Hearing Anywhere in Any Environment

Xiulong Liu [1*] Anurag Kumar[2] Paul Calamia[2] Sebastià V. Amengual[2] Calvin Murdock[2]
Ishwarya Ananthabhotla[2] Philip Robinson[2] Eli Shlizerman[1] Vamsi Krishna Ithapu[2] Ruohan Gao[3]

[1]University of Washington [2]Meta [3]University of Maryland, College Park

Abstract

In mixed reality applications, a realistic acoustic experience in spatial environments is as crucial as the visual experience for achieving true immersion. Despite recent advances in neural approaches for Room Impulse Response (RIR) estimation, most existing methods are limited to the single environment on which they are trained, lacking the ability to generalize to new rooms with different geometries and surface materials. We aim to develop a unified model capable of reconstructing the spatial acoustic experience of any environment with minimum additional measurements. To this end, we present xRIR*, a framework for cross-room RIR prediction. The core of our generalizable approach lies in combining a geometric feature extractor, which captures spatial context from panorama depth images, with a RIR encoder that extracts detailed acoustic features from only a few reference RIR samples. To evaluate our method, we introduce* AcousticRooms*, a new dataset featuring high-fidelity simulation of over 300,000 RIRs from 260 rooms. Experiments show that our method strongly outperforms a series of baselines. Furthermore, we successfully perform sim-to-real transfer by evaluating our model on four real-world environments, demonstrating the generalizability of our approach and the realism of our dataset.*

1. Introduction

Each environment echoes its own story, creating a distinct auditory experience. Imagine walking through a museum where each room's unique acoustic character brings exhibits to life, immersing you in stories through tailored soundscapes. To recreate such realistic auditory experiences across different spaces, models must seamlessly and easily adapt from one environment to another. This *cross-room* adaptability is essential for applications like virtual reality and immersive media, where authentic acoustics transform and enhance how we experience sound anywhere in any environment of interest.

[*]Work done during the internship at Meta.

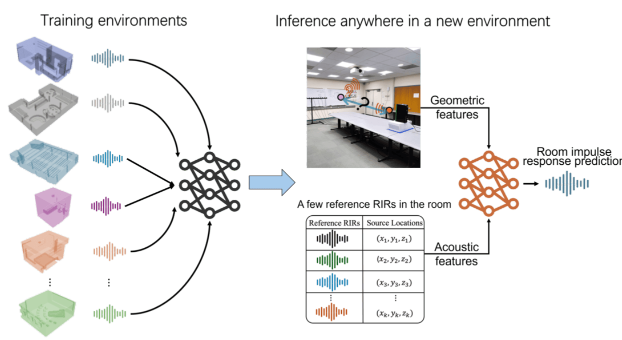

Figure 1. Our xRIR framework can predict accurate room impulse responses (RIR) of any new environment, by integrating the geometric prior learned from a large simulated RIR dataset of diverse training environments and the nuanced acoustic profile extracted from a few reference RIR measurements in the new environment.

To model how sound interacts with an environment, a room impulse responses (RIR) is often measured to capture how a perfect impulse emitted from the source location reflects, is absorbed, diffuses, and gets received at the microphone location, all according to the room's unique geometry and surface materials. Traditionally, to fully capture the immersive acoustic field in an environment, hundreds of RIR measurements are gathered by positioning speakers and microphones densely throughout the space [31, 37, 42, 49]. However, this process is labor-intensive and costly, especially when scaling across diverse real-world environments.

Recent deep learning approaches [10, 28, 31, 49] leverage implicit neural networks to "compress" these dense RIR measurements into a single model that can be queried for RIR at any source-listener pair in a specific room. However, training these models still requires a large amount of densely sampled RIRs for each room, as they are designed to overfit to a *single room*'s specific geometry and material properties. When applied to a new environment, these models must be re-trained with a similarly dense dataset, limiting their practicality for scalable cross-room applications.

Our goal is to develop a model that can handle variations in room geometry and surface material properties, which are key factors in shaping each room's unique acoustic profile. Achieving this goal presents several challenges. First, the model should utilize an easily obtainable and standard visual representation of the environment to extract geometric properties from any room. Second, given the diverse and nuanced acoustic properties of different surface materials, we need a way to quickly capture essential cues about a room's detailed acoustic characteristics (*e.g.*, energy decay and reverberation patterns). Third, to ensure generalization, we need a large-scale and high-fidelity RIR dataset that encompasses a wide range of room environments with varied acoustic materials and geometries, enabling the pre-training of a cross-room feature extractor.

To address these challenges, we introduce XRIR, a generalizable model for cross-room RIR prediction, along with ACOUSTICROOMS, a large dataset comprising over 300,000 realistic RIRs simulated from 260 rooms, specifically curated for this task. Our model features three key components: i) a Geometric Feature Extractor, which utilizes a vision transformer to process panorama depth images from the receiver's perspective, capturing the spatial relationships between the source and receiver positions within the room; ii) a Reference RIR Encoder, which extracts spatio-temporal features from a few reference RIRs, capturing the unique energy decay and reverberation characteristics associated with room materials; and iii) a Fusion and Weighting Module, which predicts the target RIR through a weighted combination of the reference RIRs. By integrating complementary geometric and acoustic features, our model effectively approximates both structural and material properties of any room, enabling precise RIR predictions not only at new locations in the training environments, but also in any new environment of interest.

We evaluate our model's performance in both seen and unseen environments from ACOUSTICROOMS, demonstrating its capability to predict RIRs at new locations within known rooms and also effectively generalize to entirely new environments. Our method consistently achieves state-of-the-art results across these scenarios, outperforming several strong baselines and prior methods. In addition, to assess the model's real-world applicability, we successfully perform sim-to-real transfer by deploying our model on a dataset comprising four real-world environments. This demonstrates the effectiveness of our framework and also the realism of RIR simulations in our dataset.

In sum, our main contributions are as follows:

- We propose XRIR, a cross-room generalizable framework that predicts accurate RIRs for any seen and unseen environment, strongly outperforming prior methods.
- We introduce ACOUSTICROOMS, a large-scale dataset tailored for this task, comprising over 300,000 high-fidelity RIRs simulated from 260 diverse rooms.
- Apart from superior performance on simulated rooms, we also successfully deploy our model in four real-world environments, showcasing effective sim-to-real transfer.

2. Related Work

Learning-Based RIR Prediction. Early machine learning methods for room impulse response (RIR) prediction, such as Image2Reverb [46] and Fast-RIR [39], utilize a generative approach conditioned on semantic information like RGB images of the environment, source and listener locations, and T60 values. While these approaches produce plausible RIRs aligned with scene semantics and basic acoustic constraints, they struggle to reproduce accurate RIRs at arbitrary locations within the target scene. Recent advances on implicit neural representations [33, 47] have inspired a series of works that approximate a function mapping spatial coordinates of source and listener locations to RIRs [31, 38, 42, 49]. Some methods [1, 2, 8, 10, 24, 27, 40] also condition on room geometry and material properties of the visual environment. By explicitly modeling the 3D scene, these methods can render precise RIRs at novel locations within the same environment they are trained on. However, they lack the ability to generalize to new environments, which is the focus of our work.

Closest to our work are Diff-RIR [55] and Few-Shot RIR [32], both of which also address cross-room RIR generalization. Diff-RIR learns material coefficients through a differentiable rendering framework based on planar room geometry and a few RIR measurements, enabling RIR rendering at any location using the image source method. However, it requires training a separate model for each room, making it computationally intensive and less scalable for large or complex environments. Few-Shot RIR leverages a limited number of RIR measurements and RGB-D observations to predict RIRs at new locations by integrating features from pose, RGB-D data, and binaural echoes. However, it does not effectively utilize geometric information as our method and is only trained and tested in simulation. We compare our approach with both methods in experiments.

Room Acoustics Simulation. Room acoustic simulation is crucial for applications in AR/VR [25], architectural design [35], and far-field speech recognition [22, 36], serving as a bridge between simulated and real-world acoustics. Approaches can be generally categorized as wave-based, geometric-based, or hybrid. Wave-based methods [13] accurately capture low-frequency phenomena but are computationally intensive. Geometric-based methods, such as [3, 6, 43, 44], efficiently trace sound rays but lack accuracy at low frequencies. Hybrid methods [51] combine both approaches to balance accuracy and efficiency. We utilize the hybrid approach [41] from the Treble simulation platform to simulate high-fidelity RIRs for our cross-room RIR prediction task, optimizing both accuracy and computational demands.

Audio-Visual Learning of Room Acoustics. Both vision and audio provide significant spatial information that reveals room properties. Prior methods have combined both modalities for a series of audio-visual learning tasks related to room acoustics, including audio spatialization using visual spatial cues from the environment [16, 19, 20, 26, 34], audio-visual navigation in environments with varying room acoustic properties [3, 4, 15, 18, 29], learning image features, scene structures, or human locations from echoes [12, 17], ambient sound [9], or music [54] in the room, and using RGB images or videos of a target environment to guide sound transfer that aligns with the space's acoustics [5, 7, 11, 48, 52]. Our work also integrates both visual and audio information, but we address a different challenging task that aims to infer accurate room acoustics in any new environment.

3. Our Approach

3.1. Problem Formulation

We tackle the *cross-room* room impulse response (RIR) prediction task, which aims to predict single-channel (omnidirectional) RIRs for any source-receiver pair across diverse room environments, including those *unseen* during training. We aim to develop a generalizable model that can accurately predict RIRs in any environment without labor-intensive data collections or training a separate model for each room. Our model (detailed in Sec. 3.2) achieves this by utilizing only minimal additional measurements from the new room, such as only a few panorama depth images and reference RIR measurements, to quickly adapt to new acoustic environments with minimal effort, thereby facilitating generalization to previously unseen environments. Next, we formally define the cross-room RIR prediction task by outlining its data, inputs, and the modeling objective.

Data. Let $R = \{R_1, R_2, \dots, R_M\}$ represent a dataset of M rooms, split into a training set, $R_{\text{train}} \subset R$, and a test set, $R_{\text{test}} = R \setminus R_{\text{train}}$. Each room R_m includes a set of receiver locations, denoted $L_m = \{P_r^{(m,1)}, P_r^{(m,2)}, \dots, P_r^{(m,N_m)}\}$, where N_m is the number of receivers in the room. For each receiver $P_r^{(m,i)}$ in room R_m, RIRs are measured at various source locations $P_s^{(m,i,j)}$, resulting in measurements $A_{m,i,j}$ of the source-receiver pair $P_s^{(m,i,j)}$ and $P_r^{(m,i)}$.

Inputs. To capture the necessary observation conditions for predicting a target RIR A_t, we define an observation tuple $\mathbf{O} = (P_s, P_r, G_r)$, where P_s is the target source location, P_r is the receiver location, and G_r represents the local geometry near the receiver location P_r, e.g., room boundary points or depth maps around the receiver. Additionally, we introduce a set of K reference RIRs measured at the target receiver location P_r from various reference source locations $\mathbf{P}_{\text{ref,s}} = \{P_{\text{ref,s}}^{(1)}, P_{\text{ref,s}}^{(2)}, \dots, P_{\text{ref,s}}^{(K)}\}$. These references, denoted as $\mathbf{A}_{\text{ref}} = \{A_{\text{ref}}^{(1)}, A_{\text{ref}}^{(2)}, \dots, A_{\text{ref}}^{(K)}\}$, are crucial for capturing essential acoustic characteristics that encode nuanced information about room materials. Note that in the above formulation, while we fix the receiver location and set reference RIRs at different source locations, exchanging the receiver and source in the input yields *an equivalent alternative formulation*.

Modeling Objective. The objective of cross-room RIR prediction is to train a model F that predicts the target RIR $\hat{A}_t$ using the observation tuple $\mathbf{O}$ along with the reference RIRs $\mathbf{A}_{\text{ref}}$ and their respective source locations $\mathbf{P}_{\text{ref,s}}$: $\hat{A}_t = F(\mathbf{O}, \mathbf{A}_{\text{ref}}, \mathbf{P}_{\text{ref,s}})$. In this formulation, $\mathbf{O}$ provides the geometric and positional context, while $\mathbf{A}_{\text{ref}}$ and $\mathbf{P}_{\text{ref,s}}$ give sparse acoustic observations that help bridge the lack of explicit material properties by capturing key room acoustics characteristics.

Unlike the *single-room* RIR prediction task [31, 49], which assumes consistent geometry and material properties and fits a separate model for each scene, our cross-room formulation aims to train a single model that generalizes across multiple scenes with diverse room geometries and materials, while with the extra condition of only a few RIR measurements. Models that are designed for single-room RIR prediction task [31, 49] must be re-trained with dense data when applied to a new environment. While our method uses one unified model to predict accurate RIRs across different rooms, seen or unseen. Our formulation can also be easily adapted to the single-room RIR prediction setting by fitting dense measurements in the room as in prior work. Please see supp. for results on single-room experiments.

3.2. The xRIR Model

To solve the cross-room RIR prediction task, we propose a new architecture, xRIR, which processes not only geometry and positional features of source and receiver, but also leverages the reference RIRs to accurately predict the target RIR. As illustrated in Fig. 2, xRIR consists of three main components: i) a *Geometric Feature Extractor* (Sec. 3.2.1), which encodes the spatial relationships among the source, receiver, and room surface geometry, capturing important geometric features that shape acoustic behavior; ii) a *Reference RIR Encoder* (Sec. 3.2.2), which processes the spatiotemporal characteristics of the reference RIRs to extract features that represent their acoustic properties within the room.; and iii) a *Fusion and Weighting Module* (Sec. 3.2.3), which integrates the spatial features from the Geometric Feature Extractor with the reference RIR features from the Reference RIR Encoder, generating a set of weights to combine reference RIRs as the predicted target RIR.

3.2.1. Geometric Feature Extractor

The Geometric Feature Extractor module captures spatial relationships among the source, receiver, and room geometries, which is important for accurate acoustic modeling. It consists of two geometric sub-modules: the *Direct Path Module*, and the *Reflection Module*. These two modules emulate the process of sound propagation. The Direct

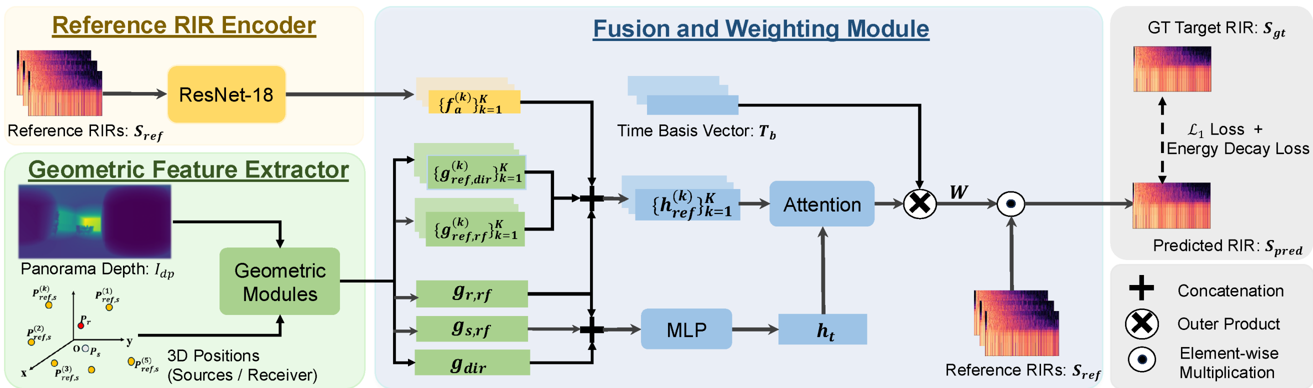

Figure 2. **System Overview of Our xRIR Model for Cross-Room RIR Prediction.** The model architecture consists of three main components: i) a *Geometric Feature Extractor*, which captures spatial relationships among the source, receiver, and room geometry; ii) a *Reference RIR Encoder*, which extracts spatiotemporal features from reference RIRs; and iii) a *Fusion and Weighting Module*, which integrates these spatial and acoustic features to predict the target RIR.

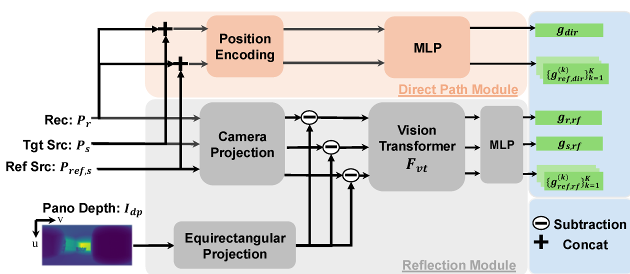

Figure 3. **Illustration of the Geometric Feature Extractor.** Rec: Receiver, Tgt Src: Target Source, Ref Src: Reference Source.

Path Module extracts the feature of the direct path between source and receiver, while the Reflection Module models the sound propagation path through the reflections from the room boundaries. A detailed overview of the Geometric Feature Extractor is illustrated in Fig. 3.

Direct Path Module. To capture the direct path between each source and the receiver, we concatenate their 3D coordinates. For the target source, we define $P_{\text{dir}} = (P_s, P_r)$, where P_s and P_r are the coordinates of the source and the receiver, respectively. For each reference source $P^{(k)}_{\text{ref,dir}}$, we define $P^{(k)}_{\text{ref,dir}} = (P^{(k)}_{\text{ref},s}, P_r)$. P_{dir} and each $P^{(k)}_{\text{ref,dir}}$ encode locations of every source-receiver pair, thereby encoding the direct path information. To extract their features, we apply sinusoidal positional encoding [53] followed by a multi-layer perceptron (MLP) to project them into high-dimensional vectors, resulting in $g_{\text{dir}} \in \mathbb{R}^{1\times C_d}$ for the target source-receiver pair and $g^{(k)}_{\text{ref,dir}} \in \mathbb{R}^{1\times C_d}$ for each reference source-receiver pair.

Reflection Module. Inspired by INRAS [49], we also model the reflection paths between source and receiver via the room boundary. But differently, instead of using a fixed set of bounce points per room, we propose to use a panorama depth map at the receiver's location as a proxy for local room geometry to unify the representation across different rooms.

Given the panorama depth map $I_{dp} \in \mathbb{R}^{H\times W\times 3}$ centered at receiver's viewpoint, we first project I_{dp} to a 3D coordinate map $I_{\text{coord}} = F_{\text{rect}}(I_{dp})$ via an equi-rectangular projection transformation F_{rect}. Each pixel in I_{coord} represents the 3D coordinate of a visible boundary point in the room from the receiver's view. Each source in the room has chance to reflect through these points until finally reaching the receiver. To model such interactions, we create a set of reflections-based maps by subtracting the 3D coordinate map from the source and receiver positions.

To perform the subtraction, it is necessary to unify the coordinates between the 3D coordinate map and the source / receiver positions. We achieve this by projecting the world coordinates of the sources and the receiver into camera coordinates at the receiver's position, resulting in the same coordinate system as the 3D coordinate map. We obtain the target source position as $P_{\text{rel},s} = R(P_s - P_r)$ and each reference source location as $P^{(k)}_{\text{rel, ref}} = R(P^{(k)}_{\text{ref},s} - P_r)$, where R is the world-to-camera transformation matrix. Then we create reflection-based maps $I_{\text{rf},s}$ for the target source, $I^{(k)}_{\text{rf, ref}}$ for reference sources, as well as $I_{r,\text{rf}}$ for the receiver by performing subtractions: $I_{s,\text{rf}} = P_{\text{rel},s} - I_{\text{coord}}$, $I^{(k)}_{\text{ref,rf}} = P^{(k)}_{\text{rel, ref}} - I_{\text{coord}}$ and $I_{r,\text{rf}} = I_{\text{coord}} - \mathbf{0}$, where $\mathbf{0}$ means the origin, where the receiver is located.

These reflection-based maps encode the dense interaction between room geometry and the sources / receiver. To further extract features, we utilize a vision transformer module [14] F_{vt} that partitions each reflection-based map into patches, aggregates local features, and builds spatial dependencies among patches. This results in compact patch-level geometry representations: $g'_{\text{r,rf}} = F_{\text{vt}}(I_{r,\text{rf}})$, $g'_{\text{s,rf}} = F_{\text{vt}}(I_{s,\text{rf}})$ and $\{g'^{(k)}_{\text{ref,rf}} = F_{\text{vt}}(I^{(k)}_{\text{ref,rf}})\}_{k=1}^{K}$, where each feature map has dimensionality $N_p \times C_p$. Finally, we apply a MLP layer to project the patch dimension N_p to 1, resulting in $g_{\text{r,rf}}$, $g_{\text{s,rf}}$, and $\{g^{(k)}_{\text{ref,rf}}\}_{k=1}^{K}$, respectively.

3.2.2. Reference RIR Encoder

To capture acoustic features related to energy decay and reverberation patterns within the room, we leverage reference RIRs as proxies for the acoustic characteristics at various source locations relative to the receiver. To encode these acoustic features, we first compute the log-magnitude spectrogram of each reference RIR using the Short-Time Fourier Transform (STFT): $\mathbf{S}_{\text{ref,k}} = \log(\|\text{STFT}(A_{\text{ref,k}})\|)$, where $\mathbf{S}_{\text{ref,k}} \in \mathbb{R}^{F\times T}$. To extract robust acoustic features, we implement the Reference RIR Encoder using ResNet-18, and use the mean pooled features $f_a^{(k)} \in \mathbb{R}^d$ from the last layer to encode each reference RIR.

3.2.3. Fusion and Weighting Module

The Fusion and Weighting Module integrates the outputs from the Geometric Feature Extractor and the Reference RIR Encoder to generate the target RIR prediction. This module combines geometric and acoustic features for reference sources as well as the geometric features of target source, finally computing the weights that are applied to reference RIRs.

Fusion of Geometric and Acoustics Features. For each reference source, we combine the geometric feature $g^{(k)}_{\text{ref,dir}}$, $g^{(k)}_{\text{ref,rf}}$, $g_{\text{r,rf}}$ and the acoustic feature $f_a^{(k)}$ by concatenating them along the feature dimension, resulting in: $\mathbf{h}^{(\mathbf{k})}_{\mathbf{ref}} = \text{Concat}(g^{(k)}_{\text{ref,dir}}, g^{(k)}_{\text{ref,rf}}, g_{\text{r,rf}}, f_a^{(k)})$.

Similarly, for the target source, we combine the geometric feature g_{dir}, $g_{\text{s,rf}}$ and $g_{\text{r,rf}}$ via concatenation, yielding: $\mathbf{h}'_{\mathbf{t}} = \text{Concat}(g_{\text{dir}}, g_{\text{s,rf}}, g_{\text{r,rf}})$. We then project the fused feature $\mathbf{h}'_{\mathbf{t}}$ to $\mathbf{h}_{\mathbf{t}}$ through a MLP to make the feature dimension the same as $\mathbf{h}^{(k)}_{\mathbf{ref}}$.

To align the target and reference features, we compute the attention between the target fused vector $\mathbf{h}_t$ and each reference fused vector $\mathbf{h}^{(k)}_{\mathbf{ref}}$. Specifically, given the reference fused features $\mathbf{H}_{\mathbf{ref}} = \{\mathbf{h}^{(k)}_{\mathbf{ref}}\}_{k=1}^{K}$ and the target fused vector $\mathbf{h}_t$, the attention output $\mathbf{Z}$ is computed as:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{H}_{\mathbf{ref}} \cdot \mathbf{h}_{\mathbf{t}}^{\mathbf{T}}}{\sqrt{C}}\right) \odot \mathbf{H}_{\mathbf{ref}},$$

where $\cdot$ and $\odot$ denote matrix multiplication and element-wise multiplication with broadcasting respectively, and $\mathbf{H}_{\mathbf{ref}} \in \mathbb{R}^{K\times C}$, $\mathbf{h}_{\mathbf{t}} \in \mathbb{R}^{1\times C}$, $\mathbf{Z} \in \mathbb{R}^{K\times C}$. These attention outputs $\mathbf{Z} = \{\mathbf{z}_k\}_{k=1}^{K}$ for each reference RIR is now attended by the fused feature of the target RIR.

Time-Aligned Weighting Matrix. Given the attention outputs $\mathbf{Z} \in \mathbb{R}^{K\times C}$, we next generate a time basis vector $\mathbf{T}_b$ based on the temporal indices of the spectrogram $[0, 1, 2, \ldots, T]$. Specifically, we compute $\mathbf{T}'_{\mathbf{b}}$ using sinusoidal positional encoding [53] and then apply a MLP layer to project $\mathbf{T}'_{\mathbf{b}}$ to C, resulting in $\mathbf{T}_{\mathbf{b}} \in \mathbb{R}^{T\times C}$. We generate the time-aligned weighting matrix $\mathbf{W} \in \mathbb{R}^{K\times T}$ by computing the outer product between $\mathbf{Z}$ and $\mathbf{T}_{\mathbf{b}}$: $\mathbf{W} = \mathbf{Z} \cdot \mathbf{T}_{\mathbf{b}}^{\mathbf{T}}$. Each row of $\mathbf{W}$ corresponds to the weights applied to the log-magnitude spectrogram of each reference RIR, adapting them to match the temporal structure of the target spectrogram. This weighting matrix $\mathbf{W}$ effectively shapes each reference spectrogram to align with the characteristics of the target RIR.

Finally, we predict the target RIR's log-magnitude spectrogram $\mathbf{S}_{\text{pred}}$ via the weighted sum of the log-magnitude spectrograms of the reference RIRs: $\mathbf{S}_{\text{pred}} = \sum_{k=1}^{K} \mathbf{W}_k \odot \mathbf{S}_{\text{ref},k}$. $\mathbf{W}_k$ is the k-th row of the weight matrix $\mathbf{W}$, applied to the corresponding log-magnitude spectrogram $\mathbf{S}_{\text{ref},k}$.

3.2.4. Training and Inference

During training, we use the magnitude STFT L_1 Loss to compute the error between the magnitude spectrograms of the predicted target RIR and the ground-truth RIR: $\mathcal{L}_{\text{STFT}} = \|\exp(\mathbf{S}_{\text{pred}}) - \exp(\mathbf{S}_{\text{gt}})\|_1$. Additionally, following [32], we incorporate an energy decay loss to optimize the decay patterns of the predicted spectrogram. The energy decay loss $\mathcal{L}_{\text{ED}}$ is defined as: $\mathcal{L}_{\text{ED}} = \|\text{EDC}(\mathbf{S}_{\text{pred}}) - \text{EDC}(\mathbf{S}_{\text{gt}})\|_1$, where $\text{EDC}(\cdot)$ denotes the energy decay curve of RIR in the frequency domain. The total loss becomes $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{STFT}} + \lambda\mathcal{L}_{\text{ED}}$, where λ is a weight to balance the contribution of the energy decay loss.

During inference, we randomly samples K RIRs $\{A_{ref,k}\}_{k=1}^{k=K}$ along with corresponding source locations $\{P_{\text{ref},k}\}_{k=1}^{k=K}$ from a test room as reference inputs. The model predicts the magnitude spectrogram of a target RIR, which is then converted back to a waveform via the Griffin-Lim [21] algorithm.

4. The ACOUSTICROOMS Dataset

To our best knowledge, there are two prior datasets with a large number of rooms: SoundSpaces MP3D [3, 32] and GWA [51]. SoundSpaces MP3D comprises only 83 rooms with limited material variety (around 100 types), a fixed one-to-one mapping between semantic labels and acoustic coefficients, and a constrained 2D configuration at fixed heights. This setup restricts real-world applicability, as actual rooms often contain diverse materials with varying acoustic properties and require 3D spatial modeling to capture realistic sound propagation. For GWA, while it includes a large number of simulated RIRs from a wide variety of synthetic rooms and explicitly models wave propagation, the wave-based method it employs, PFFDTD [23], is a lower-resolution approach. This method prioritizes computational efficiency, which comes at the cost of reduced simulation accuracy.

To address the above limitations, we introduce ACOUSTICROOMS, a new large-scale, high-quality dataset of simulated RIRs specifically designed for robust generalization across diverse room geometries, sizes, and material properties. We use Treble Technology's simulation platform[1], where a more advanced wave-based solver, i.e., the Discontinuous Galerkin (DG) Method [41], is supported. Employing such techniques to simulate RIRs in our dataset is crucial for achieving cross-room generalization and sim-to-real transfer applications. ACOUSTICROOMS simulates 260 rooms across 10 categories, featuring 300K simulated RIRs from different source-receiver pairs and full 3D spatial configurations. Each room includes a randomized material assignment from a library of 332 materials across 11 categories, ensuring diversity in acoustic properties even among similar geometries. The combination of scale, material diversity, and simulation fidelity enables ACOUSTICROOMS to accurately reflect the acoustics of real-world environments.

5. Experiments

5.1. Implementation Details

In the ACOUSTICROOMS dataset, RIRs are sampled at 22,050 Hz with a maximum length of 9600 samples (0.435 s). We compute the magnitude spectrogram S with FFT size 124, window size 62, and hop size 31, yielding a shape of 63×310. Panorama depth maps of room geometry are rendered at a resolution of 256×512 from the receiver's location, and source/receiver positions are recorded as 3D coordinates (x, y, z). For XRIR, we implement a Vision Transformer block F_{vt} with 6 multi-head attention layers (8 heads, hidden size 512). Depth maps are divided into 16×32 patches, resulting in all reflection-based features such as $g_{\text{r,rf}}$ and $g_{\text{s,rf}}$ of dimension 256×512. Direct path features are calculated using sinusoidal encoding on each 3D coordinate with 20 frequency bins, and are then projected into 256-dimensional vectors via MLP. For loss calculation, we set $\lambda = 0.01$ to balance the STFT loss and the energy decay loss.

5.2. Baselines

We compare with a series of baselines as well as prior methods [32, 55]:

- **Random Across Rooms**: Randomly sample a RIR from the entire dataset as the prediction for the target RIR.
- **Random Same Room**: Randomly sample a RIR from the same room as as the prediction for the target RIR.
- **Nearest Neighbor**: Sample k-shot reference RIRs and select the RIR with the closest spatial distance to the target source as the prediction.
- **Linear Interpolation**: Linearly interpolate between k-shot reference RIRs based on the distance between each reference and the target source location.

[1] https://www.treble.tech/

- **Few-Shot RIR [32]**: Few-Shot RIR implements a transformer architecture that fuses features from separate encoders of multi-modal conditional inputs and then generates the target RIR by decoding the transformer outputs via a UNet decoder. We adapt their model to our task by replacing the binaural echos with our single-channel reference RIRs (different source and receiver locations) and using panorama depth images as inputs to the image encoder instead of egocentric RGBD images.
- **Diff-RIR [55]**: We compare with Diff-RIR in evaluation on sim-to-real transfer. The framework utilizes the few-shot, i.e., 12 reference RIRs, to train a differentiable rendering pipeline to learn acoustics parameters of the room geometry. For fair comparison, we finetune our XRIR model pre-trained on ACOUSTICROOMS on the same set of reference RIRs as Diff-RIR in each room, and then test on the same test split. Note that Diff-RIR requires training one model per each room and the training process becomes computationally infeasible for large space with complex room geometries. Therefore, we do not include it in our comparison on ACOUSTICROOMS.

In addition, for a more complete comparison with prior methods on RIR prediction, we also adapt our method for the single-room RIR prediction task and compare with prior work [31, 49]. Please see Supp. for results.

5.3. Metrics

We evaluate the energy pattern of the generated RIRs against ground-truth RIRs using three key acoustic metrics, which are strongly correlated with hearing perception and commonly used in prior work on RIR prediction [31, 49]:

- **Early Decay Time (EDT) Error:** To evaluate early reflection characteristics, we use the EDT error, which measures the time taken for the initial 5 dB decay in the energy curve.
- **Clarity (C50):** For comparing early-to-late energy ratios, we employ the clarity metric $C50$, which provides insights into the prominence of early reflections over later reverberations.
- **T60 Error:** We evaluate the accuracy of reverberation time by comparing the $T60$ value of the predicted RIR and the ground-truth RIR. We calculate $T60$ using $T20$, based on a linear fit between -5 dB and -25 dB on the logarithmic energy decay curve obtained from Schroeder Backward Integration [45].

5.4. Quantitative Results on ACOUSTICROOMS

We show cross-room RIR prediction results in both environments seen during training as well as unseen new environments. As shown in Table 1, our XRIR model significantly outperforms all baselines across all metrics (EDT error, $C50$ error, and $T60$ error). In the seen split, our XRIR model yields the lowest errors across metrics, particularly in

Method	Seen Splits			Unseen Splits		
	EDT error (s)	C50 error (dB)	T60 error (%)	EDT error (s)	C50 error (dB)	T60 error (%)
Random Across Rooms	0.290	6.831	37.35	0.313	7.802	35.15
Random Same Room	0.129	3.567	12.80	0.172	5.440	16.08
Few-shot RIR [32] (K=1)	0.157	3.957	31.42	0.130	3.225	20.10
Few-shot RIR [32] (K=4)	0.157	4.026	31.63	0.136	3.568	19.30
Few-shot RIR [32] (K=8)	0.174	4.451	32.71	0.187	4.470	21.15
Linear Interpolation (K=8)	0.094	2.421	9.76	0.121	3.090	13.73
Nearest Neighbor (K=8)	0.064	1.717	8.94	0.090	2.667	11.64
xRIR (K=1)	0.046	1.183	9.50	0.075	1.841	13.47
xRIR (K=4)	0.040	1.005	8.15	0.068	**1.335**	13.28
xRIR (K=8)	**0.038**	**0.940**	**8.13**	**0.055**	1.457	**10.53**

Table 1. **Cross-Room RIR Prediction Results for Both the Seen and Unseen Splits.** We report EDT Error (EDT) in seconds, C50 Error (C50) in dB, and T60 percentage error (T60), with lower values indicating better performance. For Few-shot RIR [32] and xRIR (Ours), we evaluate in a few-shot manner by setting the number of reference RIRs K to 1, 4, and 8.

Method	Classroom			Dampened Room			Hallway			Complex Room		
	EDT	C50	T60	EDT	C50	T60	EDT	C50	T60	EDT	C50	T60
Random Across Rooms	0.546	8.740	19.03	0.771	18.726	-	0.874	11.025	21.71	0.472	7.392	16.01
Random Same Room	0.160	3.092	3.12	0.099	6.840	-	0.308	6.461	16.61	0.218	4.566	5.66
Linear Interpolation (K=8)	0.113	2.172	4.42	0.058	4.584	-	0.088	2.127	4.55	0.124	2.848	5.17
Nearest Neighbor (K=8)	0.108	1.949	**2.71**	**0.044**	**3.278**	-	0.068	0.990	**3.02**	0.091	1.936	**2.53**
Diff-RIR [55] (K=12)	0.113	2.147	12.39	0.100	3.796	-	0.160	2.049	14.34	0.115	2.027	12.76
xRIR (K=8) (Ours)	**0.093**	**1.628**	6.25	0.044	3.302	-	**0.062**	**0.954**	3.20	**0.077**	**1.688**	4.33

Table 2. **Sim-to-Real Transfer Results in Four Real Environments from the Hearing-Anything-Anywhere Dataset [55].** We report EDT Error (EDT) in seconds, C50 Error (C50) in dB, and T60 percentage error (T60). Due to noisy measurements in the dampened room, resulting in low SNR and invalid T60 calculations on the EDC curve, we omit this metric for the dampened room.

$C50$ and $T60$. Our gains persists in the unseen split. In particular, our model with $K = 8$ reduces $T60$ error to around 10%, while other baselines exhibit much higher errors. This result highlights our model's robustness in capturing reverberation characteristics across different room configurations and the ability to generalize to unseen environments with different room acoustic properties.

The Few-Shot RIR approach from [32] does not perform well AcousticRooms. We suspect that this is due to two factors: i) their UNet decoder struggles to reconstruct high-fidelity RIRs on our data, as it relies on highly compressed fusion features; ii) their method uses binaural echoes with co-located source and receiver positions, which fundamentally differ from our setup, where reference RIRs are measured with the source and receiver at different locations. This spatial disparity likely impacts feature relevance, limiting its performance on our dataset.

5.5. Sim-to-Real Transfer to Real Environments

To evaluate whether our model can also generalize to real-world environments, we use four real rooms from the Hearing-Anything-Anywhere Dataset [55]. We compare our method against Diff-RIR [55], a physics-based differentiable RIR rendering pipeline that utilizes 12 reference RIRs per room to predict RIRs for new locations. As shown in Table 2, our model compares favorably against all baselines. In partilar, despite using only 8-shot references, our method outperforms Diff-RIR that uses 12 reference RIRs in all acoustic metrics, demonstrating its strong generalization capabilities. We observe that our method underperforms on the T60 metric compared to the Nearest Neighbor baseline across all four rooms. We suspect this is because T60, as a global metric, is more sensitive to measurement noise due to its aggregation of all acoustic interactions within the room. Our learning-based method can struggle with low SNR beyond the early parts of the waveform, as it is trained on simulation data with higher SNR than real room measurements. In contrast, EDT focuses on early reflections with high SNR, make it less noise-sensity, and C50 is similarly robust due to noise smoothing in the integration beyond the early parts. Despite this, our results demonstrate that xRIR's effectiveness in adapting from simulated rooms to real environments, successfully capturing diverse room acoustics with fewer reference RIRs than prior methods.

5.6. Qualitative Results

We present qualitative results by comparing the predicted RIRs and acoustic maps between our model xRIR and the

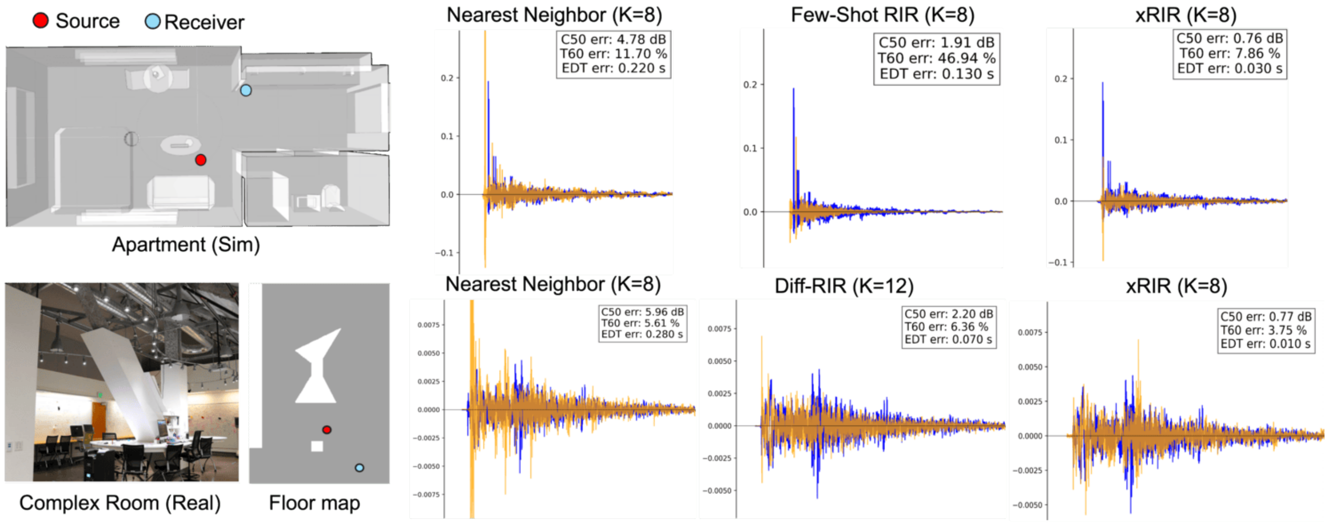

Figure 4. **Qualitative Comparisons of RIR Predictions.** We compare the performance of our method and the baselines both in simulated (top row) and real (bottom row) environments. Room geometry, sample RIR predictions, and the corresponding error metrics are included. xRIR shows more accurate RIR predictions in both settings.

Figure 5. **Qualitative Comparisons of Acoustic Map Predictions in Two Real Environments: a Hallway and a Classroom.** We visualize the acoustics maps by computing the C50 metric at dense locations in the entire room and compare with the ground-truth acoustic map. xRIR achieves C50 distributions that better matches the ground-truth.

baseline methods, in both simulated and real environments.

RIR Predictions. In Fig. 4, we visualize sample results of RIR waveforms on a simulated apartment and a real room with complex geometry. Side-by-side comparison shows that predicted RIRs from xRIR align more closely with the ground-truth RIR waveforms in the early part than baselines. This observation is consistent with the low acoustics metrics errors achieved by our method in the quantitative results shown in Table 1.

Acoustics Maps. Furthermore, we compute the RIRs at dense locations across the entire real rooms, and compute the clarity of the predicted and ground-truth RIRs to reconstruct acoustics maps according to the floor plans. As shown in Fig. 5, across dense locations in these rooms, overall xRIR achieves better C50 distribution than Diff-RIR [55] compared to the ground-truth acoustic maps, especially at moderate-to-low intensity regions. These qualitative results demonstrate the effectiveness of xRIR in accurate RIR prediction in both simulation and real-world settings.

6. Conclusion

We presented xRIR, a model designed for generalizable RIR prediction across diverse room environments. To tackle the cross-room RIR prediction task, we also introduced a large-scale, realistic RIR simulation dataset, AcousticRooms, which includes diverse room categories, geometries, and material properties. Results under the simulation settings show that our framework outperforms prior methods and strong baselines in both seen and unseen environments. Furthermore, sim-to-real transfer experiments reveal that our model, pre-trained on simulated data, effectively generalizes to real-world settings. Future work may focus on improving modeling techniques, such as using generative approach as proposed for sound generation [30, 50] to achieve better performance on acoustic modeling with minimal reference RIRs, or dynamically choose the suitable number of reference RIRs needed depending on the complexity of the environment.

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. *arXiv preprint arXiv:2310.15130*, 2023. 2

[2] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. *arXiv preprint arXiv:2406.08920*, 2024. 2

[3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 2, 3, 5

[4] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[5] 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. 3

[6] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022. 2

[7] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6409–6419, 2023. 3

[8] Mingfei Chen and Eli Shlizerman. AV-cloud: Spatial audio rendering through audio-visual cloud splatting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[9] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. *Conference on Robot Learning (CoRL)*, 2021. 3

[10] Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21886–21896, 2024. 1, 2

[11] Sanjoy Chowdhury, Sreyan Ghosh, Dasgupta Subhrajyoti, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. *ICCV*, 2023. 3

[12] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *ICRA*. IEEE, 2020. 3

[13] Elke Deckers, Onur Atak, Laurens Coox, Roberto D'Amico, Hendrik Devriendt, Stijn Jonckheere, Kunmo Koo, Bert Pluymers, Dirk Vandepitte, and Wim Desmet. The wave based method: An overview of 15 years of research. *Wave Motion*, 51(4):550–565, 2014. 2

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] 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

[16] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[17] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial visual representation learning through echolocation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[18] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. Sonicverse: A multisensory simulation platform for training household agents that see and hear. In *International Conference on Robotics and Automation (ICRA),*, 2023. 3

[19] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. In *British Machine Vision Conference (BMVC)*, 2021. 3

[20] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Visually-guided audio spatialization in video with geometry-aware multi-task learning. In *International Journal of Computer Vision (IJCV)*, 2023. 3

[21] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. 5

[22] Reinhold Haeb-Umbach, Jahn Heymann, Lukas Drude, Shinji Watanabe, Marc Delcroix, and Tomohiro Nakatani. Far-field automatic speech recognition. *Proceedings of the IEEE*, 109(2):124–148, 2020. 2

[23] Brian Hamilton and Craig J. Webb. Room acoustics modelling using gpu-accelerated finite difference and finite volume methods on a face-centered cubic grid. 2013. 5

[24] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. *arXiv preprint arXiv:2411.06307*, 2024. 2

[25] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP journal on advances in signal processing*, 2007:1–19, 2007. 2

[26] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 3

[27] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36:37472–37490, 2023. 2

[28] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arXiv preprint arXiv:2309.15977*, 2023. 1

[29] Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian. Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3765–3773, 2024. 3

[30] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see - video to audio generation through text. In *Advances in Neural Information Processing Systems*, pages 101337–101366. Curran Associates, Inc., 2024. 8

[31] 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, 3, 6

[32] 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, 5, 6, 7

[33] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[34] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 2018. 3

[35] Brady Peters. Integrating acoustic simulation in architectural design workflows: the fabpod meeting room prototype. *Simulation*, 91(9):787–808, 2015. 2

[36] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. Irgan: Room impulse response generator for far-field speech recognition. *arXiv preprint arXiv:2010.13219*, 2020. 2

[37] 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

[38] 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. 2

[39] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–575. IEEE, 2022. 2

[40] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27164–27175, 2024. 2

[41] W H Reed and T R Hill. Triangular mesh methods for the neutron transport equation. 1973. 2, 6

[42] 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

[43] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 351–355. IEEE, 2018. 2

[44] Carl Schissler, Gregor Mückl, and Paul Calamia. Fast diffraction pathfinding for dynamic sound propagation. 40 (4), 2021. 2

[45] Manfred R Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(6_Supplement):1187–1188, 1965. 6

[46] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 286–295, 2021. 2

[47] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2

[48] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[49] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022. 1, 2, 3, 4, 6

[50] Kun Su, Xiulong Liu, and Eli Shlizerman. From vision to audio and beyond: A unified model for audio-visual representation and generation. *arXiv preprint arXiv:2409.19132*, 2024. 8

[51] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. Gwa: A large high-quality acoustic dataset for audio processing. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2, 5

[52] Po-Yao Huang Andrew Owens Gopala Anumanchipalli Tingle Li, Renhao Wang. Self-supervised audio-visual soundscape stylization. In *ECCV*, 2024. 3

[53] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4, 5

[54] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: a dataset for finding humans using room acoustics. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[55] Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11790–11799, 2024. 2, 6, 7, 8