Neural Acoustic Context Field: Rendering Realistic Room Impulse Response With Neural Fields

Susan Liang[1], Chao Huang[1], Yapeng Tian[1], Anurag Kumar[2], Chenliang Xu[1]
[1]Unversity of Rochester [2]Meta Reality Labs Research
{sliang22, chuang65}@ur.rochester.edu, yapeng.tian@utdallas.edu,
anuragkr90@meta.com, chenliang.xu@rochester.edu

Abstract

Room impulse response (RIR), which measures the sound propagation within an environment, is critical for synthesizing high-fidelity audio for a given environment. Some prior work has proposed representing RIR as a neural field function of the sound emitter and receiver positions. However, these methods do not sufficiently consider the acoustic properties of an audio scene, leading to unsatisfactory performance. This letter proposes a novel Neural Acoustic Context Field approach, called NACF, to parameterize an audio scene by leveraging multiple acoustic contexts, such as geometry, material property, and spatial information. Experimental results show that NACF outperforms existing field-based methods by a notable margin. Please visit our project page for more qualitative results[1].

1. Introduction

High-fidelity audio is essential in creating an immersive experience, as it enhances the audience's perception and engagement [17, 3]. For example in movies and video games, realistic sound effects are necessary for depicting believable virtual environments. In "The Silence of the Lambs," for instance, the sound design reflects the terrifying prison environment, with metal doors clanging and footsteps echoing on concrete floors. Hence, generating audio with rich acoustic information is critical for the listener's perception of an environment.

Room impulse response (RIR), which characterizes the impact of room environment and emitter-receiver positions on sound propagation, is a valuable auditory function that aids in synthesizing audio with rich acoustic properties. RIR consists of the direct propagation and early reflection parts, revealing the occlusion and distance, as well as the late reverberation component, conveying the scene size and structure. By convolving an anechoic sound with the RIR signal, we can synthesize targeted audio that imitates the audio we would hear in this environment. In essence, RIR offers rich acoustic cues that enable the receiver to discern the sound source position and approximate geometry of the surroundings.

The research on RIR can be traced back to the 1970s. Krokstad *et al.* [9] and Vorländer [20] propose ray tracing algorithms to simulate sound propagation in rooms. Allen and Berkley [1] use a time-domain image expansion method for small-room acoustics simulation, and Borish [2] extends this image model to arbitrary polyhedra with any number of sides. Recently, several wave-based methods [6, 15, 7] have been proposed to calculate RIR using the wave equation. However, these methods either require expensive computation or generate RIR with limited realism [5, 18].

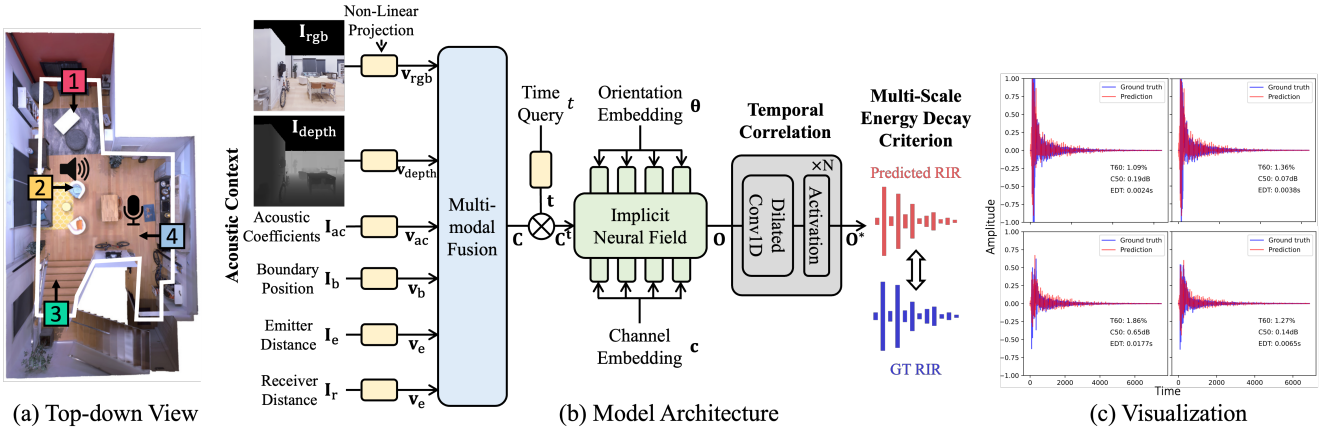
Instead, this paper proposes an effective Neural Acoustic Context Field (NACF) approach for RIR generation by implicitly representing an indoor audio scene with neural fields. Specifically, NACF learns a mapping function from the positions of the sound emitter and receiver to the desired RIR to parameterize an audio scene, akin to traditional wave field coding methods [16, 3, 17]. The learned fields can then be queried with any emitter and receiver positions of interest for RIR generation. Our inspiration comes from the visual [14, 13, 10] and auditory [11, 19] neural fields.

2. Method

2.1. Task Definition

This letter targets rendering room impulse response with implicit neural fields. Given the 2D positions of the sound emitter $\mathbf{e} \in \mathbb{R}^2$ and receiver $\mathbf{r} \in \mathbb{R}^2$, the orientation of the sound receiver $\theta \in [0°, 360°)$, and the room impulse response signal $\mathbf{O} \in \mathbb{R}^{T\times 2}$, where T is the length of the signal, and 2 is the number of channels (we use two-channel binaural sound), our goal is learning a neural field $f : (\mathbf{e}, \mathbf{r}, \theta) \rightarrow \mathbf{O}$. Afterward, we can render RIR by querying the learned field with positions and orientations of interest, including queries in the training set and unobserved (novel) queries.

[1]https://liangsusan-git.github.io/project/nacf/



(a) Top-down View (b) Model Architecture (c) Visualization

Figure 1. Method overview. The left (a) is the top-down view of an example indoor scene. We sample points evenly along the room boundary and extract various contextual information at each point, such as the RGB image, depth image, the acoustic coefficients of the surface, and several spatial information. The middle (b) is the architecture of our NACF model. First, we feed multiple acoustic contexts extracted along the room boundary (a) into the multi-modal fusion module. Then we integrate the fused contextual information with the time query as the spatial-temporal query, which is the input to the implicit neural field. After the neural field generates the RIR, we utilize a temporal correlation module to refine the RIR. Finally, we adopt the multi-scale energy decay criterion to supervise the model training. The right (c) is the visualization of predicted and ground-truth RIR together with generation errors.

2.2. Approach Overview

As shown in Fig. 1, we enhance the RIR rendering capability of neural fields through three key components: acoustic context module, temporal correlation module, and multi-scale energy decay criterion. The acoustic context module provides a comprehensive understanding of the room acoustics to the neural fields, while the temporal correlation module prevents overly smooth predictions. Additionally, the loss criterion can reinforce the energy attenuation tendency of predicted RIR at different time-frequency resolutions.

2.3. Acoustic Context

Sound propagation is mainly determined by (1) the geometry of an environment, (2) the material properties of the surface, and (3) the positions of the sound emitter and receiver. Therefore, we design an acoustic context module to encode all related acoustic information for RIR generation.

Specifically, we sample N points evenly along the room boundary for context extraction ($N = 4$ in Fig. 1 (a)). We extract the indoor depth image $\mathbf{I}_{\text{depth}} \in \mathbb{R}^{H\times W}$ and RGB image $\mathbf{I}_{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$ from each point, where H and W are the height and width of the image, respectively. The depth image $\mathbf{I}_{\text{depth}}$ depicts the geometry in the local region of each boundary point, while the RGB image $\mathbf{I}_{\text{rgb}}$ contains the semantic information of different objects and can indicate the material property of the surrounding surface. We further extract the acoustic coefficients $\mathbf{I}_{\text{ac}} \in \mathbb{R}_{+}^{P\times 3}$ of each boundary point that measures sound absorption, scattering, and transmission effects of P main frequencies. To capture the spatial information within the room, we record the position of each point $\mathbf{I}_{\text{b}} \in \mathbb{R}^2$, the distance between the emitter and each boundary point $\mathbf{I}_{\text{e}} \in \mathbb{R}^2$, and the distance between the receiver and each boundary point $\mathbf{I}_{\text{r}} \in \mathbb{R}^2$. After capturing all acoustic contexts, we embed them into latent vectors of dimension h ($\mathbf{v}_{\text{depth}}, \mathbf{v}_{\text{rgb}}, \mathbf{v}_{\text{ac}}, \mathbf{v}_{\text{b}}, \mathbf{v}_{\text{e}}, \mathbf{v}_{\text{r}} \in \mathbb{R}^h$). Consequently, we obtain six acoustic contexts from each boundary sample (Fig. 1 (b)).

We then feed all contextual information into the multi-modal fusion module for acoustic context aggregation. In detail, we use the concatenation operator to fuse acoustic contexts from different modalities and various boundary points as the holistic knowledge of room acoustics $\mathbf{C} \in \mathbb{R}^{N\times 6\times h}$:

$$\begin{aligned} \mathbf{C} &= [\mathbf{C}^1, \mathbf{C}^2, \cdots, \mathbf{C}^N] \ , \\ \mathbf{C}^i &= [\mathbf{v}^i_{\text{depth}}, \mathbf{v}^i_{\text{rgb}}, \mathbf{v}^i_{\text{ac}}, \mathbf{v}^i_{\text{b}}, \mathbf{v}^i_{\text{e}}, \mathbf{v}^i_{\text{r}}], \ 1 \le i \le N \ , \end{aligned} \tag{1}$$

where $\mathbf{C}^i \in \mathbb{R}^{6\times h}$ is the contextual information extracted from the boundary point i, N is the number of boundary points, and 6 is the number of context categories.

2.4. Implicit Neural Field

Once the acoustic context $\mathbf{C}$ has been estimated, we proceed to integrate it with the time query $t \in [1, T]$. Similar to NeRF [14], we employ positional encoding γ to project the single-value time query t to a high-dimension space $\gamma(t) \in \mathbb{R}^{2L}$, where L is the number of frequencies. Subsequently, we non-linearly embed $\gamma(t)$ into a time vector $\mathbf{t} \in \mathbb{R}^h$. Finally we calculate the dot product between the acoustic context $\mathbf{C}$ and the time vector $\mathbf{t}$, resulting in the

modified time-aware acoustic context $\mathbf{C}_t \in \mathbb{R}^{N\times 6}$:

$$\mathbf{C}_t^{(i,j)} = \mathbf{C}^{(i,j)} \cdot \mathbf{t},\ 1 \le i \le N,\ 1 \le j \le 6\ . \quad (2)$$

By incorporating the spatial context ($\mathbf{C}$) and the temporal information ($\mathbf{t}$), the modified acoustic context $\mathbf{C}_t$ can effectively serve as a spatial-temporal query for implicit neural fields.

Next, we feed $\mathbf{C}_t$ into our implicit neural field, which is instantiated as an MLP network Θ, to generate the RIR signal:

$$f_\Theta : (\mathbf{C}_t, \theta, c) \to \mathbf{O}_{t,c}\ , \quad (3)$$

where θ is the listener head orientation, c is the channel index of audio signals (c can be left or right for binaural audios), and $\mathbf{O}_{t,c}$ is the RIR of time t and channel c. To obtain the complete RIR signal $\mathbf{O}$, we query the learned neural field with all spatial-temporal queries $\mathbf{C}_{1:T}$ and all channels c.

To ease the optimization process of implicit neural fields and enhance the modeling ability, we use learnable orientation embeddings $\boldsymbol{\theta} \in \mathbb{R}^h$ and channel embeddings $\mathbf{c} \in \mathbb{R}^h$ to replace the orientation θ and the channel c, respectively. We add $\boldsymbol{\theta}$ and $\mathbf{c}$ to the input of all layers of the neural field as the orientation and channel conditions.

2.5. Multi-scale Energy Decay Criterion

Finally, we optimize NACF to predict realistic RIR with two supervision signals. Given a two-channel RIR signal $\mathbf{O}$, we use Short Time Fourier Transformation (STFT) to convert it from the time domain to the time-frequency domain and calculate its magnitude $\mathbf{M} \in {\mathbb{R}_+}^{F\times D\times 2}$, where F is the number of frequency bins, D is the number of time windows, and 2 denotes the two channels. We calculate the magnitude of ground-truth RIR $\mathbf{M}_g$ and that of predicted RIR $\mathbf{M}_p$, and measure their L1 distance as the first training objective:

$$\mathcal{L}_{\text{mag}} = ||\mathbf{M}_g - \mathbf{M}_p||_1\ . \quad (4)$$

To reinforce the energy attenuation tendency of predicted RIR, we follow Majumder *et al.* [12] using energy decay matching loss as the second training objective. Given a magnitude $\mathbf{M}$, we first compute its energy in each time window by calculating the magnitude's square and aggregating it along the frequency dimension:

$$\mathbf{M}'^{(d)} = \sum_{f=1}^{F} \left(\mathbf{M}^{(f,d)}\right)^2,\ 1 \le d \le D\ , \quad (5)$$

where $\mathbf{M}' \in \mathbb{R}_+^{D\times 2}$ represents the energy in each time window.

We then sum $\mathbf{M}'$ along the time dimension, aggregating the energy from the current step d until the end D for each time step $d \in [1, D]$ to capture the overall energy decay trend:

$$\mathbf{M}''^{(d)} = \sum_{i=d}^{D} \mathbf{M}'^{(i)},\ 1 \le d \le D\ , \quad (6)$$

where $\mathbf{M}'' \in \mathbf{R}_+^{D\times 2}$. The resulting $\mathbf{M}''$ has the same shape as $\mathbf{M}'$ since we calculate the energy sum for all time steps d. Finally, we measure the L1 distance between the ground-truth energy trend and the predicted one in the log space:

$$\mathcal{L}_{\text{dcy}} = ||\log_{10}\mathbf{M}''_g - \log_{10}\mathbf{M}''_p||_1\ . \quad (7)$$

Because both $\mathcal{L}_{\text{mag}}$ and $\mathcal{L}_{\text{dcy}}$ are computed in the time-frequency domain, their effectiveness relies on the chosen window sizes and frequency bins for STFT. Therefore, we use a set of window sizes $\{W_i\}$ and frequency bins $\{F_i\}$ to assess the prediction qualities at different scales. If we view $\mathcal{L}_{\text{mag}}$ and $\mathcal{L}_{\text{dcy}}$ as functions of window sizes and frequency bins, the overall loss can be expressed as:

$$\mathcal{L} = \sum_{i=1}^{S} (\mathcal{L}_{\text{mag}}(W_i, F_i) + \lambda \cdot \mathcal{L}_{\text{dcy}}(W_i, F_i))\ , \quad (8)$$

where λ is a hyper-parameter that controls the weight of $\mathcal{L}_{\text{dcy}}$, and S is the set size ($|\{W_i\}|$ and $|\{D_i\}|$).

3. Experiments

3.1. Experimental Settings

Dataset. We evaluate our Neural Acoustic Context Field using the SoundSpaces dataset [4]. To ensure a fair comparison, we adopt the same six representative scenes for training and evaluation, as in previous works [11, 19]: two single rooms with rectangular walls, two single rooms with non-rectangular walls, and two multi-room layouts. We maintain the same training/test split as NAF [11] with 90% data for training and 10% data for testing. For further dataset details, please refer to NAF.

Metrics. Following INRAS [19], we select three metrics, namely T60, C50, and EDT, to assess the RIR generation quality of our model. T60 measures the time it takes for the energy to decay by 60 dB. C50 captures the energy ratio between the first 50ms of RIR and the remaining portion. EDT is similar to T60 but focuses more on the early reflection of RIR. Please refer to IRNAS for details of these metrics.

3.2. Evaluation

Results. We compare our model with existing similar works, including the state-of-the-art method INRAS [19]. In line with INRAS, we include the results of traditional audio encoding methods, such as Advanced Audio Coding (AAC) [8] and Xiph Opus [21], since our model can be interpreted as an audio encoding approach. To evaluate the quality of our generated RIR, we employ the T60, C50, and

EDT metrics, where a lower score indicates better RIR quality. As depicted in Table 1, NACF outperforms all other approaches by significant margins across all metrics. Compared to INRAS, NACF reduces the T60 error by 0.78, the C50 error by 0.1 dB, and the EDT error by 0.005 sec.

Table 1. Comparison with the SOTA. We report the performance on the SoundSpaces dataset using T60, C50, and EDT metrics. A lower score indicates higher RIR generation quality.

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 [11]	3.18	1.06	0.031
INRAS [19]	3.14	0.60	0.019
NACF (Ours)	**2.36**	**0.50**	**0.014**

4. Conclusions

This letter proposes a novel method of rendering room impulse response called NACF. With the aid of the acoustic neural field, temporal correlation module, and multi-scale energy decay criterion, NACF outperforms previous work with a clear margin and sets the new state-of-the-art performance on the SoundSpaces dataset.

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, 1979.

[2] Jeffrey Borish. Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75(6):1827–1836, 1984.

[3] 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 TOG*, 39(4):44–1, 2020.

[4] 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.

[5] Thomas Funkhouser, Nicolas Tsingos, and Jean-Marc Jot. Survey of methods for modeling sound propagation in interactive virtual environment systems. *Presence: Teleoperators and Virtual Environments*, 2003.

[6] Nail A Gumerov and Ramani Duraiswami. A broadband fast multipole accelerated boundary element method for the three dimensional helmholtz equation. *The Journal of the Acoustical Society of America*, 125(1):191–205, 2009.

[7] Brian Hamilton and Stefan Bilbao. Fdtd methods for 3-d room acoustics simulation with high-order accuracy in space and time. *TASLP*, 25(11):2112–2124, 2017.

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

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

[10] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *CVPR*, pages 5521–5531, 2022.

[11] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *NeurIPS*, 2022.

[12] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In *NeurIPS*, 2022.

[13] 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 *CVPR*, pages 7210–7219, 2021.

[14] 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*, pages 405–421, 2020.

[15] Nikunj Raghuvanshi, Rahul Narain, and Ming C Lin. Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE TVCG*, 15(5):789–801, 2009.

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

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

[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, 2015.

[19] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. In *NeurIPS*, 2022.

[20] Michael Vorländer. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America*, 86(1):172–178, 1989.

[21] Xiph.Org Foundation. Xiph opus. `https://opus-codec.org/`, 2012.