in the Classroom subdataset. The area near the front of the speaker emits the loudest sound across most frequencies, as expected. The figures also confirm that higher frequencies are more directionally emitted than lower ones, evident in the narrowing yellow directivity "beam" with increasing frequency. Additionally, the fact that higher frequencies are typically emitted by the loudspeaker's tweeter at the top front of the speaker, is reflected in our heatmaps, where the yellow regions appear above the speaker's center for higher frequencies.

Reflection Amplitude Responses. The right side of Figure 3 shows the specular reflection amplitude responses that

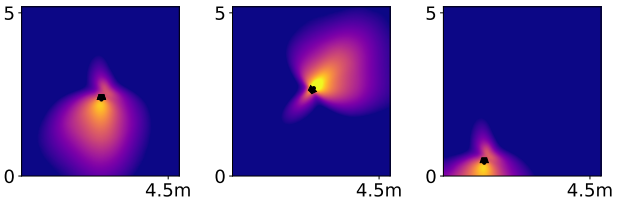
	Classroom				Dampened Room				Hallway				Complex Room			
	RIR		Music		RIR		Music		RIR		Music		RIR		Music	
	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV
NN	5.99	1.10	2.95	1.42	1.36	0.61	1.99	1.36	10.14	3.04	2.62	1.32	5.52	0.99	2.39	1.42
Linear	6.44	1.52	3.34	1.82	1.55	0.652	2.43	1.66	11.63	4.49	3.11	1.75	6.03	1.43	2.74	1.74
DeepIR	9.23	2.81	3.15	1.65	3.09	3.41	3.39	2.22	15.71	10.34	2.97	1.47	8.08	2.80	2.62	1.65
NAF	6.36	1.38	3.32	1.75	2.00	0.73	3.38	1.54	12.26	3.82	3.13	1.46	6.10	1.31	2.87	1.71
INRAS	9.99	4.52	4.45	1.75	4.20	2.48	6.22	5.35	14.52	9.19	3.70	1.58	9.02	2.58	3.61	1.66
DIFFRIR (ours)	**5.22**	**0.94**	**2.71**	**1.36**	**1.21**	**0.56**	**1.59**	**1.19**	**9.13**	**2.95**	**2.59**	**1.25**	**4.86**	**0.92**	**2.25**	**1.41**

Table 2. Experimental results on the task of predicting monaural RIRs and music at an unseen point. Lower is better for all metrics. Errors for RIRs are multiplied by 10.

	Classroom				Dampened Room				Hallway				Complex Room			
	RIR		Music		RIR		Music		RIR		Music		RIR		Music	
	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV
DIFFRIR	**5.22**	**0.94**	**2.71**	**1.36**	**1.21**	**0.56**	**1.59**	**1.19**	**9.13**	**2.95**	**2.59**	**1.25**	**4.86**	**0.92**	**2.25**	**1.41**
w/o Directivity Pattern	5.47	0.97	3.02	1.49	1.64	0.63	3.02	1.54	9.98	3.09	2.98	1.34	5.13	0.94	2.45	1.46
w/o Source IR	5.39	0.99	2.79	1.48	1.36	0.63	1.73	1.45	9.38	3.04	2.76	1.38	5.07	0.96	2.38	1.49
w/o Residual Component	6.90	1.37	3.07	1.40	1.37	0.61	1.77	1.38	15.49	4.80	2.81	1.27	6.24	1.30	2.46	1.47

Table 3. Ablation results. In each row, the ablated parameter is frozen to its initial value during training, i.e., the Source IR is assumed to be an ideal impulse, the Directivity Pattern is assumed to be uniform at all frequencies, and the Residual Component is assumed to be zero.

our model fits to some surfaces in the Classroom and Dampened Room. Our model correctly infers that the carpeted floor seems to be more absorptive than the wall, which consists of more rigid and smooth materials. The wall in the Dampened Room is even more absorptive, as our model predicts nearly no reflection above 2 kHz.



(a) Base Model (b) Virtual Rotation (c) Virtual Translation

Figure 4. RIR loudness heatmaps generated from DIFFRIR trained on 12 points in the Dampened Room's base subdataset.

Virtual Rotation and Translation. Since our model learns physically interpretable parameters, we can simulate changes to the room layout that are unseen in the training data. In Figure 4, we train our model on the Dampened subdataset, and use it to simulate virtual speaker rotation and translation. We visualize these changes by plotting RIR loudness heatmaps. Since the DIFFRIR Dataset also includes real data where the speaker is rotated or translated, we include quantitative evaluations on virtual speaker rotation and translation in the Appendix C.3, as well as evaluations on virtual panel insertion and relocation.

5.3. Ablation Study

We ablate three major components of our model (the residual, modeling the source's directivity, and modeling the source's impulse response) to determine their individual contributions. Table 3 shows our results. The results suggest that these components are all necessary for effectively rendering accurate RIRs at novel locations. More ablations experiments are in Appendix D.4.

5.4. Additional Experiments and Visualizations.

Along with additional RIR loudness maps, Appendix B.2 shows that our model can reconstruct the modal structure of the soundfield at a low frequency. In Appendix D.2, we show that our model trained on 6 points outperforms all baselines trained on 100 points. Appendix D shows that our model is robust to geometric distortions and experiments with modeling the effects of transmission.

6. Conclusions

We presented DIFFRIR, a differentiable RIR renderer capable of accurately rendering the room's acoustic impulse response at new locations, given a small set of microphone recordings and the room geometry. Future work could focus on modeling a room's acoustics implicitly by recording natural audio, thus obviating the need to measure RIRs.

Acknowledgments. The work is in part supported by NSF CCRI #2120095, RI #2211258, RI #2338203, ONR MURI N00014-22-1-2740, Adobe, Amazon, and Sony.

References

[1] Byeongjoo Ahn, Karren Yang, Brian Hamilton, Jonathan Sheaffer, Anurag Ranjan, Miguel Sarabia, Oncel Tuzel, and Jen-Hao Rick Chang. Novel-view acoustic synthesis from 3d reconstructed rooms, 2023. 2

[2] Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979. 5

[3] Adil Alpkocak and Kemal Sis. Computing impulse response of room acoustics using the ray-tracing method in time domain. *Archives of Acoustics*, 35, 2010. 2

[4] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney. A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database. *Applied Sciences*, 8(11):2029, 2018. 6

[5] Shai Avidan and Amnon Shashua. Novel view synthesis in tensor space. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1034–1040. IEEE, 1997. 1

[6] Alexis Benichoux, Laurent Simon, Emmanuel Vincent, and Remi Gribonval. Convex regularizations for the simultaneous recording of room impulse responses. *Signal Processing, IEEE Transactions on*, 62:1976–1986, 2014. 4

[7] Boualem Boashash. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic press, 2015. 7

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 3

[9] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 2

[10] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022. 3

[11] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *CVPR*, pages 6409–6419, 2023. 2, 7

[12] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[13] Mandar Chitre. Differentiable ocean acoustic propagation modeling. In *OCEANS 2023-Limerick*, pages 1–8. IEEE, 2023. 3

[14] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. Diffimpact: Differentiable rendering and identification of impact sounds. In *Conference on Robot Learning*, pages 662–673. PMLR, 2022. 3, 6, 7

[15] Simona De Cesaris, Dario D'Orazio, Federica Morandi, and Massimo Garai. Extraction of the envelope from impulse responses using pre-processed energy detection for early decay estimation. *The Journal of the Acoustical Society of America*, 138(4):2513–2523, 2015. 7

[16] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016. 6

[17] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020. 3, 6, 7

[18] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 3

[19] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Visually-guided audio spatialization in video with geometry-aware multi-task learning. *International Journal of Computer Vision*, pages 1–15, 2023. 2

[20] Michele Geronazzo, Erik Sikström, Jari Kleimola, Federico Avanzini, Amalia De Götzen, and Stefania Serafin. The impact of an accurate vertical localization with hrtfs on short explorations of immersive virtual reality scenarios. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 90–97. IEEE, 2018. 6

[21] Georg Götz, Sebastian J Schlecht, and Ville Pulkki. A dataset of higher-order ambisonic room impulse responses and 3d models measured in a room with varying furniture. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–8. IEEE, 2021. 6

[22] Brian Hamilton. Pffdtd software, 2021. https://github.com/bsxfun/pffdtd. 3

[23] D. P. Hardin, T. J. Michaels, and E. B. Saff. A comparison of popular point configurations on $\mathbb{S}^2$, 2016. 4

[24] Laszlo Hars. Frequency response compensation with dsp. *Signal Processing Magazine, IEEE*, 20:91– 95, 2003. 4

[25] Cheol-Ho Jeong. Diffuse sound field: challenges and misconceptions. *Proceedings of 45th International Congress and Exposition on Noise Control Engineering*, pages 1015–1021, 2016. 5

[26] V.D. Landon. A study of the characteristics of noise. *Proceedings of the Institute of Radio Engineers*, 24(11):1514–1521, 1936. 4

[27] Eric A. Lehmann and Anders M. Johansson. Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustical Society of America*, 124(1):269–277, 2008. 7

[28] Yan Li, Peter F. Driessen, George Tzanetakis, and Steve Bellamy. Spatial sound rendering using measured room impulse responses. In *2006 IEEE International Symposium on Signal Processing and Information Technology*, pages 432–437, 2006. 2

[29] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis, 2023. 2

[30] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022. 1, 2, 7

[31] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *Advances in Neural Information Processing Systems*, 35:2522–2536, 2022. 2

[32] J. Gregory McDaniel and Cory L. Clarke. Interpretation and identification of minimum phase reflection coefficients. *The Journal of the Acoustical Society of America*, 110(6):3003–3010, 2001. 5

[33] Thomas McKenzie, Leo McCormack, and Christoph Hold. Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis. *arXiv preprint arXiv:2111.11882*, 2021. 6

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[35] Mélanie Nolan, Marco Berzborn, and Efren Fernandez-Grande. Isotropy in decaying reverberant sound fieldsa). *The Journal of the Acoustical Society of America*, 148(2):1077–1088, 2020. 5

[36] Beth Paxton. Room acoustics, sixth ed., heinrich kuttruff. crc press (2017). isbn: 978-1-4822-6043-4. *Applied Acoustics*, 126:90–91, 2017. 5

[37] Christoph Pörschmann and Johannes M Arend. Analyzing the directivity patterns of human speakers. *Proceedings of the 46th DAGA*, pages 16–19, 2020. 3

[38] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition. In *Proc. Interspeech 2021*, pages 286–290, 2021. 1, 2

[39] Anton Ratnarajah, Zhenyu Tang, Rohith Aralikatti, and Dinesh Manocha. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 924–933, 2022. 3

[40] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. In *ICASSP*, pages 3209–3213. IEEE, 2022. 1, 2, 7

[41] Jens Holger Rindel. Modal energy analysis of nearly rectangular rooms at low frequencies. *Acta Acustica united with Acustica*, 101(6):1211–1221, 2015. 5

[42] Furniaki Satoh, Yoshito Hidaka, and Hideki Tachibana. Reverberation time directly obtained from squared impulse response envelope. In *Proc. Int. Congr. Acoust*, pages 2755–2756, 1998. 7

[43] Julius O. Smith. *Physical Audio Signal Processing*. https://ccrma.stanford.edu/~jos/pasp/, accessed 2023. online book, 2010 edition. 5

[44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[45] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 7

[46] Chiara Visentin, Matteo Pellegatti, and Nicola Prodi. Effect of a single lateral diffuse reflection on spatial percepts and speech intelligibility. *The Journal of the Acoustical Society of America*, 148(1):122–140, 2020. 5

[47] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: A dataset for finding humans using room acoustics. In *Advances in Neural Informaion Processing Systems*, 2023. 6

[48] Shu-Nung Yao. Headphone-based immersive audio for virtual reality headsets. *IEEE Transactions on Consumer Electronics*, 63(3):300–308, 2017. 6

[49] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[50] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 286–301. Springer, 2016. 1