ICCV #4

# Differentiable Room Acoustic Rendering with Multi-View Vision Priors

Anonymous ICCV submission

Paper ID 4

## Abstract

*Spatial audio is essential for immersive AR/VR applications, yet existing existing methods for room impulse response estimation either needs dense training data or expensive physics simulation. In this work, we introduce Audio-Visual Differentiable Room Acoustic Rendering (*AV-DAR*), a framework that leverages visual cues extracted from multi-view images and acoustic beam tracing for physics-based room acoustic rendering. This multi-modal, physics-based, end-to-end framework is efficient, interpretable, and accurate. Experiments across six real-world environments from two datasets demonstrate that* AV-DAR *significantly outperforms a series of prior methods. Notably, on the Real Acoustic Field dataset,* AV-DAR *achieves comparable performance to models trained on 10 times more data while delivering relative gains ranging from 16.6% to 50.9% when trained at the same scale.*

## 1. Introduction

Spatial audio is a fundamental component of immersive multimedia experiences and is often regarded as "half the experience" in VR/AR applications. Recreating the spatial acoustic experience is analogous to novel-view synthesis [13, 15] in vision, where the goal is to synthesize photorealistic images from arbitrary viewpoints based on finite observations. Similarly, novel-view acoustic synthesis [9, 11, 12, 21, 23] aims to render the sound received at any listener location within a scene. A widely used representation for this task is the *Room Impulse Response* (RIR) [4, 20], which maps an emitted impulse to its received waveform, summing direct sound and reflections.

Existing methods for estimating RIRs generally fall into two broad categories: *learning-based* and *physics-based*. Learning-based approaches [3, 10, 12, 17, 21] treat RIR estimation as a regression task trained on densely measured ground-truth RIRs, thus requires massive training data and is lack of physically grounded guarantees. Although physics-based approaches [9, 23] rely on explicit acoustic models, they become computationally impractical in large, complex scenes, limiting real-world scalability.

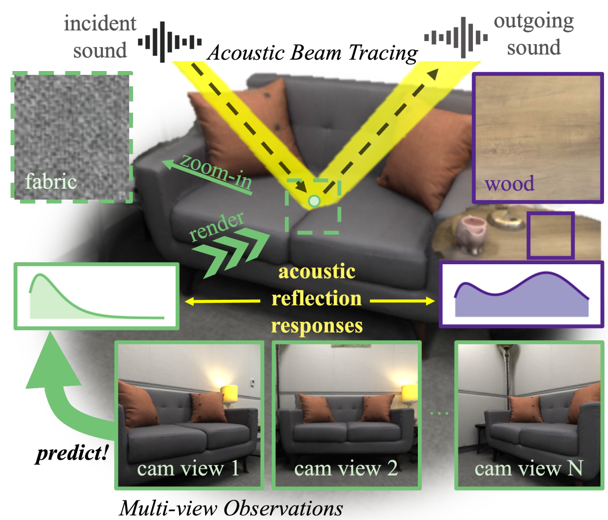


Figure 1. Our differentiable room acoustic rendering framework combines multi-view visual observations and acoustic beam tracing for efficient and accurate room impulse response (RIR) prediction. By analyzing the visual cues of surfaces (*e.g.*, fabric vs. wood), it infers acoustic reflection responses for accurately rendering RIRs through physics-based, end-to-end optimization.

Our key insight is that sound travels more slowly than light, both are influenced by the same room geometry and surface materials, therefore visual appearance and acoustic property should correlate to each other (*e.g.*, hard wooden tables reflect high-frequency sound, while soft carpets absorb it). Leveraging this link, we propose the *Audio-Visual Differentiable Room Acoustic Rendering* (AV-DAR), an end-to-end framework that leverages multi-view images to predict accurate room-impulse responses (RIRs). A cross-attention module maps image features from camera space to 3-D scene space, building a unified, material-aware representation that predicts reflection properties. On top of this, we run differentiable beam tracing to enumerate specular paths, requiring less computation and fewer training samples than existing physics-based methods, thus enabling accurate and data-efficient acoustic modeling. Across six real-world environments [5, 23], our method significantly out-

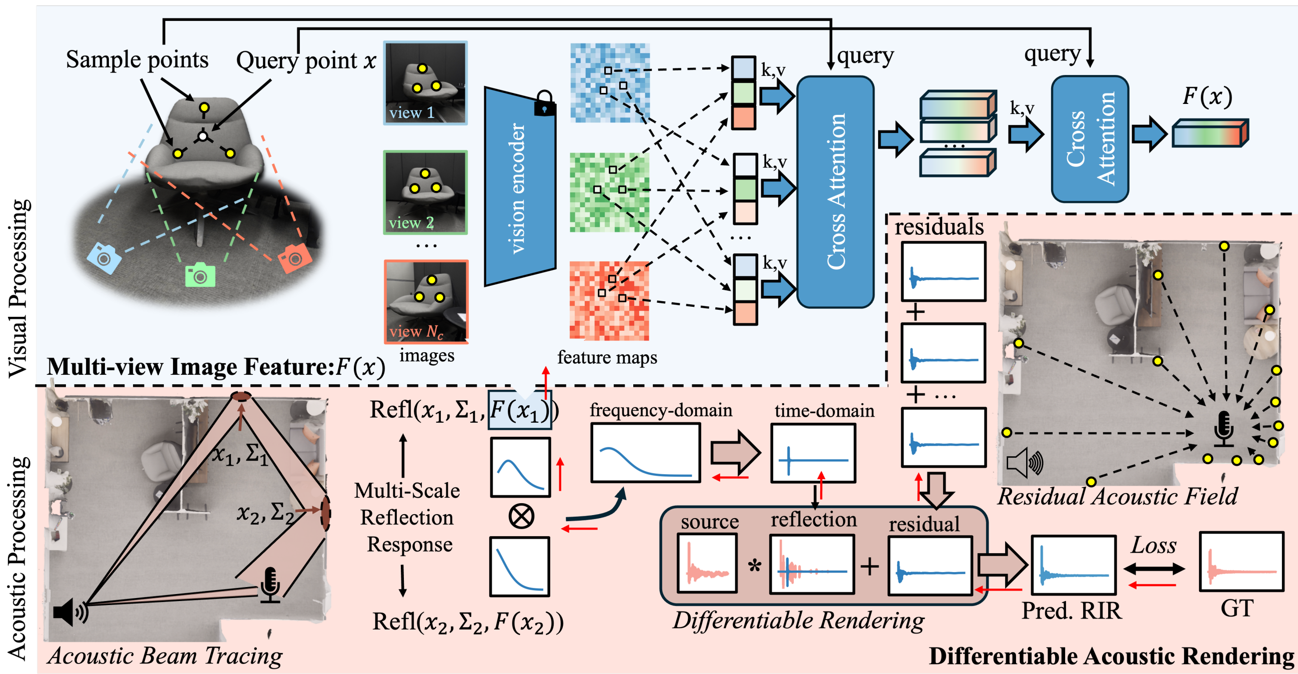


Figure 2. **Method Overview.** Our framework contains two main components for rendering the room impulse response (RIR): **(1) Visual Processing (top)**: Multi-view images of the scene are passed through a pre-trained vision encoder to extract pixel-aligned features at sampled points on the room surface. We then apply cross-attention both across views for each sampled point and across sampled points of query $\mathbf{x}$ to obtain a unified, material-aware visual feature $\mathrm{F}(\mathbf{x})$ (detailed in Section 2.5). **(2) Acoustic Processing (bottom):** On the left, we illustrate our acoustic beam tracing procedure (Section 2.3), where we sample specular paths and compute the path reflection response, conditioned on both the positional encoding (Section 2.4) and the visual feature $\mathrm{F}(\mathbf{x})$. On the right, we show how we model the residual acoustic field (Section 2.6) by treating every point on the surface as a secondary sound source and integrating its contribution via Monte-Carlo integration. The entire pipeline is fully differentiable, enabling end-to-end optimization of both acoustic and visual parameters.

performs existing baselines. On the RAF dataset [5], our model achieves comparable performance to existing methods trained on roughly 10× RIR measurements while delivering 16.6% to 50.9% improvement when trained at the same scale.

Our main contributions are threefold: First, we propose a *physics-based differentiable* room acoustic rendering pipeline that not only learn from sparse, real-world RIR measurements but is also *efficient*, *interpretable*, and *accurate*. Second, we are the first to integrate acoustic beam tracing within an end-to-end differentiable framework, enabling efficient computation of reflection responses. Third, our approach leverages multi-view images to capture material-aware visual cues that correlate with acoustic reflection properties, achieving significantly more accurate RIR rendering than prior methods.

# 2. Approach

## 2.1. Preliminaries

Our goal is to learn a time-domain room impulse response function $\mathrm{RIR}(\mathbf{x}_a, \mathbf{x}_b, \mathbf{p}_a, t)$ from sparse training data, where, $\mathbf{x}_a$, $\mathbf{x}_b$, $\mathbf{p}_a$ denote the speaker location, the listener position, and the source orientation, respectively.

Training uses a sparse set of ground-truth RIR measurements *plus* a set of multi-view images to capture the scene's visual information which contains visual material and geometric cues missing from acoustics alone. Concretely, we assume a set of $N_c$ RGB images with known intrinsics $\pi$ and extrinsics $P^{(i)}$:

$$\left\{\{I^{(i)}, \pi, P^{(i)}\}\,\middle|\, i = 1, \cdots, N_c\right\}. \quad (1)$$

Once RIR is learned, spatial audio at $\mathbf{x}_b$ for any dry signal $h(t)$ is obtained by convolution:

$$h_b(t) = h(t) \ * \ \mathrm{RIR}(\mathbf{x}_a, \mathbf{x}_b, \mathbf{p}_a, t), \quad (2)$$

enabling realistic spatial acoustic rendering.

## 2.2. Overview of the AV-DAR Framework

AV-DAR predicts room-impulse responses (RIRs) for arbitrary source–listener pairs using *sparse* measured RIRs, multi-view images, and coarse room geometry. Following [23] we decompose the target RIR as

$$\mathrm{RIR}(t) = \sum_{\tau} s(\tau; \Theta_1)\, R(t-\tau; \Theta_2) + r(t; \Theta_3), \quad (3)$$

where:

- $s(t;\Theta_1)$ is a learnable source response,
- $R(t;\Theta_2)$is the integrated reflection response, computed via *differentiable beam tracing* (Sec. 2.3) with multi-scale surface kernels (Sec. 2.4) and vision-conditioned material cues (Sec. 2.5),
- $r(t;\Theta_3)$ is the residual term capturing higher-order bounces, diffraction, and late reverberation (Sec. 2.6).

Overall, AV-DAR integrates all components into an end-to-end differentiable pipeline, enabling gradient-based optimization for accurate RIR rendering.

### 2.3. Acoustic Beam Tracing

We require a differentiable renderer that efficient and reliably captures specular paths. Image-source methods explode combinatorially [1], while stochastic ray tracing [8, 18, 19] misses specular bounces. Beam tracing [6, 7, 22] instead propagates cone-shaped volumes (see Supp.), marking a listener "hit" whenever it lies inside a beam.

From the source $\mathbf{x}_a$ we cast $N_d$ narrow Fibonacci-lattice beams; without splitting they return a set of specular paths

$$\mathcal{P}(\mathbf{x}_a, \mathbf{x}_b) = \{\tilde{\mathbf{x}}_k\}_{k=1}^{N}. \tag{4}$$

For each path $\tilde{\mathbf{x}}$ the frequency-domain attenuation is

$$\prod_{\mathbf{x}_j \in \tilde{\mathbf{x}}} \mathrm{Refl}(\mathbf{x}_j)[f]\, \mathbf{D}_{\tilde{x}}[f], \tag{5}$$

with $\mathbf{D}_{\tilde{x}}$ the source directivity. We convert this to a causal impulse via a minimum-phase transform [14]:

$$\kappa(\tilde{\mathbf{x}}, t) = \mathrm{MinPhase}\Big\{\mathbf{D}_{\tilde{x}} \circ \prod_{\mathbf{x}_j \in \tilde{\mathbf{x}}} \mathrm{Refl}(\mathbf{x}_j)\Big\}(t). \tag{6}$$

Air absorption and spherical spreading are applied with

$$\mathcal{S}_\tau\{h\}(t) = \frac{e^{-a_0\tau}}{v_{\text{sound}}\tau}\, h(t-\tau), \tag{7}$$

where $\tau$ is travel time. The room impulse response is the sum over all paths:

$$R(t) = \sum_{\tilde{\mathbf{x}}_k \in \mathcal{P}} \mathcal{S}_{\tilde{t}_k}\{\kappa(\tilde{\mathbf{x}}_k, t)\}. \tag{8}$$

This fully differentiable formulation couples beam tracing with learnable reflection responses for gradient-based optimization.

### 2.4. Multi-Scale Reflection Response

As a volumetric beam propagates, its elliptical surface footprint grows. Because the tracer returns only a hit point $\mathbf{x}$, we approximate the whole footprint by a Gaussian $\mathbf{x}' \sim \mathcal{N}(\mathbf{x}, \Sigma)$, where $\Sigma$ (derivation in Supp.) scales with path length $l$, incidence angle $\theta$, and half-aperture $\varphi$.

Following Eq. 5, we define a set of $F$ discrete key frequencies and predict their reflection magnitudes $\mathrm{Refl}\big(\gamma(\mathbf{x}, \Sigma); \Theta_2\big) \in \mathbb{R}^F$. Here $\gamma(\mathbf{x}, \Sigma)$ is the *integrated positional encoding* (IPE) from Mip-NeRF [2]:

$$\gamma(\mathbf{x}, \Sigma) = \mathbb{E}_{\mathbf{x}' \sim \mathcal{N}(\mathbf{x}, \Sigma)}\big[\gamma(\mathbf{x}')\big] = \gamma(\mathbf{x}) \circ e^{-\frac{1}{2}\operatorname{diag}(\Sigma_\gamma)}, \tag{9}$$

so the network-predicted reflections is scale-aware.

### 2.5. Multi-View Vision Feature Encoder

We guide reflection prediction with *multi-view* images via a vision encoder $\mathrm{F}(\mathbf{x}, \Sigma; \mathbf{\Theta}_2')$, so that

$$\mathrm{Refl}\big(\gamma(\mathbf{x}, \Sigma),\, \mathrm{F}(\mathbf{x}, \Sigma); \Theta_2\big) \in \mathbb{R}^F. \tag{10}$$

Given $N_s$ surface samples $\{\mathbf{z}_j\}$ on geometry $\mathcal{M}$ and $N_c$ calibrated images $\{I^{(i)}\}$, F is built in three steps:

**Per-View Featur Extraction.** Each image is encoded by a frozen backbone (*e.g.*, DINO-v2 [16]) into a feature map $W^{(i)} = \mathcal{E}(I^{(i)})$. A visible sample $\mathbf{z}_j$ is projected with known camera $P^{(i)}, \pi$ and bilinearly interpolated:

$$\mathbf{v}_j^{(i)} = \begin{cases} W^{(i)}\big(\pi(P^{(i)}\mathbf{z}_j)\big), & \text{if visible;} \\ \mathbf{0}, & \text{if occluded.} \end{cases} \tag{11}$$

**Multi-View Feature Aggregation.** Per-view features are fused per sample by a single cross-attention layer:

$$\mathbf{v}_j = \mathrm{CrossAttn}\big(Q(\mathbf{z}_j),\ KV(\{\mathbf{v}_j^{(i)}\}_{i=1}^{N}),\ M\big). \tag{12}$$

where mask $M$ is by $m_j^{(i)} = 0$ (visible) or $-\infty$ (occluded).

**Sample-Level Neighborhood Fusion.** For a query $\mathbf{x}$ we gather its $k$-nearest samples $N(\mathbf{x}) = \{(\mathbf{z}_j^*, \mathbf{v}_j^*)\}$ and apply a point-transformer [24]:

$$\mathrm{F}(\mathbf{x}, \Sigma) = \mathrm{CrossAttn}\big(Q(\mathbf{x}, \Sigma),\ KV(\{\mathbf{v}_j^*\}_{j=1}^{k})\big) \tag{13}$$

This two-level fusion (across views and local samples) delivers a geometry-, visibility-, and appearance-aware feature that drives the subsequent acoustic modules.

### 2.6. Position-Dependent Residual Component

To capture high-order reflections, diffuse reflections, diffraction, and late reverbarations, we introduce the residual component $\mathrm{r}(t; \mathbf{\Theta}_3)$ by treating every surface point $\mathbf{x} \in \mathcal{M}$ as a secondary sound sources and integrate its contribution at the listener position $\mathbf{x}_b$.

A 4-layer MLP $\epsilon$, predicts the differential time-domain response $h(t)$ per solid angle $\omega$:

$$\mathrm{h}(t) = \epsilon(\mathbf{x}, \omega, t, \mathbf{x}_a, \mathbf{p}_b; \mathbf{\Theta}_3). \tag{14}$$

The residual RIR is the integral of these responses over the unit sphere:

$$\mathrm{r}(t; \mathbf{\Theta}_3) = \int_{\mathbb{S}^2} \mathcal{S}_\tau\{\epsilon\}(\mathbf{x}'(\omega), -\omega, t; \mathbf{\Theta}_3) p_{\mathrm{u}}(\omega) d\omega \tag{15}$$

$$\approx \sum_{k=1}^{N_r} S_{\tau_k}\{\epsilon\}(\mathbf{x}_k', -\omega_k, t; \mathbf{\Theta}_3). \tag{16}$$

In Equation 15, $\mathbf{x}'(\omega)$ is the intersection of a ray in direction $\omega$ from $\mathbf{x}$ with room geometry $\mathcal{M}$, and $p_{\mathrm{u}}$ is the uniform distribution over $\mathbb{S}^2$. Equation 16 approximates Equation 15 via Monte Carlo integration with $N_r$ sampled directions $\omega_k$ from distribution $p_{\mathrm{u}}$.

| Method | Scale | RAF-Empty | | | | RAF-Furnished | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Loudness (dB) ↓ | C50 (dB) ↓ | EDT (ms) ↓ | T60 (%) ↓ | Loudness (dB) ↓ | C50 (dB) ↓ | EDT (ms) ↓ | T60 (%) ↓ |
| NAF++ [5, 12] | 1% | 6.05 | 2.10 | 94.5 | 23.9 | 6.61 | 2.10 | 74.9 | 23.0 |
| INRAS++ [5, 21] | 1% | 3.69 | 2.59 | 100.3 | 23.5 | 2.96 | 2.61 | 92.6 | 25.0 |
| AV-NeRF [10] | 1% | 3.16 | 2.52 | 96.4 | 21.8 | 2.92 | 2.64 | 96.7 | 24.5 |
| AVR [9] | 1% | 3.00 | 2.19 | 87.3 | 24.1 | 2.97 | 2.33 | 72.3 | 17.9 |
| Ours | 0.1% | 3.14 | 1.81 | 86.6 | 16.9 | 2.45 | 1.98 | 80.1 | 15.2 |
| Ours | 1% | 2.50 | 1.42 | 56.2 | 10.7 | 1.68 | 1.29 | 47.4 | 9.61 |

Table 1. Results on the Real Acoustic Field dataset [5] (0.32 s, 16 kHz). Cells highlighted in green denote the best performance, and yellow indicates the second best. Note that our model trained on only 0.1% of the data already achieves lower C50 and T60 errors than baseline methods, and significantly outperforms all baselines when using the same amount of training data.

| Method | Classroom | | | Complex Room | | | Dampened Room | | | Hallway | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Loud (dB) ↓ | C50 (dB) ↓ | T60 (%) ↓ | Loud (dB) ↓ | C50 (dB) ↓ | T60 (%)↓ | Loud (dB) ↓ | C50 (dB) ↓ | T60 (%) ↓ | Loud (dB) ↓ | C50 (dB) ↓ | T60 (%) ↓ |
| NAF++ [5, 12] | 8.27 | 1.62 | 134.0 | 4.43 | 2.25 | 44.8 | 3.88 | 4.24 | 306.9 | 8.71 | 1.36 | 21.4 |
| INRAS++ [5, 21] | 1.31 | 1.86 | 60.9 | 1.65 | 2.26 | 29.5 | 3.45 | 3.28 | 187.1 | 1.55 | 1.87 | 7.4 |
| AV-NeRF[10] | 1.51 | 1.43 | 50.0 | 2.01 | 1.88 | 36.6 | 2.40 | 3.05 | 107.9 | 1.26 | 1.03 | 9.5 |
| AVR [9] | 3.26 | 4.18 | 44.3 | 6.47 | 2.55 | 36.7 | 6.65 | 11.11 | 81.4 | 2.48 | 2.69 | 7.0 |
| Diff-RIR [23] | 2.24 | 2.42 | 39.7 | 1.75 | 2.23 | 18.5 | 1.87 | 1.56 | 44.9 | 1.32 | 3.13 | 6.8 |
| Ours | 0.99 | 1.02 | 24.3 | 0.98 | 1.44 | 10.8 | 1.11 | 1.45 | 31.9 | 0.85 | 1.15 | 6.3 |

Table 2. Results on the Hearing Anything Anywhere dataset [23] (2.0 s segments, 16 kHz), trained on 12 listener locations. Our method significantly outperforms all baseline methods in these scenes, demonstrating its effectiveness in accurately reconstructing room acoustics in few-shot settings. See Supp. for EDT error results.

# 3. Experiments

**Datasets.** We evaluate our method on two real-world datasets: the *Real Acoustic Field* (RAF) [5] dataset and the *Hearing Anything Anywhere* (HAA) [23] dataset, which are the only available real-world RIR datasets with accompanying visual capture.

**Evaluation Metrics.** Following [5, 9, 21], we evaluate perception-related energy decay patterns using *Clarity* (C50), *Early Decay Time* (EDT), and *Reverberation Time* (T60). To account for differences in overall RIR magnitude, we also adopt a loudness metric defined as:

$$\text{Loudness Error} = \left| 10 \log_{10} \left( \frac{E_{\text{pred}}}{E_{\text{gt}}} \right) \right|, \tag{17}$$

where $E = \int_0^\infty h^2(t)\,\mathrm{d}t$ is the energy of the signal $h(t)$.

## 3.1. Quantitative Results

**Results on the RAF Dataset.** To fully exploit the dense samples in RAF [5], we split the original training split (80% of all data) into 9 nested subsets ranging from 0.01% to 100% of the data (3–30k RIRs) and evaluate every model on the *unchanged* test set. As reported in Table 1, our model trained on just 0.1% of the data is comparable to state-of-the-art baselines trained on 10× more data. At equal training scales we achieve the best scores across all metrics, *e.g.*, improving Loudness by 16.6% and T60 by 50.9% in the RAF Empty scene.

**Results on the HAA Dataset.** Table 2 shows similar gains on the four real-world scenes of HAA dataset [23], where our method significantly outperforms all baseline methods. The only exception is the C50 metric in *Hallway*, where AV-NeRF exhibits particularly strong performance. This is likely due to AV-NeRF uses depth as an input, which is especially beneficial in this simple, constrained geometry.

# 4. Conclusion

We presented AV-DAR, an audio-visual differentiable pipeline for synthesizing room impulse responses (RIRs). By combining beam tracing with visually-guided reflection modeling, our approach learns RIRs from sparse real-world measurements and outperforms state-of-the-art baselines while reducing data requirements. Our work opens new possibilities for immersive AR/VR applications. As future work, we plan to extend our framework to handle multi-scene scenarios for few-shot or zero-shot reflection response prediction. We also aim to explore implicit acoustic modeling from only raw audio data, leveraging much larger corpora for training more generalizable models.

## References


[1] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 04 1979. 3

[2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844, 2021. 3

[3] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[4] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Trans. Graph.*, 35(6), Dec. 2016. 1

[5] 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, 4

[6] Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, page 21–32, New York, NY, USA, 1998. Association for Computing Machinery. 3

[7] John Kenneth Haviland and Balakrishna D. Thanedar. Monte carlo applications to acoustical field solutions. *The Journal of the Acoustical Society of America*, 54(6):1442–1448, 12 1973. 3

[8] A. Krokstad, S. Strom, and S. Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968. 3

[9] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 4

[10] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 4

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

[12] 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, 4

[13] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 1

[14] J. Gregory McDaniel and Cory L. Clarke. Interpretation and identification of minimum phase reflection coefficients. *The Journal of the Acoustical Society of America*, 110(6):3003–3010, 12 2001. 3

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

[16] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3

[17] 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 (CVPR)*, pages 27164–27175, June 2024. 1

[18] Lauri Savioja and U. Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 08 2015. 3

[19] Carl Schissler, Ravish Mehra, and Dinesh Manocha. High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *ACM Trans. Graph.*, 33(4), July 2014. 3

[20] Samuel Siltanen, Tapio Lokki, Sami Kiminki, and Lauri Savioja. The room acoustic rendering equation. *The Journal of the Acoustical Society of America*, 122:1624, 10 2007. 1

[21] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 4

[22] Dirk van Maercke and Jacques Martin. The prediction of echograms and impulse responses within the epidaure software. *Applied Acoustics*, 38(2):93–114, 1993. 3

[23] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. 1, 2, 4

[24] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 3