MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Deep Neural Room Acoustics Primitive

He, Yuhang; Cherian, Anoop; Wichern, Gordon; Markham, Andrew

TR2024-072 June 18, 2024

Abstract

The primary objective of room acoustics is to model the intricate sound propagation dynamics from any source to receiver position within en- closed 3D spaces. These dynamics are encapsulated in the form of a 1D room impulse response (RIR). Precisely measuring RIR is diffi- cult due to the complexity of sound propagation encompassing reflection, diffraction, and absorption. In this work, we propose to learn a contin- uous neural room acoustics field that implicitly encodes all essential sound propagation primitives for each enclosed 3D space, so that we can infer the RIR corresponding to arbitrary source- receiver positions unseen in the training dataset. Our framework, dubbed DeepNeRAP, is trained in a self-supervised manner without requiring direct access to RIR ground truth that is often needed in prior methods. The key idea is to design two cooperative acoustic agents to actively probe a 3D space, one emitting and the other receiving sound at various locations. Analyzing this sound helps to inversely characterize the acoustic primitives. Our framework is well-grounded in the fundamental physical principles of sound propagation, including reciprocity and globality, and thus is acoustically interpretable and meaningful. We present experiments on both synthetic and real- world datasets, demonstrating superior quality in RIR estimation against closely related methods.

International Conference on Machine Learning (ICML) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Deep Neural Room Acoustics Primitive

Yuhang He [1] **Anoop Cherian** [2] **Gordon Wichern** [2] **Andrew Markham** [1]

Abstract

The primary objective of room acoustics is to model the intricate sound propagation dynamics from any source to receiver position within enclosed 3D spaces. These dynamics are encapsulated in the form of a 1D room impulse response (RIR). Precisely measuring RIR is difficult due to the complexity of sound propagation encompassing reflection, diffraction, and absorption. In this work, we propose to learn a continuous neural room acoustics field that implicitly encodes all essential sound propagation primitives for each enclosed 3D space, so that we can infer the RIR corresponding to arbitrary source-receiver positions unseen in the training dataset. Our framework, dubbed *DeepNeRAP*, is trained in a self-supervised manner without requiring direct access to RIR ground truth that is often needed in prior methods. The key idea is to design two cooperative acoustic agents to actively probe a 3D space, one emitting and the other receiving sound at various locations. Analyzing this sound helps to inversely characterize the acoustic primitives. Our framework is well-grounded in the fundamental physical principles of sound propagation, including reciprocity and globality, and thus is acoustically interpretable and meaningful. We present experiments on both synthetic and real-world datasets, demonstrating superior quality in RIR estimation against closely related methods.

1. Introduction

Acoustically characterizing an enclosed room scene (Kuttruff, 1979) demands estimating all the acoustic primitives related to the physical space and is key in enabling a wide range of applications, including architectural acoustics (Long, 2014), audio-based virtual and augmented reality (Verron et al., 2010; Brinkman et al., 2015) and geometric room structure estimation from audio (Kuster, 2008). One prominent primitive is to identify the sound propagation dynamics underpinning the room that models the interaction of sound waves with entities in the room during its propagation from a source position to a receiver position. Due to the nature of sound waves, this interaction between sound and the room is highly complex and is sensitive to: (i) the sound source and receiver positions, (ii) room architecture, (iii) geometric layout, including furniture placement, and (iv) material properties. Further, sound waves undergo reflection, scattering, and absorption, complicating their dynamics. All of these challenges make quantifying and measuring sound propagation primitives laborious and difficult, especially using classical methods (Szöke et al., 2019).

The acoustic effects of a room scene can be well modeled as a linear time-invariant (LTI) system (Gardner, 1998) and thus the sound propagation primitive can be expressed as a one dimensional room impulse response (RIR) function. The received signal can then be obtained by convolving the source sound with the RIR. Due to the complex nature of sound propagation as described above, the RIR is extremely nonsmooth and arbitrarily long in the time domain. Existing RIR measurement methods either require one to physically collect a discrete RIR in the room scene by sending an impulse sound (*e.g.*, starter pistol, balloon exploding, etc.) or chirp sound at one position and recording the response at another, or to approximate the RIR using geometry or wave-based approaches (Savioja & Svensson, 2015; Bilbao & Hamilton, 2017). While, the former approach based on physical measurements is inefficient and unscalable (*e.g.*, an RIR is available only at the measured locations), the latter approximation methods are computationally expensive and require detailed knowledge of the room scene acoustics. Recent works (Ratnarajah et al., 2022; Luo et al., 2022; Steinmetz et al., 2021) propose to learn RIRs with deep neural networks in a fully supervised manner, but assume access to massive RIR datasets and evaluate in small room scene (Straub et al., 2019) settings.

In this work, we take a fresh look at the sound propagation primitive estimation problem. Given the challenge in directly deriving the sound propagation primitive, we propose an indirect framework for which data is much more

Work done while YH was interning at MERL. [1] Department of Computer Science, University of Oxford, Oxford, UK [2] Mitsubishi Electric Research Labs, Cambridge, MA, US. Correspondence to: Yuhang He <yuhang.he@cs.ox.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

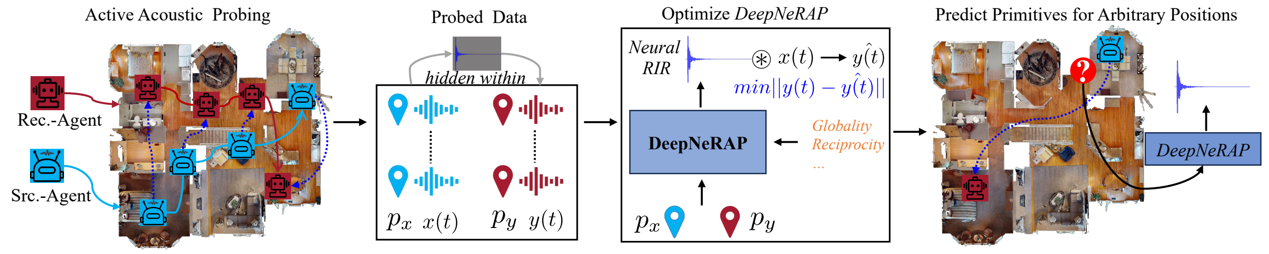

Figure 1. **DeepNeRAP Pipeline**: To learn neural sound propagation primitive, we make two agents to actively probe the room scene acoustically by emitting and receiving sounds at varied locations. Such a data collection strategy requires no prior knowledge of room acoustic properties, instead needs only the agents' position. The sound propagation primitive is implicit in the collected dataset. Our DeepNeRAP takes as input two positions and outputs a neural RIR that encodes the primitive. Convolving this neural RIR with the source sound gives the predicted receiver sound. The entire DeepNeRAP model is optimized by minimizing the discrepancy between the receiver recorded sound and predicted neural RIR effected sound. During inference, the learned DeepNeRAP can predict the RIR for any source and receiver positions.

readily accessible and is also agnostic to the room scene acoustic characteristics. Our underlying insight is that, although directly measuring the primitive governing sound propagation is difficult, the *effects* of the primitive are more easily accessible because it simply requires moving to a room scene to receive sound after propagation. Thus, by analyzing the source and receiver sounds appropriately, we can inversely estimate the propagation primitive. Motivated by this idea, we propose to use two cooperative agents that are temporally synchronized, one serving as a sound source agent and the other as sound receiver agent, to probe a room scene by moving around independently. At arbitrary agents' locations, the source sends a sound signal, which the receiver is assumed to receive. This active room scene acoustic probing strategy can be easily executed in real scenarios because it requires no prior knowledge of room scene acoustic properties and the agents' position can be easily obtained using existing and mature localization frameworks such as SLAM (Khairuddin et al., 2015). Assuming the two agents reach most of the traversable area in the scene, we can easily obtain a probing dataset to inversely learn the sound propagation primitive. We illustrate our approach in Fig. 1.

With the probed dataset, we propose a novel framework, called **Deep Ne**ural **R**oom **A**coustics **P**rimitive (DeepNeRAP), to learn sound propagation implicitly in a self-supervised manner. It takes as input two positions, viz., the source and the receiver agents' positions, and predicts the corresponding neural RIR that essentially captures the sound propagation dynamics between the two positions. During training, our *neural RIR* is convolved with the source sound (using 1D convolutions) to predict the receiver sound. DeepNeRAP is then optimized by minimizing the discrepancy between the ground truth received sound and the predicted receiver sound produced using the neural RIR. The DeepNeRAP network design incorporates fundamental room acoustics principles, including *Globality*, *Reciprocity* and *Superposition*; adhering to these principles makes DeepNeRAP primitive closer towards capturing the physics of sound propagation, while being acoustically explainable.

To empirically validate the superiority of DeepNeRAP, we conduct experiments on both synthetic and real-world datasets. For the former, we use the large scale SoundSpaces 2.0 dataset (Chen et al., 2022; Chang et al., 2017) consisting of indoor scenes with an average room area $> 100m^2$ and enriched with room acoustics. For the latter, we use the real-world MeshRIR dataset (Koyama et al., 2021). Our experiments on these datasets demonstrate state-of-the-art RIR estimation performances over closely related approaches. Below, we summarize the main contributions of this paper:

- We present DeepNeRAP to implicitly encode sound propagation primitives in a room scene. It is spatially continuous, and capable of predicting a neural RIR for arbitrary source and receiver positions.
- DeepNeRAP is trained in a self-supervised and data efficient manner, requiring neither massive RIRs nor detailed prior knowledge of room scene acoustic properties.
- The DeepNeRAP design is guided by fundamental room acoustics physical principles, resulting in the learned primitive being acoustically explainable and meaningful. We show DeepNeRAP's superiority on both synthesized large-scale room scenes and a real-world dataset.

2. Related Work

Room Scene Acoustic Characterization. Compared with vision based scene characterization that has received a significant attention in the recent years (Chang et al., 2017; Savva et al., 2019; Straub et al., 2019), the corresponding field of acoustic characterization has lagged behind. Existing acoustic characterization methods include localizing sound sources (He et al., 2021; He & Markham, 2023) and relying on the received reverberation for tasks such as speech enhancement (Zhao et al., 2017), room volume (Kuster, 2008) and sound-involved virtual reality (Brinkman et al., 2015). Closely related to DeepNeRAP, (Luo et al., 2022) propose to learn an implicit neural acoustic field from massive RIR data.

Room Acoustics models the sound propagation from one position to another in an enclosed room scene. There are primarily two classical approaches that both require access to detailed prior knowledge about room scene acoustic properties: wave-based (Bilbao & Hamilton, 2017) and geometry-based (Savioja & Svensson, 2015). Geometry-based methods rely on the wave equation, and thus the derived RIR is of high accuracy, but needs large compute, whereas geometry-based methods instead treat sound as being analogous to light rays and adopt approximation methods such as ray tracing (Krokstad et al., 1968), image source method (Allen & Berkley, 1979), beam tracing (Funkhouser et al., 2003), delay lines (De Sena et al., 2015), and acoustic radiosity (Hodgson & Nosal, 2006; Nosal et al., 2004) to compute RIRs with reduced accuracy and lower computational cost. Related to the problem of RIR interpolation (Das et al., 2021), a few recent works (Ratnarajah et al., 2022; 2021b;a; Luo et al., 2022; Richard et al., 2022; Nossier et al., 2020; Pepe et al., 2020; Ratnarajah et al., 2023; Majumder et al., 2022) propose to directly learn RIRs to characterize 3D room scenes from RIR data with deep neural networks. However, the assumption that massive RIR datasets are available does not hold in real scenarios (as detailed in Appendix A.4). There are also recent works (Kim et al., 2017; 2019) that first use vision to explicitly reconstruct 3D room scene geometry before predicting spatial audio.

Implicit Neural Representations are yet another very active research area that has seen significant progress in the recent times both in vision (Mildenhall et al., 2020; Müller et al., 2022) and in 3D shape modelling (Xu et al., 2022; Park et al., 2019; Takikawa et al., 2021), and has made initial explorations into spatial acoustics (Luo et al., 2022). Instead of directly representing the object of interest, an implicit representation tries to learn a continuous function parameterized by neural networks that is capable of representing the object at various resolutions. Our DeepNeRAP also learns an implicit spatially continuous representation, however for predicting sound propagation primitive at arbitrary source and receiver positions.

Sound Synthesis. DeepNeRAP partially relates to prior works in sound synthesis (Oord et al., 2016; Donahue et al., 2019; Engel et al., 2019; Clarke et al., 2021; Prenger et al., 2019). WaveNet (Oord et al., 2016) learns to predict future sound based on previous waveform samples. WaveGAN (Donahue et al., 2019) and GANSynth (Engel et al., 2019) adopt generative adversarial networks (Goodfellow et al., 2014) to learn to generate sound. Different from pure sound synthesis, we focus on learning a sound propagation primitive that can bring sound with spatial effects.

3. Deep Neural Room Acoustics Primitive

Given an enclosed 3D room scene in $\mathbb{R}^{3D}$, that is assumed to be a linear time invariant (LTI) system, our task is to learn an implicit deep neural room acoustics field $\boldsymbol{\mathcal{F}}_\theta$ that encodes the sound propagation primitive underlying the scene. In our case, the primitive is expressed as a one dimensional neural RIR. $\boldsymbol{\mathcal{F}}_\theta$ is spatially continuous so that it can predict the neural RIR $h(t)_{p_s \to p_r}$ for any arbitrary source position p_s and receiver position p_r. The received sound at p_r then can be derived by convolving $h(t)_{p_s \to p_r}$ with the sound at position p_s and is entirely agnostic to the sound class, i.e.,

$$h(t)_{p_s \to p_r} = \boldsymbol{\mathcal{F}}_\theta(p_s, p_r); \quad p_s, p_r \in P, \tag{1}$$

where P indicates all source or receiver (reachable) positions in the room scene, and θ represents the learnable parameters of $\boldsymbol{\mathcal{F}}$. Due to the high complexity of sound propagation dynamics, the measured RIR $h(t)$ is a highly nonsmooth and long signal (usually more than 20k points), making it difficult to collect RIR data directly to optimize $\boldsymbol{\mathcal{F}}_\theta$. Alternatively, we propose to learn $\boldsymbol{\mathcal{F}}_\theta$ in a more readily accessible way: we use two cooperative agents, one a source agent carrying an omnidirectional loudspeaker and the other a receiver agent carrying an omnidirectional microphone receiver, to actively probe the room scene independently (see Fig. 1). At each step, the source agent emits a sound $x(t)$ at one position p_x and the receiver agent receives the response sound $y(t)$ at another position p_y accordingly. This active probing strategy is practical and easy to execute in real scenarios because it requires neither detailed prior knowledge of room scene's acoustic properties nor direct collection of RIR data. Since the received sound $y(t)$ implicitly carries the room scene's sound propagation primitive conditioned on the two agents' position, we can utilize it to inversely estimate the sound propagation primitive (*e.g.,* in our case $h(t)$). That is, $\boldsymbol{\mathcal{F}}_\theta$ is learned using N samples of active probing data $\mathcal{D} = \{p_x, p_y, x(t), y(t)\}_{i=1}^{N}$,

$$\boldsymbol{\mathcal{F}}_\theta \leftarrow \boldsymbol{\mathcal{F}}_\theta(\{p_x, p_y, x(t), y(t)\}_{i=1}^{N}); \quad \{p_x\}, \{p_y\} \subseteq P. \tag{2}$$

In our setting, the source agent emits a sine sweep (Farina, 2020) sound so as to cover the whole frequency range. The agents' spatial position can be easily retrieved by either SLAM systems (Khairuddin et al., 2015) or an inertial measurement unit (IMU) with high accuracy. Our framework is "self-supervised" in the sense that $\mathcal{F}_\theta$ is learned using only the data D we collected without involving any sort of data annotation (especially the RIR data). Our self-supervision cue lies in the difference between the emitted sound and the received sound. By enforcing the predicted neural RIR effected sound to be close to the receiver agent recorded sound, we naturally encourage $\mathcal{F}_\theta$ to essentially approximate the room acoustics primitive of the room scene, that is, θ is estimated as:

$$\underset{\theta}{\operatorname{argmin}} \sum_{(p_x,p_y,x(t),y(t))\in\mathcal{D}} d\big((h(p_x,p_y) \circledast x)\,(t), y(t)\big) \qquad (3)$$
$$\text{where } h(p_x,p_y) = \mathcal{F}_\theta(p_x,p_y),$$

where h is the $\mathcal{F}_\theta$ predicted neural RIR, $\circledast$ is the 1D convolution[1] indicating the RIR effect applied on the sound (in our case, the source agent sound), and $d(\cdot)$ measures the discrepancy between the $\mathcal{F}_\theta$ predicted neural RIR effect and the ground truth observed effect (i.e., the receiver agent recorded sound); in our case, we measure the discrepancy in both frequency and time domain using the ℓ_2 loss.

3.1. LTI Room Acoustics Physical Principles

Before introducing DeepNeRAP, we present four fundamental room acoustics physical principles (Kuttruff, 1979; Rayleigh & Lindsay, 1945) that will guide the DeepNeRAP network design. In an LTI room scene, the measured RIR has to satisfy the following principles.

Principle 1, **Globality:** Unlike in computer vision where a camera captures a localized neighborhood, sound propagation relates to an entire room scene. Once emitted at the source position, sound waveform traverses isotropically[2] to interact with the whole scene before reaching the receiver position. The recorded sound thus acoustically characterizes the whole scene (Kuttruff, 1979).

Principle 2, **Reciprocity**: The reciprocity principle (Rayleigh & Lindsay, 1945; Samarasinghea & D. Abhayapala, 2017) states that in an LTI system, the RIR corresponding to the source and receiver positions is exactly the same in terms of both magnitude and phase should the source and receiver positions be swapped. Therefore, the DeepNeRAP needs to be source-receiver position permutation invariant. For example, in Eqn. 1, $h(t)_{p_s\to p_r} = h(t)_{p_r\to p_s}$.

Principle 3, **Superposition**: The superposition principle relating to room acoustics states that the RIR responses caused by more than one sound source is simply the linear combination of the response caused by each single sound source individually. Under this principle, we only need to model the neural RIR for one source and one receiver setting. The polyphonic situation where multiple sound sources are co-emitting sound can be easily derived by linearly adding individual sounds convolved with their associated propagation dynamics (neural RIR) together.

Principle 4, **Sound Independence**: This principle encompasses two key aspects. Firstly, the room acoustics primitive remains intrinsic to a room scene, irrespective of the specific sound used to probe the scene. Secondly, the neural RIR is completely sound-class agnostic. It can be universally applied to any sound to accurately capture propagation effects.

3.2. DeepNeRAP Neural Network Architecture

Following the aforementioned physical principles and the recent advances in neural implicit representations, we present the DeepNeRAP network architecture (also illustrated in Fig. 2). DeepNeRAP takes as input two spatial positions and outputs a neural RIR that implicitly encodes the sound propagation primitive for this specific position pair. Specifically, DeepNeRAP comprises five modules, which are detailed below in the sequence of data processing flow.

1. A learnable room acoustic representation $\mathcal{M}$, which is represented by a 2D spatial grid representation covering the whole room scene area. Each entry in the spatial grid is registered to a physical position in the room scene and associated with a learnable feature representation.

2. A source and receiver position pair feature extractor $\mathcal{P}$ that emphasizes both each single position's individuality and the global interaction with the whole room scene conditioned on single position.

3. An encoder $\mathcal{E}$ that learns room acoustic primitives. In our case, it is multi-layer perceptrons (MLP).

4. A neural primitive prediction decoder $\mathcal{D}$ that predicts a multi-scale neural RIR in frequency domain, where the neural RIR is represented by a complex 2D map at multiple resolutions.

5. A loss $\mathcal{L}$ that optimizes the whole neural network by minimizing the discrepancy between the neural RIR effected sound and the receiver agent recorded sound.

Room Acoustic Representation. $\mathcal{M}$ is expressed as a 2D $N \times N \times k$ spatial grid representation (rather than a

[1] In room acoustics, the received sound is obtained by convolving the source sound with the corresponding RIR in the time domain. For example, in our case, $y(t) = h(t) \circledast x(t)$.

[2] The isotropic assumption is made under the fact that the emitter is not highly directional. If the orientation of a directional emitter was also known, it would be possible to condition the network on this orientation.

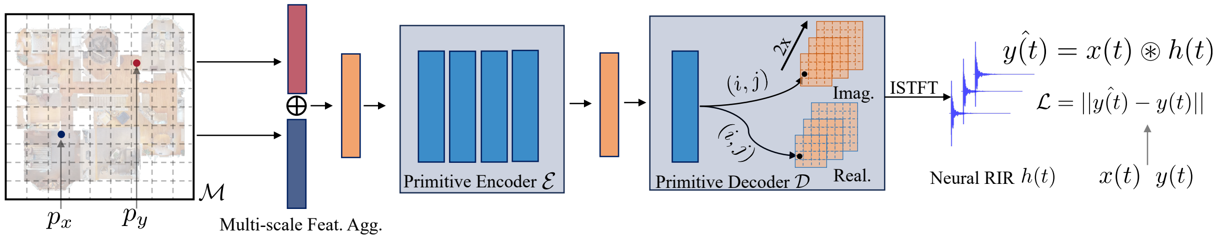
$\mathcal{M}$
p_x
p_y
Multi-scale Feat. Agg.
Primitive Encoder $\mathcal{E}$
Primitive Decoder $\mathcal{D}$
(i,j)
2π
Imag.
Real.
ISTFT
$\hat{y(t)} = x(t) \circledast h(t)$
$\mathcal{L} = ||\hat{y(t)} - y(t)||$
Neural RIR $h(t)$
$x(t)$ $y(t)$

Figure 2. **DeepNeRAP Model**: We construct a learnable spatial grid feature $\mathcal{M}$ so that the two agents' positions can be registered to the grid map. By querying features in a multi-scale manner (see Fig. 3), we relate the agent's position to the whole room scene satisfying the *globality* principle. The source and receiver position features are merged by a permutation-invariant operator `add`, satisfying the *reciprocity* principle, and then fed to the primitive encoder $\mathcal{E}$ for further refinement. The primitive decoder $\mathcal{D}$ then decodes the refined primitive features into a multi-scale neural RIR expressed in the time-frequency domain. The inverse short-time Fourier transform (ISTFT) is used to convert the neural RIR to the time domain. DeepNeRAP is optimized by minimizing the discrepancy between the neural RIR effected receiver sound and the actual recorded sound without accessing RIR data.

3D voxel grid, $\mathcal{M} \in \mathbb{R}^{N \times N \times k}$) because the two horizontal axes (topdown map) extent of most large room scenes is much larger than the vertical extent. Each entry in the grid $\mathcal{M}$ corresponds to a physical 2D position in the room scene, and is associated with a learnable small feature of size k. Such constructed grid representation is responsible for learning the sound propagation related representation underpinned by the room scene. Its grid-wise feature organization helps to learn the position-aware representation that is vital for modeling sound propagation. It is worth noting that, although $\mathcal{M}$ is geo-registered to the room scene, we do not explicitly require knowledge of the precise room geometry. We simply need to ensure that the grid map covers the whole room. Alternatively, we can construct a grid map simply covering the two agents' traversed area, obviating knowing the prior room scene size information.

Multi-Scale Position-aware Feature Extraction. The feature extraction procedure for either source position p_x or receiver position p_y should be: 1) position-aware so that the extracted features intrinsically reflect the position's uniqueness, and 2) related to the global room scene for the sake of the globality principle. To this end, we propose a multi-scale position-aware feature extraction strategy $\mathcal{P}$. Specifically, for the input position p $(= \{p_x, p_y\})$, we retrieve L-scale bounding boxes on the grid map that center at p but are of different sizes. Given a scale resolution r, the l-th bounding box's size is $l \cdot r$. By adjusting the scale resolution r and scale number L, we can correspond p to the whole grid map features. For the l-th scale bounding box, we take the four farthest grid features within the bounding box to p and further adopt bilinear interpolation to get the corresponding feature for p at scale l. By concatenating the interpolated features arising from L scales, we obtain the room acoustic representation $f(p)$ for position p,

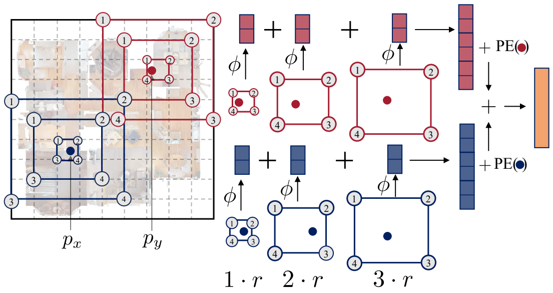

Figure 3. Mulit-scale position-aware global grid feature extraction. We just show three scales for clear visualization.

$$\begin{aligned} f(p) &= f_{l \cdot r}(p) \oplus f_{(l+1) \cdot r}(p), \\ f_{l \cdot r}(p) &= \phi(f^1_{l \cdot r}, f^2_{l \cdot r}, f^3_{l \cdot r}, f^4_{l \cdot r}); l = 1, \cdots, L-1, \end{aligned} \tag{4}$$

where $\oplus$ indicates the concatenation operation along feature dimension. $f^i_{l \cdot r}$ indicates the i-th farthest grid feature in the l-th scale bounding box ($i = 1, 2, 3, 4$). We visualize this multi-scale feature extraction strategy in Fig. 3. To reduce the computation overhead and storage cost without sacrificing the expressiveness of $\mathcal{M}$, we instantiate d with a small value but obtain a much larger and complex representation for p by concatenating small features arising from multiple scales. Such multi-scale aggregation strategy relates p to the whole room scene so it satisfies the globality principle. In addition to just querying room acoustic features (Eqn. 4), we explicitly add position encoding (Vaswani et al., 2017) to both source and the receiver positions to capture each position's uniqueness.

$$f(p_x) \leftarrow f(p_x) + \text{PE}(p_x); \ f(p_y) \leftarrow f(p_y) + \text{PE}(p_y), \tag{5}$$

where $\text{PE}(\cdot)$ is the sine/cosine position encoding. It is worth noting that the position's individuality is encoded by both the position encoding $\mathrm{PE}(\cdot)$ and the concatenation operation $\oplus$ because the concatenation operation concatenates the interpolated features from different scales in an order decided by the position (from closest to farthest). Given the two extracted features, we adopt a permutation-invariant operation (element-wise add) to merge them to get the fused feature representation so that the reciprocity principle is satisfied, *e.g.*, $f(p_{xy}) = f(p_x) + f(p_y)$.

Primitive Encoder $\boldsymbol{\mathcal{E}}$. After obtaining $f(p_{xy})$, we further adopt a multiple layer perceptron (MLP) to further encode the sound propagation essentials. In our case, the MLP consists of 6 fully-connected layers with hidden unit size of 512, batch normalization, and ReLU activation are used.

Primitive Decoder $\boldsymbol{\mathcal{D}}$ takes as input the features learned by $\boldsymbol{\mathcal{E}}$, and directly outputs a neural RIR. Due to the non-smoothness and large dimensionality of neural RIRs in the time domain, we propose to predict it in the time-frequency domain. Unlike prior works that predicts magnitude and phase maps (Luo et al., 2022) where the phase map is often chaotic, we predict real and imaginary 2D maps because the two maps are comparatively more smooth and easy to predict (see Sec. A.3). By applying inverse short time Fourier transform (ISTFT), we can convert the neural RIR from frequency domain to time domain and the conversion operation is differentiable. Specifically, $\boldsymbol{\mathcal{D}}$ combines $f(p_{xy})$ and position encoded row and column index $[\mathrm{PE}(i), \mathrm{PE}(j)]$ to predict the real and imaginary values indexed at $[i, j]$,

$$\begin{aligned} h(t) &= \mathrm{ISTFT}(H(\omega)), \\ H(\omega)[i,j] &= \boldsymbol{\mathcal{D}}(f(p_{xy}) + \mathrm{PE}(i) + \mathrm{PE}(j)), \end{aligned} \tag{6}$$

where $H(\omega)$ is the representation of $h(t)$ in the time-frequency domain, $H(\omega) = [\mathrm{Real}(\omega), \mathrm{Imag}(\omega)]$, which contains a real part map and an imaginary part map of shape $[w, h]$ ($w = h = 128$). Instead of just predicting $H(\omega)$ at one resolution, we propose to predict multiple $H(\omega)$ to express the same $h(t)$ at multiple time-frequency resolutions. We consecutively predict three $h(t)$ representations in frequency domain: $[H(\omega), H(\omega)_{2\mathrm{x}}], H(\omega)_{4\mathrm{x}}]$ by doubling the time and frequency dimension. We adopt 2D transposed convolution `TransConv` to 2× scale up (double) $H(\omega)$ resolution. $H(\omega)_{4\mathrm{x}} = \mathtt{TransConv}(H(\omega)_{2\mathrm{x}})$, $H(\omega)_{2\mathrm{x}} = \mathtt{TransConv}(H(\omega))$. By adjusting the ISTFT parameters such as hop length and window size, we get three $[h(t), h(t)_{2\mathrm{x}}, h(t)_{4\mathrm{x}}]$ from the three learned time-frequency maps, respectively. During training, we deeply supervise the three neural RIRs (He et al., 2023). In test, we merge the three neural RIRs to obtain the final neural RIR.

Loss calculator $\boldsymbol{\mathcal{L}}$ computes the discrepancy between $\hat{y}(t)$ and $y(t)$ in both time domain using the ℓ_2 loss) and frequency domain (with multi-resolution time-frequency ℓ_2 loss) (Defossez et al., 2020).

4. Experiments

Synthetic Dataset. We depend on SoundSpaces 2.0 (Chen et al., 2022) supported Matterport3D (Chang et al., 2017) dataset to collect the synthetic data. Matterport3D is a large-scale 3D indoor environment dataset with multiple rooms and complex furniture layout (with averaging size $> 100m^2$), so it contains sophisticated acoustic characteristics. We collect the dataset from all the 54 indoor scenes in the Matterport3D train set designed audio-visual navigation task (Chen et al., 2020). For each scene, we randomly sample 100 navigable probing positions covering the whole scene, these positions serve as the positions the two agents can traverse to. By randomly pairing two positions (assume the two agents stand on), we call SoundSpaces 2.0 to simulate the corresponding RIR. Convolving the RIR with sine sweep sound gets the received reberverant sound. Finally, we have obtained 4000 probing data which is further split into 3000/1000 for train/test separately by guaranteeing no position pair in the test set is close enough to any position pair in the training set. More discussion data creation is given in Sec. A.2 in Appendix.

Real-world Dataset. We adopt MeshRIR S32-M441 dataset (Koyama et al., 2021), which contains 32 source positions and 441 receiver positions (14 k data points, with 10k/4k split for train/test). The data collecting room dimension is 7.0 m × 6.4 m × 2.7 m. The sampling rate is 48kHz and RIR length is 32768 points. For this dataset, we create the sine sweep signal to match the 48 kHz sampling frequency. We predict the same RIR map size but adjusting ISTFT parameters to get longer RIR length in time domain.

4.1. Evaluation and Metrics

There are two main evaluation aspects for measuring the quality of the estimated neural RIR, namely: (i) directly comparing the predicted neural RIR with ground truth RIR, which helps to understand how well the learned primitive approximates the room acoustics primitive, and (ii) comparing the neural RIR effected sound, which helps to test how the learned primitive performs in a real acoustic environment. We adopt VCTK (Yamagishi et al., 2019) anechoic speech dataset uttered by 110 English speakers with various accents. By convolving RIR (either ground truth RIR or learned neural RIR) with the anechoic speech, we get the corresponding RIR effected (reverberant) speech sound.

Evaluation Metrics. For evaluating RIR, we incorporate six metrics, three of which are evaluations on time domain and the other three in the frequency domain. In time domain, we use: (i) **t-MSE**, measuring the difference of the predicted neural RIR and ground truth RIR in time domain with mean square error, (ii) **SDR** (signal-to-distortion ratio) in which, in accordance with the metric outlined in (Richard et al., 2022), we also report SDR to appraise the fidelity of

Table 1. Quantitative Result on Matterport3D Dataset. t-MSE: 10^{-7}, f-MSE: 10^{-2}.

Method	Neural RIR						Speech
	t-MSE (↓)	SDR (↑)	T_{60} Error (↓)	f-MSE (↓)	PSNR (↑)	SSIM (↑)	PSEQ (↑)
NAF (Luo et al., 2022)	1.01± 0.27	5.16± 0.09	7.84± 0.40	4.09 ± 0.00	15.17± 3.57	0.996± 0.00	1.40 ± 0.41
IR-MLP (Richard et al., 2022)	1.02± 0.32	4.09± 0.10	8.68± 0.12	5.68± 0.01	13.67± 3.51	0.994± 0.01	1.40 ± 0.14
S2IR-GAN (Ratnarajah et al., 2023)	1.09± 0.27	3.81± 0.10	9.19± 0.02	6.55 ± 0.12	12.98± 3.46	0.994± 0.01	1.38 ± 0.17
DeepNeRAP	**0.93 ± 0.34**	**6.62± 0.12**	**6.04± 0.08**	**1.68± 0.02**	**18.95± 3.02**	**0.998± 0.00**	**1.53 ± 0.41**

the predicted neural RIR in comparison to the ground truth RIR, and (iii) $\mathbf{T_{60}}$ **Error**, where T_{60} indicates the time to decay by 60 dB; we measure the T_{60} difference between ground truth RIR and predicted neural RIR. In frequency domain, we convert RIR to time-frequency 2D magnitude map of size 256×256, and evaluate on the magnitude map using: (i) **f-MSE**, we compute mean square error between ground truth magnitude map and predicted neural magnitude map, (ii) **PSNR** (Peak Signal-to-Noise Ratio) quantifying the quality of the magnitude map within the frequency domain (and is widely recognized for its applicability), and (iii) **SSIM** (structural similarity index measure), a perception-based metric which offers insight into the perceptual similarity between the magnitude map originating from the ground truth RIR and that derived from the predicted neural RIR. For the evaluation of sound influenced by RIR, we augment our assessment framework with the **PESQ** (perceptual evaluation of speech quality (Rix et al., 2001)) metric (using the speech convolved with the true RIR as reference) to have a human-centric perspective on perceptual similarity.

4.2. Comparison Methods

Currently there are no existing methods sharing exactly the same problem setting with our framework. Although sharing the same focus on neural RIR prediction, they largely differ in three aspects: 1. if they require ground truth RIR to train their model, 2. if they require extra prior knowledge of the room scene acoustic properties to train their model, and 3. the way to predict RIR (either in time domain or frequency domain, RIR length). For meaningful comparison, we compare with three spatial-position input based methods with appropriate modifications so as to be suitable for our setting. 1. **NAF** (Luo et al., 2022), a work that is most similar to ours. The main difference is that NAF requires access to massive RIR to train its model and it predicts binaural RIR for relatively small room scenes (Replica Dataset (Straub et al., 2019)), ours instead simply requires more readily accessible source and receiver sound. Moreover, we incorporate acoustic physical principles into our framework. We modify NAF to accept two positions and predict monoaural RIR. 2. **IR-MLP** (Richard et al., 2022), we modify it to accept two positions as input and output the neural RIR. 3. **S2IR-GAN** (Ratnarajah et al., 2023) which is an encoder-decoder architecture, we also modify it to accept two positions and output the neural RIR. More details are in Table IV.

Implementation Details. We implement DeepNeRAP in Pytorch. The detailed network architecture is illustrated in Sec. A.5 in Appendix and the source code is given in supplementary material. We train DeepNeRAP on A40 GPU with Adam optimizer (Kingma & Ba, 2015) with an initial learning rate 0.0005 but decays at every 50 epochs with decaying rate 0.5. We train all models for 300 epochs. For the comparing prior methods, we adopt their proposed training strategy. We train each model for each dataset three times independently, and report the mean and variance.

4.3. Experimental Results

The quantitative results on the synthetic Matterport3D dataset is given in Table 1 and on the real-world MeshRIR dataset is given in Table 2. From the two tables, we can clearly see that our proposed DeepNeRAP outperforms all the comparison methods across all evaluation metrics significantly and consistently. In direct neural RIR evaluation, DeepNeRAP receives much higher scores in SDR, PSNR and SSIM, and much lower score in t-MSE, f-MSE and T_{60} error. In terms of the quality of the estimated neural RIR when convolved with real-world speech, DeepNeRAP has achieved higher PESQ scores than the three competing methods. Although PESQ is an imperfect metric, as it was originally developed for quantifying audio coding artifacts, it provides evidence of the superiority of neural RIRs from DeepNeRAP while including perceptual weighting. Moreover, we have noticed all methods have achieved slightly better performance on MeshRIR dataset than on Matterport3D dataset. We hypothesize that this is due to the fact the meshRIR dataset is collected in a much smaller and simpler indoor environment and the data in MeshRIR dataset is much more densely collected than the way we used to collect on the Matterport3D dataset. Moreover, owing to the dense sampling data collection strategy in MeshRIR, we have more training data (10k) for MeshRIR than the data (3k) in Matterport3D room scenes.

We further provide qualitative visualizations of the neural predicted RIR across five room scenes in Fig. 4 (as well as in Fig. III in the Appendix). From these figures, we can see that DeepNeRAP is capable of predicting neural RIRs that best match the ground truth RIRs, even under complex room scenes and arbitrary source and receiver positions. Comparing with the other three methods, we observe that

Table 2. Quantitative Result on MeshRIR dataset. t-MSE: 10^{-8}, f-MSE: 10^{-2}.

Method	Neural RIR						Speech
	t-MSE (↓)	SDR (↑)	T_{60} Error (↓)	f-MSE (↓)	PSNR (↑)	SSIM (↑)	PSEQ (↑)
NAF (Luo et al., 2022)	2.21± 0.17	5.36± 0.12	7.44± 0.31	4.01 ± 0.02	16.19± 1.27	0.996± 0.00	1.54 ± 0.32
IR-MLP (Richard et al., 2022)	2.32± 0.12	4.22± 0.11	7.87± 0.11	5.43± 0.02	14.39± 2.41	0.995± 0.01	1.51 ± 0.20
S2IR-GAN (Ratnarajah et al., 2023)	2.55± 0.10	4.27± 0.07	8.21± 0.11	6.32 ± 0.09	13.88± 3.17	0.994± 0.01	1.41 ± 0.09
DeepNeRAP	**1.13 ± 0.20**	**7.31± 0.10**	**5.01± 0.03**	**1.27± 0.01**	**20.15± 1.01**	**0.999± 0.00**	**1.77 ± 0.30**

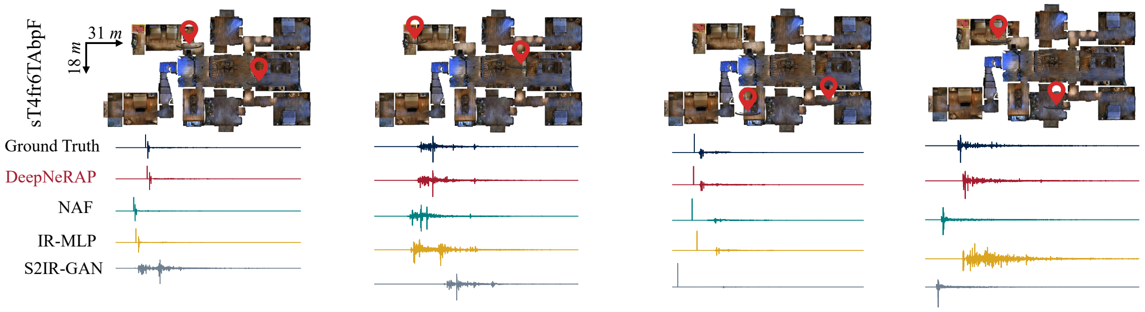

Figure 4. Vis. of learned neural RIRs in time domain on one room scene. The source/receiver position is denoted by the red position logo.

DeepNeRAP better encodes important properties of reverberation, such as the correct time delay between the source and receiver (alignment with the ground truth RIR), and the diffuse reverberation tail, which contributes most to human perception of reverberation (Traer & McDermott, 2016). The ability of DeepNeRAP to more accurately model late reflections, that contribute to diffuse tails, is confirmed by the lower T_{60} errors in Tables 1 and 2.

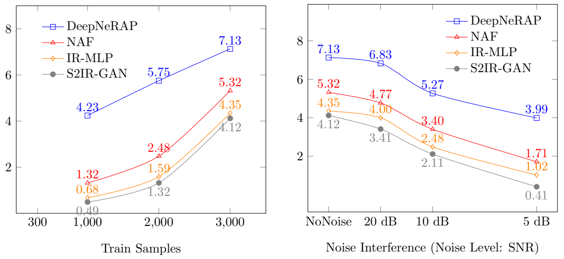

Figure 5. SDR variation against training sample (left) and noise interference test (right).

4.4. Ablation Studies

We first want to figure out the performance under different training samples or noise interference, which is important to show the robustness of DeepNeRAP. To this end, we run experiment on the scene `17DRP5sb8fy` by either varying the training samples (1,000/2,000/3,000) or adding white noise (the noise level is measured by signal-to-noise ratio, SNR, in dB). We report the SDR variation in Fig. 5, from which we can see that 1) while all methods have observed performance drop with fewer training samples, DeepNeRAP still far outperforms all other methods. It thus shows DeepNeRAP is capable of learning better neural room acoustics primitive with less data. 2) adding noise leads to performance drop, and DeepNeRAP suffers the least from the noise by significantly outperforming all other methods.

We then do six ablations on the room scene with id:`17DRP5sb8fy` and real-world MeshRIR data to assess the necessity of each component in DeepNeRAP.

1. **No Room Acoustic Feature $\mathcal{M}$ Learning**. We validate if involving a learnable grid feature is necessary; this variant is denoted $\mathcal{M}$ (DNeRAP_noRF).

2. **No Multi-Scale Feature Aggregation**. In Eqn. 4, we adopt a position-aware multi-scale feature aggregation to relate a position to the global room scene. We test one variant without multi-scale aggregation (DNeRAP_noMS).

3. **Single resolution RIR Prediction**. In Eqn. 6, we jointly predict three neural RIR maps in frequency domain. To understand the implications of this choice, we test a variant by just predicting a single resolution (of size 128 × 128) neural RIR map (DeepNeRAP_singR).

4. **No Position Encoding**. In Eqn. 5, we introduce position encoding to emphasize each position's individuality. We test one variant without position encoding (DNeRAP_noPE).

The quantitative results are given in Table V for t-MSE, SDR, and T_{60},Tables VI and VII in Appendix. The tables collectively reveal that all four ablations exhibit a decline in performance. Specifically, DNeRAP_noRF shows the a significant drop on both datasets, underscoring the need to incorporate a learnable grid. DNeRAP_noMS leads to a notable decrease highlighting the benefits of position-aware multi-scale feature aggregation. Reduced performance is also observed for DeRAP_singR and DNeRAP_noPE em-

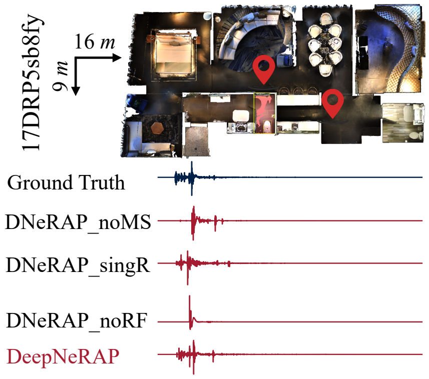

Figure 6. Three DeepNeRAP variants learned neural RIRs vis..

phasizing the necessity of multi-resolution neural RIR learning and position encoding. Visualizations of the neural RIRs for these variants in Fig. 6, clearly demonstrating their inferior quality. More results are in Appendix A.9.

5. Conclusions and Limitations

In this work, we propose a novel framework DeepNeRAP, to learn a sound propagation primitive in a self-supervised way using a data collection approach that is easy to execute. Our approach circumvents the difficulty in modeling room impulse response and we show its superiority on both synthetic data and real-world data against prior methods. The main limitation of our approach is that we assume the two probing agents can actively explore to all areas in the room within a limited step budget, which in real-world scenarios may require implementing efficient exploration algorithms.

Impact Statement

This paper introduces a novel approach aimed at enhancing the modeling of room acoustics within enclosed spaces through the collaborative exploration of two agents. Notably, our research poses no societal or ethical concerns, as our experiments are conducted using publicly synthetic and available datasets. The findings of this study hold promise for significant advancements in the realm of augmented and virtual reality (AR/VR) technologies.

References

Allen, J. B. and Berkley, D. A. Image Method for Efficiently Simulating Small-Room Acoustics. In *The Journal of the Acoustical Society of America*, 1979.

Bilbao, S. and Hamilton, B. Wave-based Room Acoustics Simulation: Explicit/Implicit Finite Volume Modeling of Viscothermal Losses and Frequency-Dependent Boundaries. *Journal of the Audio Engineering Society*, 2017.

Brinkman, W.-P., Hoekstra, A., and Vanegmond, R. The Effect Of 3D Audio And Other Audio Techniques On Virtual Reality Experience. In *Studies in Health Technology and Informatics*, 2015.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

Chen, C., Jain, U., Schissler, C., Gari, S. V. A., Al-Halah, Z., Ithapu, V. K., Robinson, P., and Grauman, K. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision (ECCV)*, 2020.

Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P. W., and Grauman, K. SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.

Clarke, S., Heravi, N., Rau, M., Gao, R., Wu, J., James, D., and Bohg, J. DiffImpact: Differentiable Rendering and Identification of Impact Sounds. In *Annual Conference on Robot Learning (CoRL)*, 2021.

Das, O., Calamia, P., and Gari, S. V. A. Room impulse response interpolation from a sparse set of measurements using a modal architecture. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, 2021.

De Sena, E., Hüseyin, H., Zoran, Cvetković, Z., and Smith, J. O. Efficient Synthesis of Room Acoustics via Scattering Delay Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2015.

Defossez, A., Synnaeve, G., and Adi, Y. Real Time Speech Enhancement in the Waveform Domain. In *Interspeech*, 2020.

Donahue, C., McAuley, J., and Puckette, M. Adversarial Audio Synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. GANSynth: Adversarial Neural Audio Synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Farina, A. Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. In *Audio Engineering Society Convention*, 2020.

Funkhouser, T., Tsingos, N., Carlbom, I., Elko, G., Sondhi, M., West, J., Pingali, G., Min, P., and Ngan, A. A Beam Tracing Method for Interactive Architectural Acoustics. *Journal of the Acoustical Society of America*, 2003.

Gardner, W. G. Reverberation Algorithms. In *Applications of digital signal processing to audio and acoustics*, pp. 85–131. Springer, 1998.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeruIPS)*, 2014.

He, Y. and Markham, A. SoundSynp: Sound Source Detection from Raw Waveforms with Multi-Scale Synperiodic Filterbanks. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*. PMLR, 25–27 Apr 2023.

He, Y., Trigoni, N., and Markham, A. SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform. In *International Conference on Machine Learning (ICML)*, 2021.

He, Y., Fang, I., Li, Y., Shah, R. B., and Feng, C. Metric-Free Exploration for Topological Mapping by Task and Motion Imitation in Feature Space. In *Proceedings of Robotics: Science and Systems*, 2023.

Hodgson, M. and Nosal, E.-M. Experimental Evaluation of Radiosity for Room Sound-Field Prediction. *Journal of the Acoustical Society of America*, 2006.

Khairuddin, A. R., Talib, M. S., and Haron, H. Review on simultaneous localization and mapping (slam). In *IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2015.

Kim, H., Remaggi, L., Jackson, P. J., Fazi, F. M., and Hilton, A. 3D Room Geometry Reconstruction Using Audio-Visual Sensors. In *International Conference on 3D Vision (3DV)*, 2017.

Kim, H., Remaggi, L., J.B. Jackson, P., and Hilton, A. Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images. In *IEEE VR*, 2019.

Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representation (ICLR)*, 2015.

Koyama, S., Nishida, T., Kimura, K., Abe, T., Ueno, N., and Brunnström, J. MESHRIR: A Dataset of Room Impulse Responses on Meshed Grid Points for Evaluating Sound Field Analysis and Synthesis Methods. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

Krokstad, A., Strøm, S., and Sorsdal, S. Calculating the Acoustical Room Response by The Use Of a Ray Tracing Technique. *Journal of Sound and Vibration*, 1968.

Kuster, M. Reliability of Estimating the Room Volume from a Single Room Impulse Response. In *Journal of the Acoustics Society of America*, 2008.

Kuttruff, H. Room Acoustics. In *Applied Science Publishers*, 1979.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Long, M. Architectural Acoustics (Second Edition). In *Architectural Acoustics (Second Edition)*, pp. xxix, Boston, 2014. Academic Press.

Luo, A., Du, Y., Tarr, M. J., Tenenbaum, J. B., Torralba, A., and Gan, C. Learning Neural Acoustic Fields. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Majumder, S., Chen, C., Al-Halah, Z., and Grauman, K. Few-Shot Audio-Visual Learning of Environment Acoustics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

Müller, T., Evans, A., Schied, C., and Keller, A. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 2022.

Nosal, E.-M., Hodgson, M., and Ashdown, I. Investigation of The Validity of Radiosity for Sound-Field Prediction in Cubic Rooms. *Journal of the Acoustical Society of America*, 2004.

Nossier, S. A., Wall, J., Moniri, M., Glackin, C., and Cannings, N. A Comparative Study of Time and Frequency Domain Approaches to Deep Learning based Speech Enhancement. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. In *arXiv:1609.03499*, 2016.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Pepe, G., Gabrielli, L., Squartini, S., and Cattani, L. Designing Audio Equalization Filters by Deep Neural Networks. *Applied Sciences*, 2020.

Prenger, R., Valle, R., and Catanzaro, B. WaveGlow: a Flow-based Generative Network for Speech Synthesis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Ratnarajah, A., Tang, Z., and Manocha, D. TS-RIR: Translated Synthetic Room Impulse Responses for Speech Augmentation. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021a.

Ratnarajah, A., Zhang, S.-X., Yu, M., Tang, Z., Manocha, D., and Yu, D. Fast-RIR: Fast Neural Diffuse Room Impulse Response Generator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Ratnarajah, A., Ananthabhotla, I., Ithapu, V. K., Hoffmann, P., Manocha, D., and Calamia, P. Towards Improved Room Impulse Response Estimation for Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Ratnarajah, A. J., Tang, Z., and Manocha, D. IR-GAN: Room impulse Response Generator for Far-field Speech Recognition. *Interspeech*, 2021b.

Rayleigh, J. W. S. and Lindsay, R. B. *The Theory of Sound*. Dover Publications, New York, 2nd Edition Revised and Enlarged edition, 1945.

Richard, A., Dodds, P., and Ithapu, V. K. Deep Impulse Responses: Estimating and Parameterizing Filters with Deep Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Rix, A., Beerends, J., Hollier, M., and Hekstra, A. Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of telephone Networks and Codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, 2001.

Samarasinghea, P. and D. Abhayapala, T. Acoustic Reciprocity: An Extension to Spherical Harmonics Domain. In *The Journal of the Acoustical Society of America*, 2017.

Savioja, L. and Svensson, U. P. Overview of Geometrical Room Acoustic Modeling Techniques. *Journal of the Acoustical Society of America*, 2015.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., and Batra, D. Habitat: A Platform for Embodied AI Research. In *International Conference on Computer Vision (ICCV)*, 2019.

Steinmetz, C. J., Ithapu, V. K., and Calamia, P. Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., Nardi, R. D., Goesele, M., Lovegrove, S., and Newcombe, R. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Szöke, I., Skácel, M., Mošner, L., Paliesek, J., and Černocký, J. Building and Evaluation of a Real Room Impulse Response Dataset. *IEEE Journal of Selected Topics in Signal Processing*, 2019.

Takikawa, T., Litalien, J., Yin, K., Kreis, K., Loop, C., Nowrouzezahrai, D., Jacobson, A., McGuire, M., and Fidler, S. Neural Geometric Level of Detail: Real-time Rendering with Implicit 3D Shapes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Traer, J. and McDermott, J. H. Statistics of Natural Reverberation Enable Perceptual Separation of Sound and Space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jone, L. J., N. Gomez, A., and Kaiser, L. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Verron, C., Aramaki, M., Kronland-Martinet, R., and Pallone, G. A 3-D Immersive Synthesizer for Environmental Sounds. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2010.

Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., and Neumann, U. Point-NeRF: Point-based Neural Radiance Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yamagishi, J., Veaux, C., and MacDonald, K. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), 2019.

Zhao, Y., Wang, Z.-Q., and Wang, D. A Two-Stage Algorithm for Noisy and Reverberant Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

A. Appendix

A.1. Matterport3D Train/Test Data Distribution Visualization

We visualize the part of the train/test data positional distribution in terms of the source and receiver position in Fig. II, from which we can clearly see that the source/receiver positions in the train and test dataset largely vary. It thus ensures the the difference between train and test dataset.

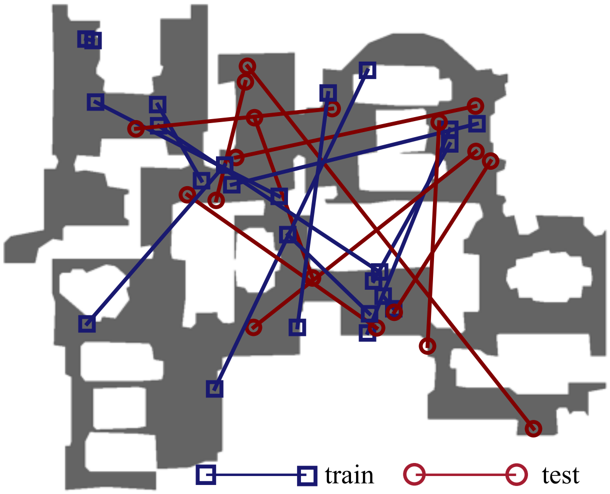

Figure II. Train/Test position pair data distribution on top of room scene `S9hNv5qa7GM` topdown map. Grey color indicates traversible area.

A.2. Matterport3D Data Synthesis Discussion

Based on SoundSpaces 2.0 (Chen et al., 2022), we can simulate the ground truth RIR for two arbitrary positions in any given Matterport3D room scene. We find that the SoundSpaces 2.0 simulated RIR does not completely satisfy *Reciprocity* principle, which means the simulated RIR will change slightly if we swap the source and receiver position. To guarantee the *Reciprocity* principle, we explicitly divide the added two RIRs, one is the RIR from source-receiver and the other from receiver-source, by 2 to get the final RIR. To get the 100 probing positions for each room scene, we iteratively call SoundSpace 2.0 (Chen et al., 2022) randomly sample navigable position API 100 times, we find such 100 probing positions are enough to cover the space's whole navigable area. We use such randomly sampled 100 probing positions to imitate the two agents' actively explored positions.

A.3. Discussion on neural RIR Representation in Frequency Domain

We represent neural RIR in frequency domain by directly predicting the real-part 2D map and imaginary-part 2D map. We choose to do so because the alternative representation of magnitude 2D map and phase 2D map is much more non-smooth and chaotic than its real- and imaginary- 2D map, resulting in the difficulty directly predicting magnitude and phase maps. We show the **mean** and **standard deviation** comparison between them in Table II. We can clearly see that the phase map has much larger standard deviation than the other three maps (high nonsmoothness). We thus choose to predict the real-part map and imaginary-part map.

Table II. Mean and standard deviation deviation comparison between the two ways representing neural RIR in frequency domain.

Metric	Real Map	Imaginary Map	Magnitude Map	Phase Map
Mean	0.000	0.000	0.002	0.012
Std.	0.005	0.005	0.007	1.763

A.4. DeepNeRAP Comparison with Existing RIR Prediction Methods

In this work, we formulate sound propagation primitive as neural RIR as we assume the room scene is linear time invariant (LTI). We have noticed there are several existing deep neural network based RIR prediction methods (Ratnarajah et al., 2022; Steinmetz et al., 2021; Ratnarajah et al., 2021b;a; De Sena et al., 2015; Ratnarajah et al., 2023; Luo et al., 2022). However, these methods differ in either problem setting or basic assumption. For example, most of these methods assume massive RIR data is accessible to train the neural network, which falls out of our assumption. We show the comparison between our framework DeepNeRAP and those relevant methods in Table III, from which we can see that all of those comparing methods require ground truth RIR data to train their own model. Some of them even require prior knowledge of the room scene's acoustic properties, such as room dimension and reverberation time. Our framework DeepNeRAP requires no RIR and is parsimonious to room scene acoustic properties.

Table III. Room acoustics modelling methods comparison. g.t. RIR means ground truth RIR, which can be either synthetic RIR or real-world collected RIR.

Methods	Need RIR ?	Network Input	More Information
NAF (Luo et al., 2022)	✓	g.t. RIR, Room Dimension	Small Room Scene
Fast-RIR (Ratnarajah et al., 2022)	✓	Position, Room Dimension; Reverb. Time	None
IR-GAN (Ratnarajah et al., 2021b)	✓	Real RIR	RIR length = 16,384
GanSynth (Engel et al., 2019)	✓	g.t. RIR	None
FiNS (Steinmetz et al., 2021)	✓	g.t. RIR, Reverb. Speech	None
IR-MLP (Richard et al., 2022)	✓	Position, Source Sound	None
S2IR-GAN (Ratnarajah et al., 2023)	✓	Reverb. Speech	RIR length = 4,096
TS-RIRGAN (Ratnarajah et al., 2021a)	✓	Synthetic RIR	RIR length = 16,384
Few-ShotRIR (Majumder et al., 2022)	✓	Synthetic RIR, RGB, Depth, Pose, Echo	RIR length = 16,000
Ours DeepNeRAP	✗	Position Pairs	RIR length = 20,001

A.5. SoundNeRAP Neural Architecture Illustration

The learnable room acoustic representation $\mathcal{M}$ consists of $500 \times 500 \times 2$, which means the grid number is 500×500, each entry associates with a learnable feature of size 2. The scale number $L = 256$ and scale resolution $r = 2.1$. The aggregated room acoustic feature representation for one position is 512.

The primitive encoder $\mathcal{E}$ network consists of 6 multi-layer perceptron (MLP) layer, each of which consists of a fully-connected layer, batch normalization layer and a ReLU activation layer. Each MLP layer's hidden unit number is 512.

The primitive decoder $\mathcal{D}$ network consists of one fully-connected layer, which fuses real/imaginary map row and column index position encoded feature (via sine/cosine position encoding) to construct an initial 2D feature map of shape $256 \times 128 \times 128$ (a 2D convolution is added to reduce the channel dimension from 512 to 256). Two learnable 3×3 2D Transposed Convolution consecutively applied to get larger 2D feature maps: one is $256 \times 256 \times 256$ and the other is $256 \times 512 \times 512$. Given the three 2D maps, another real/imaginary map prediction head neural network is used to predict multi-resolution neural RIR maps. The prediction head consists of two 3×3 2D convolutions that gradually reduce the dimension from 256 to 128, and finally to 2.

The loss calculator $\mathcal{L}$ computes the ℓ_2 loss between neural RIR effected sound and receiver agent recorded sound at both time domain and frequency domain. In frequency domain, we adopt multi-scale (in our case, three scales) frequency loss calculation strategy by converting the sound waveform to frequency domain with various sizes (by adjusting the STFT parameters appropriately).

A.6. Evaluation Metrics Definition

Given a predicted neural RIR $h(t)$ and the corresponding ground truth RIR $h\hat{(t)}$, the signal-to-distortion ratio (SDR (Richard et al., 2022)) can be defined as,

$$\text{SDR}(h(t), h\hat{(t)}) = 10 \log_{10}\left(\frac{||h\hat{(t)}||^2}{||h(t) - h\hat{(t)}||^2}\right) \tag{7}$$

A.7. More Discussion on Comparing Method

The trainable parameter number of various comparing methods is given in Table

Table IV. Comparing Methods Trainable Parameter Number Comparison

NAF (Luo et al., 2022)	12.77 M
IR-MLP (Richard et al., 2022)	4.37 M
S2IR-GAN (Ratnarajah et al., 2023)	12.78 M
DeepNeRAP	4.71 M

Table V. Ablation study on Matterport3D `17DRP5sb8fy` (t-MSE: 10^{-7}, f-MSE: 10^{-2}) room scene and real-world MeshRIR data (t-MSE: 10^{-8}, f-MSE: 10^{-2}).

Variants	Matterport3D `17DRP5sb8fy` room scene				MeshRIR Dataset			
	Neural RIR			Speech	Neural RIR			Speech
	t-MSE (↓)	SDR (↑)	T_{60} Error (↓)	PESQ (↑)	t-MSE (↓)	SDR (↑)	T_{60} Error (↓)	PESQ (↑)
DNeRAP_noRF	1.02	3.10	7.89	1.32	2.54	4.01	8.12	1.43
DNeRAP_noMS	1.01	3.14	7.90	1.36	2.50	4.00	8.13	1.40
DNeRAP_singR	0.94	4.78	7.23	1.51	2.10	5.21	6.33	1.52
DNeRAP_noPE	0.96	4.98	7.45	1.48	2.08	5.24	6.30	1.49
DeepNeRAP	**0.90**	**7.13**	**6.41**	**1.63**	**1.13**	**7.31**	**5.01**	**1.77**

Table VI. Ablation study on Matterport3D `17DRP5sb8fy` room scene. t-MSE: 10^{-7}, f-MSE: 10^{-2}

Method	Neural RIR						Speech
	t-MSE (↓)	SDR (↑)	T_{60} Error (↓)	f-MSE (↓)	PSNR (↑)	SSIM (↑)	PSEQ (↑)
NDeRAP_noRF	1.02± 0.12	3.10± 0.05	7.89± 0.20	6.85 ± 0.01	14.33± 1.22	0.993± 0.01	1.31 ± 0.21
NDeRAP_noMS	1.01± 0.32	3.14± 0.11	7.90± 0.09	6.50± 0.02	15.01± 1.14	0.994± 0.01	1.32 ± 0.21
NDeRAP_singR	0.94± 0.20	4.78± 0.08	7.23± 0.00	3.01 ± 0.09	17.90± 2.22	0.995± 0.01	1.49 ± 0.14
NDeRAP_noPE	0.96± 0.17	4.98± 0.11	7.45± 0.02	3.29 ± 0.13	18.18± 2.41	0.995± 0.01	1.51 ± 0.20
DeepNeRAP	**0.90 ± 0.12**	**7.13± 0.11**	**6.41± 0.0**	**1.70± 0.03**	**20.33± 2.01**	**0.998± 0.01**	**1.67 ± 0.21**

A.8. More Ablations

The detailed quantitative ablation study result on both Matterport3D and MeshRIR dataset are given in Table VI and Table VII respectively.

A.9. More Qualitative Result

We provide more qualitative visualization in Fig. III. We also provide bad case visualization in Fig. IV. From those bad cases, we can see that in some cases all methods inevitably predict inaccurate neural RIR, which shows designing more robust and generalized neural RIR prediction framework remains as a future challenge.

Table VII. Ablation study on MeshRIR dataset. t-MSE: 10^{-8}, f-MSE: 10^{-2}.

Method	Neural RIR						Speech
	t-MSE (↓)	SDR (↑)	T_{60} Error (↓)	f-MSE (↓)	PSNR (↑)	SSIM (↑)	PSEQ (↑)
NDeRAP_noRF	2.54± 0.12	4.01± 0.11	8.13± 0.22	6.65 ± 0.01	14.01± 2.11	0.993± 0.01	1.43 ± 0.17
NDeRAP_noMS	2.50± 0.07	4.00± 0.10	8.13± 0.09	6.33± 0.02	14.12± 1.89	0.994± 0.01	1.40 ± 0.12
NDeRAP_singR	2.10± 0.12	5.21± 0.02	6.33± 0.07	4.34 ± 0.04	18.01± 2.21	0.996± 0.00	1.52 ± 0.12
NDeRAP_noPE	2.08± 0.09	5.24± 0.03	6.30± 0.10	4.12 ± 0.06	17.11± 1.22	0.996± 0.01	1.49 ± 0.03
DeepNeRAP	**1.13 ± 0.20**	**7.31± 0.10**	**5.01± 0.03**	**1.27± 0.01**	**20.15± 1.01**	**0.999± 0.00**	**1.77 ± 0.30**

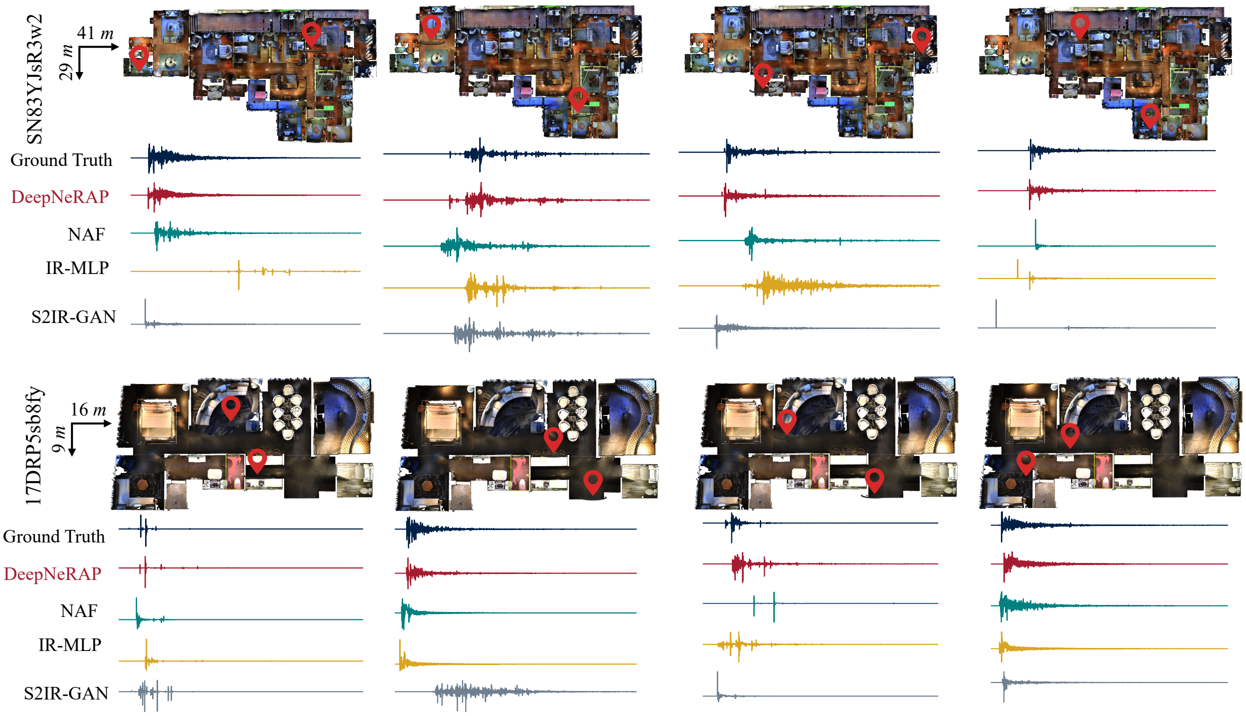

Figure III. Qualitative Result Visualization on room scene `SN83YJsR3w2` (top) and `17DRP5sb8fy` bottom. The source and receiver position are labelled by the "red position" indicator on the room scene topdown map.

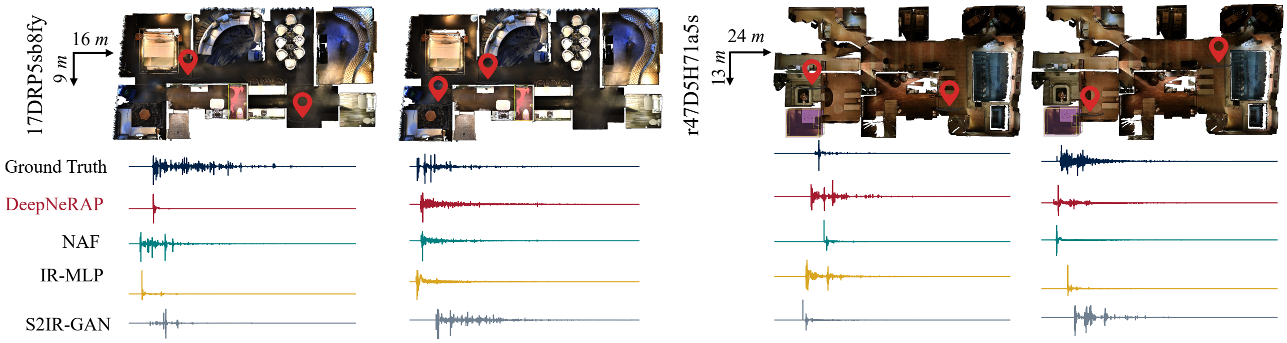

Figure IV. Qualitative Bad Case Visualization on room scene `17DRP5sb8fy` (left) and `r47D5H71a5s` (right). The source and receiver position are labelled by the "red position" indicator on the room scene topdown map.