SOAF: Scene Occlusion-aware Neural Acoustic Field

Huiyu Gao[1], Jiahao Ma[1,2], David Ahmedt-Aristizabal[2], Chuong Nguyen[2], Miaomiao Liu[1]
[1]Australian National University, [2]CSIRO Data61
{huiyu.gao, jiahao.ma, miaomiao.liu}@anu.edu.au
{david.ahmedtaristizabal, chuong.nguyen}@data61.csiro.au

Abstract

This paper tackles the problem of novel view acoustic synthesis along an arbitrary trajectory in an indoor scene, given the audio-video recordings from other known trajectories of the scene. Existing methods often overlook the effect of room geometry, particularly wall occlusions on sound propagation, making them less accurate in multi-room environments. In this work, we propose a new approach called Scene Occlusion-aware Acoustic Field (SOAF) for accurate sound generation. Our approach derives a global prior for the sound field learning through distance-aware parametric sound propagation modeling and then transforms it based on the scene structure learned from the input video. We extract features from the local acoustic field centered at the receiver using a Fibonacci Sphere to generate binaural audio for novel views with a direction-aware attention mechanism. Extensive experiments on the real dataset RWAVS *and the synthetic dataset* SoundSpaces *demonstrate that our method achieves superior performance in spatial audio generation.*

1. Introduction

We live in a world rich with audio-visual multimodal information. Audio-visual scene synthesis [30] aims to generate videos with synchronized spatial audio along arbitrary novel camera trajectories, given a source video with associated binaural audio. This process involves reconstructing the scene both visually and acoustically from recordings captured from known poses, thereby enabling the synthesis of what a person would see and hear from novel positions and orientations while navigating within the environment. While Neural Radiance Fields (NeRF) [7, 24, 36, 38, 42, 51, 69] have achieved remarkable progress in novel view synthesis, they focus exclusively on the visual modality, disregarding the accompanying audio track. Yet the real world is inherently multimodal. Synthesizing spatial audio that matches visual content from a new perspective is therefore crucial for creating richer and more immersive experiences in audio-visual, virtual reality (VR), and augmented reality (AR) applications.

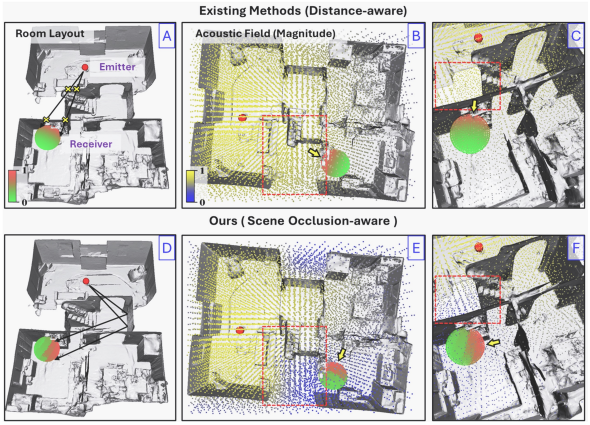

Figure 1. Pure distance-aware acoustic field [30, 34] *vs.* our proposed Scene Occlusion-aware Acoustic Field (SOAF). **Left column** (A & D) shows sound propagation in a room: the small ball represents the emitter and the large red-green ball represents the receiver. Coded colors indicate sound intensity, with red to green denoting high to low. **Middle column** (B & E) visualizes the magnitude distribution of the acoustic field, with yellow to blue indicating high to low. The comparison of sound attenuation through walls, highlighted by red dashed bounding boxes in sub-figs B, C, E and F emphasizes our consideration of wall obstruction. **Right column** (C & F) highlights the existing methods' neglect of obstruction in sound propagation in C while the proposed method gives higher sound intensity near the door than near the wall in F.

In principle, reverberant spatial audio can be obtained by convolving the Room Impulse Response (RIR) from the source to the receiver with the emitted audio signal [27]. However, acquiring ground-truth RIRs is both labor-intensive and time-consuming, as it demands specialized equipment and carefully controlled conditions such as sine-sweep excitations in acoustically quiet spaces. Although RIRs can be generated with acoustic simulators [8, 11, 57], these methods typically depend on complex hybrid algorithms and are restricted to synthetic scenes with predefined object materials and acoustic parameters. Such con-

straints substantially limited the applicability and scalability of approaches [3, 17, 28, 31, 32, 35, 48–50, 63] that optimized with RIR-based loss functions (*e.g.*, energy decay loss), especially when the environments are only observable through images and not physically accessible, making both RIR measurement and simulation impractical.

Instead of relying on RIR-specific formats, alternative methods [2, 10, 12, 14, 30] synthesize spatial audio by learning acoustic masks directly from reverberant recordings, which better reflect daily listening environments but are harder to model due to their entanglement with source waveform variations. These methods leverage visual cues from the receiver's viewpoint, enabling simpler data collection and integration with visual rendering pipelines. However, their performance relies primarily on local perspective cues and positional coordinates, which are insufficient to capture the broader scene context. As shown in Figure 1, when the sound source is occluded or located in a separate room, local features alone cannot capture critical spatial information, such as structural occlusions, essential for realistic acoustic rendering. In this work, we address this limitation by leveraging visual sequences to infer complete room geometry, thereby introducing effective geometric cues that capture the spatial relationship between the source and receiver. Together with our direction-aware attention, these features provide useful heuristics for learning the scene's acoustic field, leading to improved audio synthesis quality.

More specifically, as a sound wave propagates through space, its intensity attenuates with distance and is reflected, absorbed, or transmitted by surfaces [25, 27]. The intensity received at any 3D position is determined by the overall scene geometry and material properties [8, 29, 49, 57]. Therefore, given video sequences, we first learn a representation of the scene's geometry and appearance from visual frames using NeRF [38, 64, 71]. To enhance the sound field learning, we derive a scene occlusion-aware prior, termed the *global acoustic field*, based on simplified distance-aware parametric sound propagation modeling centered at the sound source and transformed by the extracted scene structure. We then extract the acoustic feature from the *local acoustic field* around the receiver using a Fibonacci Sphere, followed by a direction-aware attention mechanism to obtain binaural features. These features are used to generate binaural audio at novel views, demonstrating superior performance.

In summary, our contributions are as follows: (i) We introduce a scene occlusion-aware global prior for the sound field, enabling us to explicitly incorporate scene geometry and occlusions for accurate audio generation. (ii) Our direction-aware attention captures useful local features to enhance binaural audio synthesis. Extensive experiments on real and synthetic datasets, including RWAVS and SoundSpace, demonstrate the effectiveness of our approach.

2. Related Work

Neural Radiance Fields and Implicit Surface. Neural Radiance Fields (NeRF) [38] has emerged as a promising representation of scene appearance and has been widely used in novel view synthesis. Subsequent works [7, 24, 36, 42, 51, 69] extend NeRF in various aspects, including faster training [7, 24, 42], faster inference [51, 69], and handling in-the-wild images [36]. However, these methods struggle to extract high-quality surfaces due to insufficient surface constraints during optimization. To solve this issue, NeuS [64] and VolSDF [66] propose utilizing the signed distance function (SDF) as an implicit surface representation and developing new volume rendering methods to train neural SDF fields. Some following works, like MonoSDF [71] and NeuRIS [62], demonstrate the effectiveness of incorporating monocular depth priors [71] and normal priors [62, 71] as additional geometric cues for learning implicit surface representation of indoor scenes from sequences of scene images. In our work, we follow [64, 71] to achieve surface reconstruction from input images for our occlusion-aware spatial audio generation.

***RIR-based* Acoustic Fields.** The representation of spatial sound fields has been studied extensively. Traditional methods either approximate acoustic fields with handcrafted priors [1, 37, 59] or model perceptual cues with parametric representations [6, 46, 47], but both rely on strong assumptions. To overcome this, researchers have shifted toward data-driven approaches that model scene acoustics more generally. Neural Acoustic Field (NAF)[34] pioneers modeling RIRs across emitter–listener pairs using an MLP, and INRAS[55] extends this by disentangling emitter, geometry, and listener features under known room boundaries. NACF [31] incorporates multiple acoustic contexts and introduces a multi-scale energy decay criterion to supervise generated RIRs. Few-shotRIR [35] employs a transformer-based network to extract perceptual features and predict RIRs for queried source–receiver pairs with a decoder module. DIFF-RIR [63] develops a differentiable RIR rendering framework that reconstructs spatial acoustic characteristics from RIR measurements and planar scene geometry. AVR [28] proposes acoustic volume rendering to model sound propagation and enforce multi-view acoustic consistency. However, these methods [3, 17, 28, 31, 32, 34, 35, 55, 63] achieve good performance with specialized RIR-based loss functions, but perform poorly—or cannot be applied at all—when RIRs are unavailable. In contrast, our approach remains robust with reverberant audio recordings, making it practical and effective for real-world applications.

Audio-visual Learning. Recent works have explored learning acoustic information from multimodal data for diverse tasks, including sound localization [39, 40, 58], audio-visual navigation [8, 9, 68], visual-acoustic matching [10, 53], dereverberation [13, 18], and audio separation [67, 72].

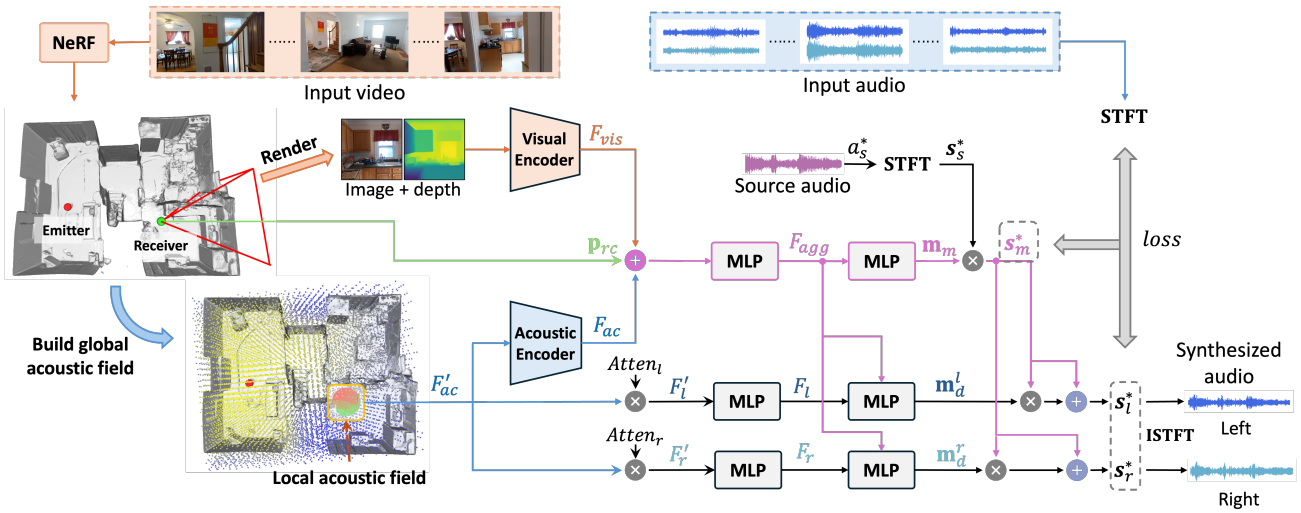

Figure 2. The pipeline of our proposed SOAF method. We first reconstruct the scene using NeRF from a calibrated video and build the global acoustic field. For audio synthesis at a new receiver pose $\mathbf{p_{rc}}$, we extract the acoustic feature F_{ac} from the local acoustic field around the receiver, and combine it with F_{vis} obtained from synthesized novel view images and $\mathbf{p_{rc}}$ to predict F_{agg} and mixture acoustic mask $\mathbf{m}_m$. To distinguish left and right channels, we propose a direction-aware attention mechanism to generate channel-specific features F'_l, F'_r based on distinct attention $Atten_l, Atten_r$ to the local acoustic field. Then the difference masks $\mathbf{m}_d^l$ and $\mathbf{m}_d^r$ are estimated with F_{agg} combined with the refined channel feature F_l or F_r, separately. Finally, we synthesize the binaural audio by combining the source audio magnitude $\mathbf{s}_s^*$ and the predicted masks $\mathbf{m}_m$, $\mathbf{m}_d^l$, and $\mathbf{m}_d^r$.

For novel-view acoustic synthesis, existing work relies on audio–visual data either from a single camera–microphone rig or synchronized multi-rig configurations. Under the latter, BEE [15] and SoundVista [16] perform audio reconstruction from a set of audio–visual references but require simultaneous recordings from multiple cameras and microphones, sparsely deployed at strategically selected locations within the scene. Under a single camera–microphone setup, ViGAS [12] fuses audio and image cues from one viewpoint to render sounds at a target view, assuming the source is visible in the input image and restricting synthesis to a few views. AV-NeRF [30] employs vanilla NeRF to render novel-view images and extracts visual features to synthesize novel-view audio, while AV-GS [2] and AV-Cloud [14] adopt point-based scene representations for greater efficiency. In contrast, our approach derives effective spatial features from visual images, which is orthogonal to existing contributions and agnostic to the underlying representation, yielding enhanced audio generation.

3. Problem Definition

Novel-view acoustic synthesis [12, 30] aims to generate spatial audio for unseen viewpoints in a scene from observations at multiple known viewpoints. We consider audio-visual recordings from a single camera-microphone rig as input, where the audio may consist of arbitrary sounds and does not require known RIRs for each source–receiver pair. Formally, we define a set of audio-visual observations $O = \{O_1, O_2, \ldots, O_N\}$ in an environment E, where each O_i includes an image $\mathbf{I}$ and a binaural audio clip a_t recorded at a receiver pose $\hat{\mathbf{p}}_{rc} = (\mathbf{p}_{rc}, \mathbf{d}_{rc})$, with $\mathbf{p}_{rc}, \mathbf{d}_{rc} \in \mathbb{R}^3$ denoting position and direction. For each a_t, the environment provides the corresponding monaural source audio a_s and source position $\mathbf{p}_{sr} \in \mathbb{R}^3$. Given such observations O, our goal is to learn the scene's acoustic field such that, for a novel receiver pose $\hat{\mathbf{p}}^*_{rc}$ and source audio a^*_s, the model can synthesize binaural audio a^*_t that reflects how the sound would be perceived at the target pose within E. It can be formulated as

$$a_t^* = f_\theta(\hat{\mathbf{p}}^*_{rc},\ a^*_s \mid O,\ E), \tag{1}$$

where f_θ denotes a parameterized neural network that models the acoustic field of environment E and generalizes to unseen listener poses within it.

4. Method

4.1. Overview

We adopt the acoustic masks for binaural audio prediction. Specifically, the acoustic masks consist of $\mathbf{m}_m, \mathbf{m}_d^l, \mathbf{m}_d^r \in \mathbb{R}^{F\times W}$, where F represents the frequency bins and W is the number of time frames. $\mathbf{m}_m$ captures changes in audio magnitude at the receiver position $\mathbf{p}_{rc}$ relative to the sound source position $\mathbf{p}_{sr}$, $\mathbf{m}_d^l$ and $\mathbf{m}_d^r$ characterize the changes for left and right channels of the binaural audio. Given the spectrogram magnitude of the input audio clip a^*_s after applying the Short-Time Fourier Transform (STFT), defined as $\mathbf{s}^*_s = \text{STFT}(a^*_s) \in \mathbb{R}^{F\times W}$, and predicted acoustic masks

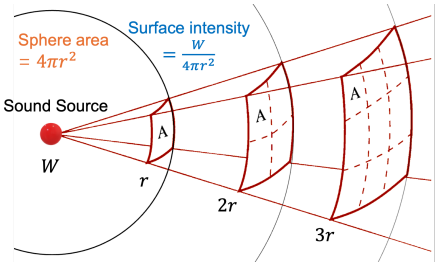

Figure 3. Illustration of the inverse square law [21, 61]. As the sound wave travels away from its source, the energy twice as far away from the source is distributed over four times the area, hence the intensity is one-quarter.

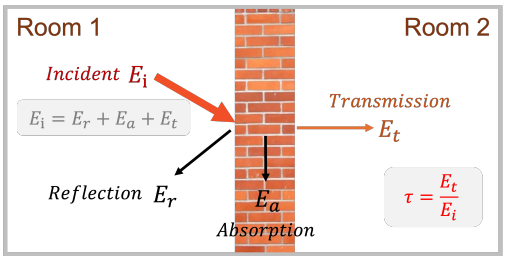

Figure 4. Illustration of the sound transmission coefficient τ [4, 56], which represents the ratio of transmitted sound energy when the sound wave travels through the barrier.

$\mathbf{m}_m$, $\mathbf{m}_d^l$, $\mathbf{m}_d^r$, we can synthesize the changed magnitude of the binaural audio as

$$\begin{aligned} \mathbf{s}_m^* &= \mathbf{s}_s^* \odot \mathbf{m}_m, \\ \mathbf{s}_l^* &= \mathbf{s}_m^* \odot (1 + \mathbf{m}_d^l), \\ \mathbf{s}_r^* &= \mathbf{s}_m^* \odot (1 + \mathbf{m}_d^r), \end{aligned} \tag{2}$$

where $\odot$ denotes the element-wise multiplication operation, $\mathbf{s}_l^*$ and $\mathbf{s}_r^*$ represent the magnitude of the synthesized audio in the left and right channels, respectively. Finally, we can obtain the binaural audio as $a_t^* = [\text{ISTFT}(\mathbf{s}_l^*), \text{ISTFT}(\mathbf{s}_r^*)]$, where ISTFT denotes the inverse STFT.

The overall pipeline of our approach is illustrated in Figure 2. Beyond extracting visual features from novel-view images, our framework reconstructs scene geometry from the input video, then builds the global and local acoustic field to derive auditory features for acoustic mask prediction. Details are provided below.

4.2. Image Synthesis and Visual Feature Extraction

Following AV-NeRF [30], we achieve novel view image synthesis and visual feature extraction by learning a NeRF from the input video. In particular, we render an RGB image and a depth map from the receiver's pose, then extract color and depth features from the rendered images with a pre-trained encoder of ResNet-18 [26]. These extracted features are concatenated as the visual feature F_{vis}, which provides important visual cues about the environment.

4.3. Global and Local Acoustic Field

Global Acoustic Field. The global acoustic field describes the distribution of sound intensity generated by the sound source throughout the scene. Since the 3D geometry and material properties determine sound propagation in an environment [8, 27, 29, 49, 57], we derive a global prior of the acoustic field to enhance the audio synthesis quality based on the scene geometry extracted from the video frames. Given the sound source position $\mathbf{p}_{sr}$, we calculate the prior value at any point $\mathbf{p}_i$ in the scene by considering (i) the *distance-aware* sound intensity attenuation; (ii) the *occlusion-aware* sound intensity transmittance.

Distance-aware prior. When sound propagates in the air, its intensity decreases as the distance from the sound source increases [25]. It obeys the *inverse square law* [21, 61] in free space, which describes that the intensity of the sound is inversely proportional to the square of the distance from the sound source. As illustrated in Figure 3, the finite amount of energy created by the sound source is spread thinner and thinner along the expanding surface area of the sphere. Inspired by this, we propose to explicitly quantify the sound attenuation based on the propagation distance, and calculate a prior value $E_0(\mathbf{p}_i, \mathbf{p}_{sr})$ at the point $\mathbf{p}_i$ as

$$E_0(\mathbf{p}_i, \mathbf{p}_{sr}) = \frac{1}{4\pi d(\mathbf{p}_i, \mathbf{p}_{sr})^2}, \tag{3}$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance between points.

Occlusion-aware prior. Unlike light waves, sound waves have much longer wavelengths, which allows them to diffract around small objects and propagate further [43]. Large obstacles, particularly walls, significantly influence the acoustics in an indoor setting [33]. Therefore, in this work, we check wall occlusions at the middle height of the scene to reduce the influence of small obstacles on the ground. To obtain the scene geometry, we learn a neural SDF field [70] from the visual images, and find the locations of scene walls with the zero-level set of SDF. As shown in Figure 4, when sound waves travel through a wall from one side to another, the ratio of transmitted sound energy is termed as the *sound transmission coefficient* τ [4, 56]. It can be expressed as

$$\tau = \frac{E_t}{E_i}, \tag{4}$$

where E_i and E_t represent the sound energy before and after the sound waves traverse the barrier, respectively.

Finally, for any point $\mathbf{p}_i$ in the scene, we generate a prior value $E(\mathbf{p}_i, \mathbf{p}_{sr})$ that combines distance and wall occlusion information by

$$E(\mathbf{p}_i, \mathbf{p}_{sr}) = E_0(\mathbf{p}_i, \mathbf{p}_{sr}) \times \tau^n, \tag{5}$$

where n is the number of walls found between the point $\mathbf{p}_i$ and the sound source $\mathbf{p}_{sr}$ in the SDF field. We convert $E(\mathbf{p}_i, \mathbf{p}_{sr})$ to a logarithmic scale, following standard practice in acoustics, and normalize it to $\hat{E}(\mathbf{p}_i, \mathbf{p}_{sr}) \in [0, 1]$ with the maximum and minimum values across the scene. While τ typically depends on wall material and thickness, we set a uniform $\tau = 0.25$ to simply indicate the presence of occlusions. Note that we are not aiming to model precise sound propagation for RIR rendering, but rather to introduce geometric cues that guide the acoustic field learning.

Local Acoustic Field. The local acoustic field depicts the distribution of sound intensity around the receiver. Inspired by the design of spherical microphone arrays [45], we utilize a Fibonacci Sphere [52] around the receiver to collect the sound intensity in the global acoustic field. Specifically, we first generate a unit Fibonacci Sphere with G points on the sphere's surface centered at the receiver, then emit rays from the sphere center toward each surface point, in the directions $\mathbf{d}_{Fib} \in \mathbb{R}^{3\times G}$. To build the local acoustic field, we uniformly sample H points along each ray within the radius range $[r_{min}, r_{max}]$, resulting in a total of $G \times H$ sampled points. For each sampled point $\mathbf{p}_i$, we apply Equation 5 and obtain its prior $\hat{E}(\mathbf{p}_i, \mathbf{p}_{sr})$. By integrating the prior values of the H points along each direction, we obtain the local acoustic feature $F'_{ac} \in \mathbb{R}^G$, representing the priors across G directions. Rather than treating all H points equally along the same direction, we assign greater importance to points closer to the receiver by applying weights $w_i = e^{-d(\mathbf{p}_i, \mathbf{p}_{rc})} \in [0, 1]$ and computing a weighted sum.

As shown in Figure 2, our local acoustic feature F'_{ac} is fed into an acoustic encoder to reduce the feature dimension and get the refined feature F_{ac}. Combining it with the visual feature F_{vis} and the positional encoding of the receiver position $\mathbf{p}_{rc}$, we aggregate these features as F_{agg} and estimate the mixture acoustic mask $\mathbf{m}_m$ with MLPs.

4.4. Direction-aware Attention Mechanism

Given the local acoustic field, we propose a direction-aware attention mechanism to distinguish the left and right channel features to generate binaural audio. Specifically, we calculate the similarity between the left or right channel directions $\mathbf{d}_l, \mathbf{d}_r \in \mathbb{R}^3$ with the sphere points directions $\mathbf{d}_{Fib}$ to obtain the attention $Atten_l, Atten_r \in \mathbb{R}^G$ for each channel, then combine these attentions with local acoustic feature F'_{ac} to obtain binaural features. It can be expressed as:

$$Atten_l = \mathbf{d}_{Fib}^T \mathbf{d}_l, \quad Atten_r = \mathbf{d}_{Fib}^T \mathbf{d}_r, \tag{6}$$

$$F'_l = F'_{ac} \odot Atten_l, \quad F'_r = F'_{ac} \odot Atten_r, \tag{7}$$

where $\odot$ denotes element-wise multiplication. After further transformation of F'_l and F'_r to F_l and F_r, respectively, to align their dimension with F_{agg}, we combine F_{agg} with F_l or F_r separately to estimate $\mathbf{m}_d^l$ or $\mathbf{m}_d^r$.

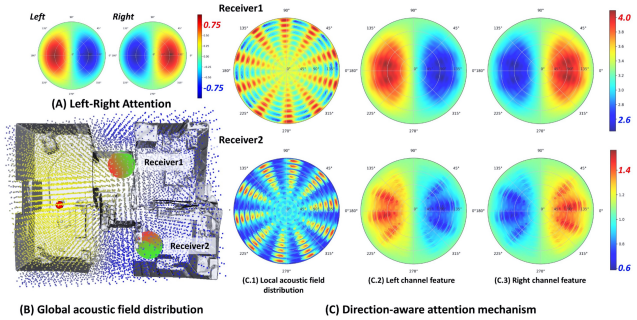

Figure 5. Direction-aware attention mechanism. (**A**) Predefined left-right attention. (**B**) Local distribution of two receivers: Receiver 1 in the hallway, close to the sound source with higher intensity; Receiver 2 in the kitchen, further away and obstructed with lower intensity. Intensity comparison is highlighted in the subfigure C color bar. (**C**) The binaural features describe the spatial and directional sound characteristics generated by the combination of the left-right attention and the local acoustic field distribution.

Figure 5 compares the local acoustic fields and binaural channel features for two receivers at different positions. The color bar shows the differing sound intensities of their left and right channels. Binaural features are generated by considering the channel directions and the sound intensity in the local sound field comprehensively.

4.5. Learning Objective

On the RWAVS dataset, we predict the acoustic masks $\mathbf{m}_m$, $\mathbf{m}_d^l$, $\mathbf{m}_d^r$ and obtain the estimated magnitudes $\mathbf{s}_m^*$, $\mathbf{s}_l^*$, $\mathbf{s}_r^*$ via Equation 2. We optimize the network using the following loss function:

$$\mathcal{L}_A = \|\mathbf{s}_m - \mathbf{s}_m^*\|^2 + \|\mathbf{s}_l - \mathbf{s}_l^*\|^2 + \|\mathbf{s}_r - \mathbf{s}_r^*\|^2, \tag{8}$$

where $\mathbf{s}_m$, $\mathbf{s}_l$ and $\mathbf{s}_r$ denote the ground-truth magnitudes, corresponding to the mixture, left, and right channels, respectively. The mixture $\mathbf{s}_m$ is defined as the average of $\mathbf{s}_l$ and $\mathbf{s}_r$. The first term of $\mathcal{L}_A$ encourages the network to predict the masks reflecting spatial effects caused by distance and geometry occlusion. The second and third terms encourage the network to generate masks that capture the differences between the left and right channels.

4.6. Implementation Details

Our model is implemented using PyTorch [44] and optimized by the Adam [19] optimizer, with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 5×10^{-4} and is exponentially decreased to 5×10^{-6}. The training process spans 200 epochs, with a batch size of 32. When building the Fibonacci Sphere, we set $G = 1024$ (uniformly generate 1024 rays around the sphere center) and $H = 10$ (uniformly sample 10 points along each ray), with the radius range $r_{min} = 0.01$ and $r_{max} = 1$. The acoustic transmission coefficient τ is set to 0.25. All our experiments are conducted on an RTX 4090 GPU.

Methods	Office w/o *occ*		House w/ *occ*		Apartment w/ *occ*		Outdoors w/o *occ*		Mean	
	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓
Mono-Mono	9.269	0.411	11.889	0.424	15.120	0.474	13.957	0.470	12.559	0.445
Mono-Energy	1.536	0.142	4.307	0.180	3.911	0.192	1.634	0.127	2.847	0.160
Stereo-Energy	1.511	0.139	4.301	0.180	3.895	0.191	1.612	0.124	2.830	0.159
INRAS [55]	1.405	0.141	3.511	0.182	3.421	0.201	1.502	0.130	2.460	0.164
NAF [34]	1.244	0.137	3.259	0.178	3.345	0.193	1.284	0.121	2.283	0.157
ViGAS [12]	1.049	0.132	2.502	0.161	2.600	0.187	1.169	0.121	1.830	0.150
AV-NeRF [30]	0.930	0.129	2.009	0.155	2.230	0.184	0.845	0.111	1.504	0.145
SOAF (Ours)	**0.795**	**0.125**	**1.952**	**0.153**	**2.098**	**0.182**	**0.737**	**0.108**	**1.396**	**0.142**

Table 1. Quantitative results on the RWAVS dataset. Our method consistently achieves improved performance across diverse environments. "w/ *occ*" denotes a multi-room indoor scene with *occ*lusion.

Methods	T60 (%) ↓	C50 (dB) ↓	EDT (sec) ↓
Opus-nearest	10.10	3.58	0.115
Opus-linear	8.64	3.13	0.097
AAC-nearest	9.35	1.67	0.059
AAC-linear	7.88	1.68	0.057
NAF [34]	3.18	1.06	0.031
INRAS [55]	3.14	0.60	0.019
AV-NeRF [30]	2.47	0.57	0.016
SOAF (Ours)	**2.29**	**0.54**	**0.014**

Table 2. Quantitative results on the Soundspaces dataset. Our method achieves improved results in all metrics.

5. Experiments

5.1. Datasets, Baselines & Metrics

Datasets. We evaluate our method on the real-world dataset RWAVS and the synthetic dataset SoundSpaces.

RWAVS dataset. The Real-World Audio-Visual Scene (RWAVS) dataset is collected by the authors of AV-NeRF [30] from diverse real-world scenarios, divided into four categories: *office*, *house*, *apartment*, and *outdoor* environments. In particular, the *office* category includes indoor scenes with single-room layouts, while the *house* and *apartment* categories present scenes with multi-room layouts. To capture various acoustic and visual signals along different camera trajectories, the data collector moved randomly through the environment while holding the recording device. For each scene, RWAVS contains multimodal data including source audio, collected high-quality binaural audio, video, and camera poses, ranging from 10 to 25 minutes. Camera positions are densely distributed throughout the scene, and camera directions are sufficiently diverse. For a fair comparison, we maintain the official training/test split, which contains 9,850 samples for training and 2,469 samples for testing, respectively.

SoundSpaces dataset. SoundSpaces [8, 54] is a synthetic dataset generated based on hybrid sound propagation methods [5, 20, 60] that simulates fine-grained acoustic properties by simultaneously considering the effects of room geometry and surface materials on sound propagation in a 3D environment. Following previous methods [30, 34, 55], we validate our method on the same six representative indoor scenes, including two single rooms with rectangular walls, two single rooms with non-rectangular walls, and two multi-room layouts. For each scene, SoundSpaces provides binaural impulse responses for extensive emitter and receiver pairs sampled within the room at a fixed height from four different head orientations (0°, 90°, 180°, and 270°). To validate the effectiveness of our approach on this dataset, we modify our model to estimate binaural impulse responses instead of acoustic masks. We keep the same training/test split as previous works by using 90% data for training and 10% data for testing.

Baselines. We compare our approach with baseline methods [12, 30, 34, 55] that also learn a neural acoustic field with implicit representation. Among these methods, NAF [34] learns audio signals with a trainable local feature grid while INRAS [55] disentangles scene-dependent features from audio signals and reuses them for all emitter-listener pairs. ViGAS [12] and AV-NeRF [30] are multimodal approaches that leverage the visual feature of a single image for audio generation. On the RWAVS dataset, we include three additional baselines for reference: Mono-Mono, Mono-Energy, and Stereo-Energy. Mono-Mono simply repeats the source audio twice to achieve a binaural effect. Mono-Energy scales the energy of the source audio to match the average energy of the ground truth target audio then duplicates it to obtain a binaural audio. Stereo-Energy first duplicates the source audio and then scales the two channels separately to match the energy of each channel of the ground truth target audio. On SoundSpaces, we also compare our model with the linear and nearest neighbor interpolation results of two widely used audio coding methods: Advanced Audio Coding (AAC) [22] and Xiph

Methods	Office		House		Apartment		Outdoors		Mean	
	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓	MAG↓	ENV↓
Ours - w/o *geo, dir*	0.928	0.129	2.015	0.155	2.249	0.184	0.831	0.111	1.506	0.145
Ours - w/o *dir*	0.868	0.127	1.977	0.154	2.168	0.183	0.785	0.109	1.450	0.143
Ours - *full*	**0.795**	**0.125**	**1.952**	**0.153**	**2.098**	**0.182**	**0.737**	**0.108**	**1.396**	**0.142**

Table 3. Ablation study of proposed components on RWAVS. "Ours - w/o *geo, dir*" refers to results reproduced from official AV-NeRF. "*geo*" denotes global-local acoustic field, "*dir*" denotes direction-aware attention mechanism and "*full*" denotes all proposed modules.

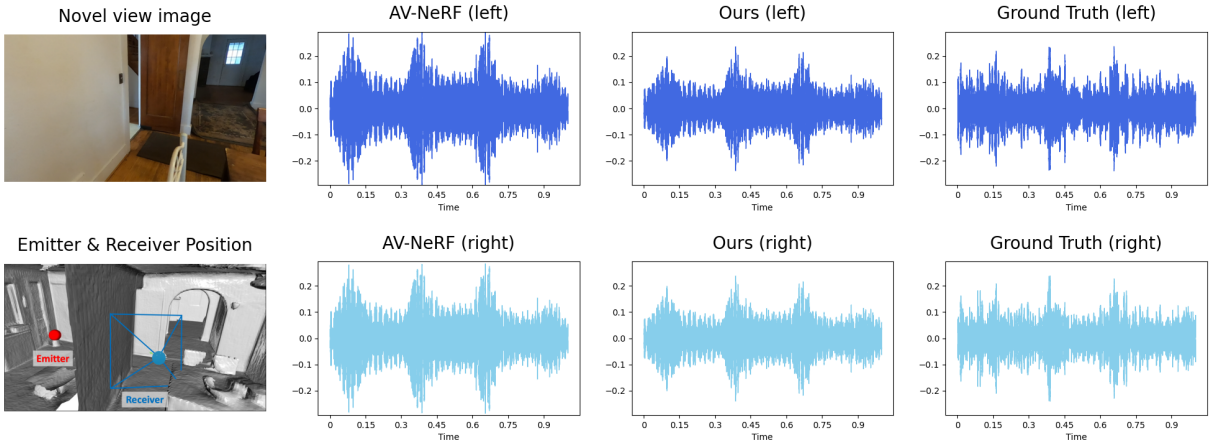

Figure 6. Example visual comparison of novel view audio synthesis. For the receiver, the emitter is blocked by a wall. Compared to AV-NeRF (MAG: 0.816; ENV: 0.124), our method (MAG: 0.471; ENV: 0.097) generates audio with more accurate energy attenuation.

Opus [23]. All these methods are evaluated with the same train/test split for each dataset.

Metrics. On the RWAVS dataset, we follow AV-NeRF to utilise the magnitude distance (MAG) [65] and envelope distance (ENV) [41] as evaluation metrics. MAG measures the audio quality of the generated sound in the time-frequency domain after applying the short-time Fourier transform, while ENV measures it in the time domain. On SoundSpaces, we follow [30, 55] to evaluate our method with the reverberation time (T60), acoustic parameter clarity (C50), and early decay time (EDT) metrics. T60 is the percentage error between the time it takes for the synthesised RIR to decay by 60 dB in the time domain with the ground truth T60 reverberation time. C50 describes the clarity and loudness of audio by quantifying the energy ratio between early reflections and late reverberation. EDT reflects people's perception of reverberation by focusing on the early reflections of impulse responses. For all of these metrics, lower is better. The detailed definitions of these metrics are provided in the supplementary material.

5.2. Results & Ablation study

For a fair comparison, we integrate our priors into AV-NeRF while keeping all other components (visual features, estimated poses, or optimizer settings) unchanged. This controlled setup isolates and validates the effect of our contributions. We present quantitative results on the RWAVS dataset in Table 1. Our model consistently outperforms baselines across all environments. Specifically, we achieve an overall 7.2% and 23.7% reduction in the MAG metric compared to AV-NeRF and ViGAS, respectively. This demonstrates that our approach extracts more comprehensive spatial cues from visual inputs and effectively guides learning of the neural acoustic field. Figure 6 shows a qualitative comparison of rendered audio. On Soundspaces, we adapt our method to predict RIRs without any architectural changes or additional supervision (e.g., energy decay loss), and provide results in Table 2. Compared to previous works, our approach achieves better performance in all metrics. In particular, we obtain an overall 7.3% reduction in T60 error compared to AV-NeRF and a 27.1% reduction compared to INRAS, respectively. The consistent improvement further shows the generality of our proposed priors. To further verify the agnosticism of scene representation, we also compare with AV-Cloud [14] (point-based approach) and include results in the supplementary material.

Ablation of Proposed Components. We conducted an ablation study on the RWAVS dataset. In Table 3, "w/o *geo*, *dir*" uses AV-NeRF's default input (visual features, receiver location, and relative direction). "w/o *dir*" only adds our global-local acoustic field (*geo*), showing improvements in all scenarios, demonstrating the effectiveness of our spatial geometric prior. "Ours - w/o *dir*" uses AV-NeRF's relative direction information and *geo*, while "Ours - full" incorpo-

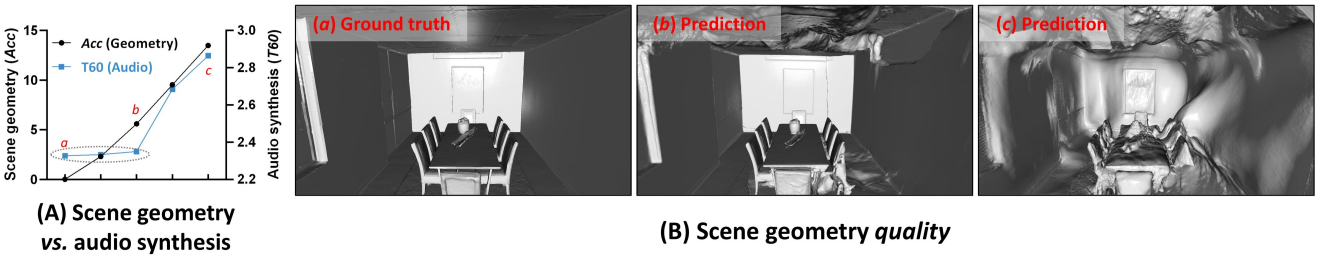

Figure 7. Robustness to scene reconstruction quality. **Left**: audio synthesis performance with various reconstructions. Small *Acc* and *T60* are better. **Right**: visual differences in reconstruction results. *a*, *b*, and *c* correspond to the different geometry quality shown on the left.

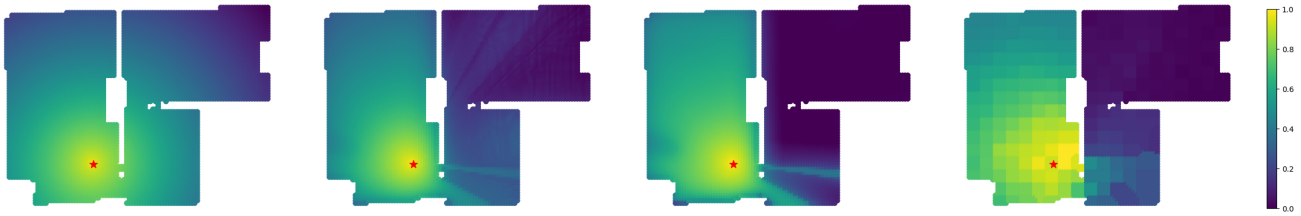

Figure 8. Visualization of varying values of transmission coefficient τ in our global acoustic field. $\tau = 1$ means all the sound energy will be transmitted to the other side of the wall, while $\tau = 0$ represents no sound energy will be transmitted.

rates all proposed modules, achieving the best performance.

Robustness to Reconstruction Quality. Our method reconstructs wall layouts to ensure that transmission attenuation occurs at the correct locations in our global priors. To evaluate the robustness of our approach to 3D reconstruction quality, we visualize audio synthesis results in relation to scene reconstruction accuracy in Figure 7. Specifically, we employ a vanilla NeRF [38] (Figure 7.B.*c*) and MonoSDF [71] (Figure 7.B.*b*) to obtain 3D scene reconstructions of varying quality. As shown in Figure 7.A, our audio synthesis performance is only slightly affected when the reconstruction error remains below 6.5.

Contribution of Transmission Coefficient τ. We evaluate our method across four scenes with wall *occ*lusions and four without, varying τ values. The average results are reported in Table 4. For single-room scenes, our method achieves a steady improvement with different τ values, demonstrating the effectiveness of combining our distance-aware priors with direction-aware attentions. However, for multi-room scenes, setting $\tau = 1$ produces pure distance-aware priors for receivers in other rooms, resulting in complete ignorance of wall occlusions, while $\tau = 0$ uniformly assigns zero values to all points behind the wall, leading to undifferentiated prior knowledge between these points. Based on experiments, we achieve a balanced integration of distance and occlusion effects by setting $\tau = 0.25$ for representative priors across receivers. A visualization of the impact of τ in our global acoustic field is shown in Figure 8.

Spatial Effect. We report the Left-Right Energy (LRE) error [12] on the RWAVS dataset in Table 5. The consistent decrease in all metrics, including the LRE error, shows that our direction-aware attention improves the spatial effect by providing informative binaural features.

Methods	Single-room (w/o *occ*)		Multi-room (w/ *occ*)	
	MAG↓	ENV↓	MAG↓	ENV↓
AV-NeRF [25]	0.780	0.119	1.963	0.159
Ours ($\tau = 1$)	0.672	0.113	1.902	0.158
Ours ($\tau = 0.5$)	0.672	0.113	1.864	0.157
Ours ($\tau = 0.25$)	**0.672**	**0.113**	**1.856**	**0.157**
Ours ($\tau = 0$)	0.673	0.113	1.871	0.157

Table 4. Ablation study of transmission coefficient τ.

Methods	MAG↓	ENV↓	LRE↓
Ours - w/o *geo, dir*	1.506	0.145	0.988
Ours - w/o *dir*	1.450	0.143	0.982
Ours - *full*	**1.396**	**0.142**	**0.956**

Table 5. Performance comparison on the RWAVS dataset.

6. Conclusion

In this work, we introduced effective geometric priors for the sound field learning, derived from distance-aware parametric sound propagation modeling and scene occlusions extracted from the input video. Our proposed direction-aware attention mechanism captures useful local features for binaural channels. Tested on the real dataset RWAVS and the synthetic dataset SoundSpaces, our approach outperforms existing works in audio generation.

Limitations. Following previous approaches, our model is scene-specific and deals with a static sound source. Currently, the proposed priors focus solely on geometric aspects—distance, occlusion, and direction—which have proven effective for audio synthesis. We believe that expanding them to incorporate factors such as reverberation or time of flight is a promising direction for future research.

Acknowledgement. This research was supported in part by the Australia Research Council ARC Discovery Grant (DP200102274).

References

[1] Niccolo Antonello, Enzo De Sena, Marc Moonen, Patrick A Naylor, and Toon Van Waterschoot. Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1929–1941, 2017. 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. *Advances in Neural Information Processing Systems*, 37:28920–28937, 2024. 2, 3

[3] Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. Neraf: 3d scene infused neural radiance and acoustic fields. *arXiv preprint arXiv:2405.18213*, 2024. 2

[4] Carmen Bujoreanu, Florin Nedeff, Marcelin Benchea, and Maricel Agop. Experimental and theoretical considerations on sound absorption performance of waste materials including the effect of backing plates. *Applied Acoustics*, 119:88–93, 2017. 4

[5] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 6

[6] Chakravarty R Alla Chaitanya, Nikunj Raghuvanshi, Keith W Godin, Zechen Zhang, Derek Nowrouzezahrai, and John M Snyder. Directional sources and listeners in interactive sound propagation using reciprocal wave field coding. *ACM Transactions on Graphics (TOG)*, 39(4):44–1, 2020. 2

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 1, 2

[8] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 1, 2, 4, 6

[9] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021. 2

[10] 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

[11] 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. 1

[12] 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. 2, 3, 6, 8

[13] 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

[14] Mingfei Chen and Eli Shlizerman. Av-cloud: Spatial audio rendering through audio-visual cloud splatting. *Advances in Neural Information Processing Systems*, 37:141021–141044, 2024. 2, 3, 7

[15] Mingfei Chen, Kun Su, and Eli Shlizerman. Be everywhere-hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7853–7862, 2023. 3

[16] Mingfei Chen, Israel D Gebru, Ishwarya Ananthabhotla, Christian Richardt, Dejan Markovic, Jake Sandakly, Steven Krenn, Todd Keebler, Eli Shlizerman, and Alexander Richard. Soundvista: Novel-view ambient sound synthesis via visual-acoustic binding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8331–8341, 2025. 3

[17] 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. 2

[18] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7884–7896, 2023. 2

[19] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 5

[20] M David Egan, JD Quirt, and MZ Rousseau. Architectural acoustics, 1989. 6

[21] TFW Embleton. Mean force on a sphere in a spherical sound field. i.(theoretical). *The Journal of the Acoustical Society of America*, 26(1):40–45, 1954. 4

[22] International Organization for Standardization. Advanced audio coding (aac). *ISO/IEC 13818-7:2006*, 2006. 6

[23] Xiph.Org Foundation. Xiph opus. *https://opus-codec.org/*, 2012. 7

[24] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 1, 2

[25] Steven L Garrett and Steven L Garrett. Attenuation of sound. *Understanding Acoustics: An Experimentalist's View of Sound and Vibration*, pages 673–698, 2020. 2, 4

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[27] Heinrich Kuttruff. *Room acoustics*. Crc Press, 2016. 1, 2, 4

[28] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *Advances in Neural Information Processing Systems*, pages 44600–44623. Curran Associates, Inc., 2024. 2

[29] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2, 4

[30] 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, 2023. 1, 2, 3, 4, 6, 7

[31] 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. 2

[32] Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastia V Amengual, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5732–5741, 2025. 2

[33] Marshall Long. *Architectural acoustics*. Elsevier, 2005. 4

[34] 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, 6

[35] 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

[36] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 1, 2

[37] Rémi Mignot, Gilles Chardon, and Laurent Daudet. Low frequency interpolation of room impulse responses using compressed sensing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):205–216, 2013. 2

[38] 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. 1, 2, 8

[39] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022. 2

[40] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10565–10574, 2023. 2

[41] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018. 7

[42] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 1, 2

[43] Robert Pasnau. What is sound? *The Philosophical Quarterly*, 49(196):309–324, 1999. 4

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[45] Boaz Rafaely. Analysis and design of spherical microphone arrays. *IEEE Transactions on speech and audio processing*, 13(1):135–143, 2004. 5

[46] Nikunj Raghuvanshi and John Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 2

[47] Nikunj Raghuvanshi and John Snyder. Parametric directional coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

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

[49] 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, 4

[50] 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

[51] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14335–14345, 2021. 1, 2

[52] Edward B Saff and Amo BJ Kuijlaars. Distributing many points on a sphere. *The mathematical intelligencer*, 19:5–11, 1997. 5

[53] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[54] 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. 6

[55] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022. 2, 6, 7

[56] Wei-Hong Tan, EA Lim, HG Chuah, EM Cheng, and CK Lam. Sound transmission loss of natural fiber panel. *Interna-*

tional Journal of Mechanical & Mechatronics Engineering, 16(6):33–42, 2016. 4

[57] 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. 1, 2, 4

[58] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 2

[59] Natsuki Ueno, Shoichi Koyama, and Hiroshi Saruwatari. Kernel ridge regression with constraint of helmholtz equation for sound field interpolation. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–440. IEEE, 2018. 2

[60] Eric Veach and Leonidas Guibas. Bidirectional estimators for light transport. In *Photorealistic Rendering Techniques*, pages 145–167. Springer, 1995. 6

[61] Nikolaos Voudoukis and Sarantos Oikonomidis. Inverse square law for light and radiation: A unifying educational approach. *European Journal of Engineering and Technology Research*, 2(11):23–27, 2017. 4

[62] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, pages 139–155. Springer, 2022. 2

[63] 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

[64] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[65] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 7

[66] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[67] Yuxin Ye, Wenming Yang, and Yapeng Tian. Lavss: Location-guided audio-visual spatial audio separation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5508–5519, 2024. 2

[68] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *IEEE Robotics and Automation Letters*, 8(2):928–935, 2023. 2

[69] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1, 2

[70] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. 4

[71] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 2, 8

[72] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer, 2020. 2