Learning Neural Acoustic Fields

Andrew Luo [1] **Yilun Du** [2] **Michael J. Tarr** [1] **Joshua B. Tenenbaum** [2] **Antonio Torralba** [2] **Chuang Gan** [3]

Abstract

Our environment is filled with rich and dynamic acoustic information. When we walk into a cathedral, the reverberations as much as appearance inform us of the sanctuary's wide open space. Similarly, as an object moves around us, we expect the sound emitted to also exhibit this movement. While recent advances in learned implicit functions have led to increasingly higher quality representations of the visual world, there have not been commensurate advances in learning spatial auditory representations. To address this gap, we introduce Neural Acoustic Fields (NAFs), an implicit representation that captures how sounds propagate in a physical scene. By modeling acoustic propagation in a scene as a linear time-invariant system, NAFs learn to continuously map all emitter and listener location pairs to a neural impulse response function that can then be applied to arbitrary sounds. We demonstrate that the continuous nature of NAFs enables us to render spatial acoustics for a listener at an arbitrary location, and can predict sound propagation at novel locations. We further show that the representation learned by NAFs can help improve visual learning with sparse views. Finally we show that a representation informative of scene structure emerges during the learning of NAFs. Project site: andrew.cmu.edu/user/afluo/Neural_Acoustic_Fields/

1. Introduction

The sound of the ball leaving the bat, as much as its visible trajectory, tells us whether the hit is likely to be a home run or not. Our experience of the world around us is rich and multimodal, depending on integrated input from multiple sensory modalities. In particular, spatial acoustic cues provide us with a sense of the direction and distance of a sound source without needing visual confirmation, allow us to estimate the properties of a surrounding environment, and are critical to subjective realism in gaming and virtual simulations.

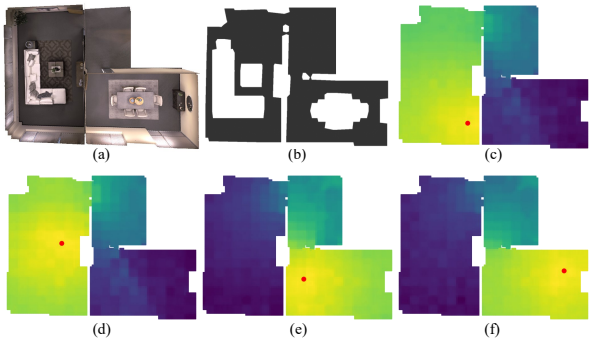

Figure 1. Neural Acoustic Field (NAF) learns an implicit representation for acoustic propagation. **(a)** A 3D top-down view of the house with two rooms. **(b)** Floor of the rooms shown in grey. **(c)-(f)** The loudness of acoustic field as predicted by our NAF is visualized for an emitter located at the red dot. Notice how sound does not leak through walls, and the portaling effect open doorways can have. Louder regions are shown in yellow.

Recent progress in implicit neural representations has enabled the construction of continuous, differentiable representations of the visual world directly from raw image observations (Sitzmann et al., 2019; Mildenhall et al., 2020; Niemeyer et al., 2020; Yariv et al., 2020). These models typically utilize a neural renderer in combination with a learned implicit representation to jointly capture and render images of a scene. By leveraging the multiview consistency between visual observations, these methods can infer images of the same scene from novel viewpoints.

However, our perception of the physical world is informed not only by our visual observations, but also by the spatial acoustics cues present in the environment. As a preliminary step in learning the acoustic properties of scenes, we explore an implicit model that represents the magnitude of the underlying impulse response of audio reverberations. As shown in Figure 1, our model can model the spatial propagation of sound in a physical scene.

Past work has explored capturing the underlying acoustics of a scene (Raghuvanshi & Snyder, 2014; 2018; Chaitanya et al., 2020). These models, however, require manually de-

[1]Carnegie Mellon University [2]Massachusetts Institute of Technology [3]MIT-IBM Watson AI Lab.

signing acoustic functions which, critically, prevent such approaches from being applied to arbitrary scenes. In this work, we extend this approach by constructing an implicit neural representation which captures, in a *generic manner*, the underlying acoustics of a scene. In particular, following (Raghuvanshi & Snyder, 2014), we define the acoustic modeling problem as modeling the impulse-response a listener receives given a sound emitted at an emitter location (as illustrated in Figure 1) and across all possible emitter-listener pairs in a scene. This learned representation inherently captures the pattern of all acoustic reverberations in a scene. Because a complete field of acoustic reverberations of a complex scene can be hundreds of gigabytes in size, the compact and continuous nature of our learned representation enables useful virtual reality and gaming applications.

Learning a representation of scene acoustics poses several challenges compared to the visual setting. First, how do we generate plausible audio impulse responses at each emitter-listener position? While we may represent the visual appearance of a scene with an underlying three-dimensional vector, an acoustic reverberation (represented as an impulse response) can consist of over 20000 values and, thus, is significantly harder to capture. Second, how do we learn an acoustic neural representation that densely generalizes to novel emitter-listener locations? In the visual setting, ray-tracing can enforce view consistency across large portions of a visual scene (modulo occlusions). While in principle, in a similar manner, we may reflect acoustic "rays" in our scene to obtain an impulse response, a intractable number of rays are necessary to obtain the desired representation.

To address both challenges, we propose Neural Acoustic Fields (NAFs). To capture the complex signal representation of impulse responses, NAFs encode and represent an impulse-response in the Fourier frequency domain. Motivated by the strong influence of nearby geometry on anisotropic reflections (Raghuvanshi & Snyder, 2018), we propose to condition NAFs on local geometric information present at both the listener and emitter locations when decoding the impulse response. In our framework, local geometric information is learned directly from impulse responses. Such a decomposition facilitates the transfer of local information captured from training emitter-listener pairs to novel combinations of emitters and listeners

By modeling the dense acoustic fields of an environment, NAFs learns a useful representation that enables us to extract structural information about the scene. In the cross-modal setting, we demonstrate how the learned acoustic structure can be utilized to aid learned visual representations, improving novel view synthesis. Further, by directly utilizing the learned latent representation in our NAFs, we demonstrate how one can infer the structure of a scene.

In summary, we present Neural Acoustic Fields (NAFs), a neural implicit field which captures the underlying acoustics of a scene in a compact and spatially continuous fashion. We show that NAFs are able to outperform baselines in modeling scene acoustics, and provide detailed analysis of the design choices in NAFs. We further illustrate how the structure learned by NAFs can improve cross-modal generation of novel visual views of a scene. Finally, we illustrate how the learned representation of NAFs enable the downstream application of inferring scene structure.

2. Related Work

Audio Field Coding There is a rich history of encoding methods for 3D spatial audio. These approaches largely fall into two categories. The first approach encodes the sound field at a user-centric location by capturing the sound from spatially distributed sources (Gerzon, 1973; Breebaart et al., 2005; Pulkki, 2007; Richard et al., 2021). While they may leverage perceptual cues to create the sense of spatial audio, they do not allow the listener to freely traverse the scene and experience sound from different locations. The second approach aims to model the sound heard as a listener moves in a scene (Raghuvanshi & Snyder, 2014; Mehra et al., 2014; Raghuvanshi & Snyder, 2018; Chaitanya et al., 2020; Ratnarajah et al., 2021). Since the complete acoustic field of a scene is computationally prohibitive to simulate in real time, and expensive to store in full fidelity, these methods have relied on a handcrafted encoding of the acoustic field, prioritizing efficiency above reproduction fidelity. Our work allows a listener to move and experience sounds that come from anywhere in a scene, and can represent the acoustic field continuously at high fidelity by directly learning from data.

Implicit representations Our approach towards modeling the underlying acoustics a scene relies on the use of a neural implicit representations. Implicit representations have emerged as a promising representation of 3D geometry (Niemeyer et al., 2019; Chen & Zhang, 2019; Park et al., 2019; Saito et al., 2019) and appearance (Sitzmann et al., 2019; Mildenhall et al., 2020; Niemeyer et al., 2020; Yariv et al., 2020) of a scene. Compared to traditional discrete representations, implicit representations are a continuous mapping capable of capturing data at an "infinite resolution". (Jiang et al., 2020) proposed a grid based representation for implicit scene reconstruction, while more recently (DeVries et al., 2021) has adopted spatial conditioning for 3D image synthesis, where in both settings, the grid enables a higher-fidelity encoding of the scene. Our work also leverages local grids to model acoustics, but as an inductive bias and way to generalize to novel inputs.

Audio-Visual Learning Our work is also closely related to joint modeling of vision and audio. By leveraging the correspondence between vision and audio, work has been done to learn unsupervised video and audio representations

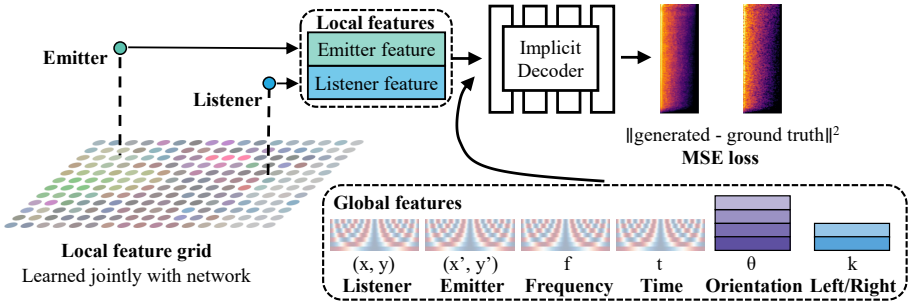

Figure 2. Overview of our NAF architecture. Given a listener position and an emitter location, we first query a grid for local features which are learned together with the network during training. We compute the sinusoidal embedding of the positions, frequency, and time, and query a discrete embedding matrix using the orientation and left/right ear. These features are fed to an implicit decoder. Our method is trained with a MSE loss with impulse responses.

(Aytar et al., 2016; Arandjelovic & Zisserman, 2017), localize objects that emit sound (Senocak et al., 2018; Zhao et al., 2018; 2019; Gan et al., 2020a), and jointly use vision and audio for navigation (Chen et al., 2020; Gan et al., 2020b; 2021). Recent work aims to propose plausible reverberations or sounds from image input (Singh et al., 2021; Du et al., 2021), these approaches model the phase-free log-magnitude STFT using either convolution or implicit functions, which we also utilize. Different from them, our work leverages the geometric features learned by modeling acoustic fields to improve the learning of 3D view generation.

3. Methods

We are interested in learning a generic acoustic representation of an arbitrary scene, which can capture the underlying sound propagation of arbitrary sound sources across both seen and unseen locations in a scene. We first review relevant background information towards modeling environment reverberations. We then describe Neural Acoustic Fields (NAFs), a neural field which we show can capture, in a generic manner, the acoustics of arbitrary scenes. We further discuss how we can parameterize NAF in a manner so that it can capture acoustics property even at unseen sound sources and listener positions. Finally, we discuss the underlying implementation details of our model.

3.1. Background on Environmental Reverberation

The sound emitted by a sound source undergoes decay, occlusion, and scattering due to both the geometric and material properties of a scene. For a fixed location pair $(\boldsymbol{q}, \boldsymbol{q}')$, we define the impulse-response at a listener position $\boldsymbol{q}$, as the sound pressure $p(t; \boldsymbol{q}, \boldsymbol{q}')$ induced by an impulse at $\boldsymbol{q}'$. Such behavior can be concisely and elegantly modeled utilizing the linear wave equation (Pierce, 2019):

$$\left[\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right] p(t, \boldsymbol{q}, \boldsymbol{q}') = \delta(t)\delta(\boldsymbol{q} - \boldsymbol{q}'), \quad (1)$$

where c is the speed of sound, p is the sound pressure, (q, q') being the listener and emitter location respectively, and δ the Dirac delta representing the forcing function, where we refer to sound pressure $p(t; \boldsymbol{q}, \boldsymbol{q}')$ as the impulse-response observed at listener position $\boldsymbol{q}$.

Given an accurate model of the impulse-response $p(t; \boldsymbol{q}, \boldsymbol{q}')$ described in Eqn (1), we may model audio reverberation of any sound waveform $s(t)$ emitted at $\boldsymbol{q}'$, by computing the response $r(t, \boldsymbol{q}, \boldsymbol{q}')$ at listener location $\boldsymbol{q}$ by querying the continuous field and using temporal convolution:

$$r(t; \boldsymbol{q}, \boldsymbol{q}') = s(t) \circledast p(t; \boldsymbol{q}, \boldsymbol{q}') \quad (2)$$

3.2. Neural Acoustic Fields

We are interested in constructing a continuous representation of the underlying acoustics of a scene, which may specify the reverberation patterns of an arbitrary sound source. The parameterization of an impulse-response introduced in Section 3.1 provides us with a method to model audio propagation when given an omnidirectional listener and emitter. To construct a model of a directional listener, we need to further model the 3D head orientation $\theta \in \mathbb{R}^2$, and ear $k \in \{0, 1\}$ (binary left or right) of a listener, in addition to the spatial position $\boldsymbol{q} \in \mathbb{R}^3$ of the listener and $\boldsymbol{q}' \in \mathbb{R}^3$ of the emitter.

We may then model the time domain impulse response $\boldsymbol{v}$ using a neural field Φ which takes as input the listener and emitter parameters:

$$\Phi : \mathbb{R}^8 \times \{0, 1\} \rightarrow \mathbb{R}^T$$
$$(\boldsymbol{q}, \theta, k, \boldsymbol{q}') \rightarrow \Phi(\boldsymbol{q}, \theta, k, \boldsymbol{q}') = \boldsymbol{v} \quad (3)$$

Directly outputting the impulse-response $\boldsymbol{v} \in \mathbb{R}^T$ with a neural network is difficult die to its high dimensional (over 20000 elements) and chaotic nature. A naïve solution would be further add t as an additional argument our neural field,

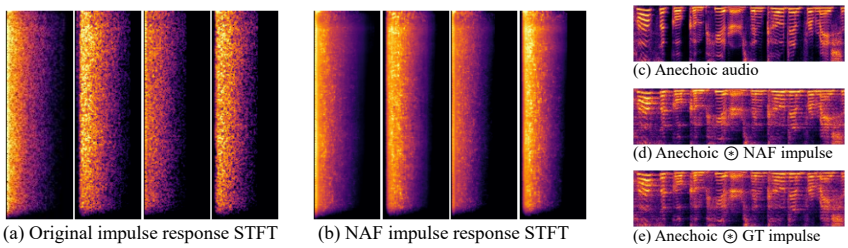

Figure 3. **Qualitative Visualization of Test Set Impulse Response Prediction.** Example log-STFT of impulse responses and predictions from NAF are shown in **(a)** and **(b)**. **(c)** shows the log-STFT of an anechoic audio (without any reverberation). **(d)** The sound with a reverberation impulse response from our NAF. **(e)** The sound with the ground truth reverberation impulse response applied.

but we found that such a solution worked poorly, due to the highly non-smooth representation of the waveform (see Table A2). We instead encode the impulse-response utilizing a short-time Fourier transform (STFT) to create a log-magnitude spectrogram denoted v_{STFT}, which we find to be significantly more amenable to neural network prediction, due to the smooth nature of the frequency space. In Figure 3 we show spectrograms for ground truth impulse responses and those learned by our network.

Thus, our parameterization of NAF is a neural field Ω that is trained to estimate the impulse response function ϕ, and outputs $\boldsymbol{v}_{\text{STFT}}$ for a given time and frequency coordinate:

$$\Omega : \mathbb{R}^{10} \times \{0, 1\} \rightarrow \mathbb{R}$$
$$(\boldsymbol{q}, \theta, k, \boldsymbol{q}', t, f) \rightarrow \Omega(\boldsymbol{q}, \theta, k, \boldsymbol{q}', t, f) \approx \boldsymbol{v}_{\text{STFT}}(t, f) \quad (4)$$

We train our model using MSE loss between the generated and ground truth log-spectrograms $\boldsymbol{v}_{\text{STFT}}$:

$$\mathcal{L}_{\text{NAF}} = \|\Omega(\boldsymbol{q}, \theta, k, \boldsymbol{q}', t, f) - \boldsymbol{v}_{\text{STFT}}(t, f)\|^2 \quad (5)$$

across spectrogram coordinates t and f.

The phase generally contains no correlation in a spatial impulse response. We follow prior work in utilizing random phase for waveform reconstruction with learned log-magnitude spectrograms (Singh et al., 2021). Phase free models are typical in spatial acoustic modeling utilizing both handcrafted and learned approaches (Pulkki, 2007; Raghuvanshi & Snyder, 2018; Singh et al., 2021).

3.3. Generalization through Local Geometric Conditioning

We are interested in parameterizing the underlying acoustic field, so that we may not only accurately represent impulse-response at emitter-listener pairs we see during training, but also at novel combinations of emitter and listener seen at test time. Such generalization may be problematic when directly parameterizing NAFs utilizing a MLP with inputs specified in Eqn (4), as the network may learn to directly overfit and entangle the relation between emitter and listener impulse-responses.

What generic information may we extract from a given impulse-response between an emitter and listener? In principle, extracting the full dense geometric information in a scene would enable us to robustly generalize to new emitter and listener locations. However, the amount of geometric information available in a particular impulse-response, especially for positions far away from either current emitter and listener is limited, since these positions have little impact on the underlying impulse-response. In contrast, the local geometry near either emitter and listener positions will have a strong influence in the impulse-response, as much of the anisotropic reflection comes from such geometry (Paasonen et al., 2017). Inspired by this observation, we aim to capture and utilize local geometric information, near either emitter or listener locations, as a means to predict impulse-responses across novel combinations.

To parameterize and represent these local geometric features, we learn a 2D grid of of spatial latents which we illustrate in Figure 2. When predicting an impulse-response at a given emitter and offset position, we query the learned grid features at both emitter and listener positions, and provide it as additional context into our NAF network Ω . Such features provide rich information on the impulse-response, enabling NAF to generalize better to unseen combinations of both emitter and listener locations. In the rest of this work, we refer to the NAFs with local geometric features as Ω_{grid}. We learn grid latent features jointly with the underlying parameters of NAF. Additional details can be found in Appendix B.

Such a design choice, however, still requires us to consider how to further combine local geometric information captured separately from either listeners or emitters. A naïve implementation would be to maintain separate feature grids for both listener and emitter positions. Such an approach fails to account for the fact that the local geometric information captured by emitter may also inform the local geometric information around a listener. Examining Eqn (1), we note that it is in fact symmetric with respect to exchanging either listener or emitter positions (Chaitanya et al., 2020), indicating that the impulse-response does not change when omnidirectional listener and emitters are swapped. Such a result means that we may in fact utilize the local geometric information captured near an emitter position interchangeably for either emitters and listeners. Thus, we propose our

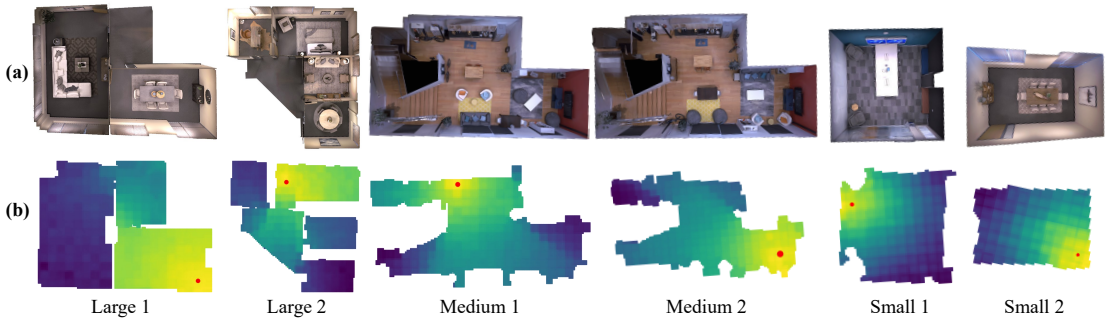

Figure 4. **Qualitative Visualization of Neural Acoustic Fields. (a)** Top down view of the rooms. **(b)** Results as inferred by our neural acoustic field. Loudness of a sound given a emitter location indicated in red, lighter color indicates louder sound. Note how openings and walls lead to portaling and occlusion of the sound respectively.

local geometric information as a single latent grid, which we show to outperform the naïve dual grid implementation.

4. Experiments

In this section, we demonstrate that our model can faithfully represent the acoustic impulse response at seen and unseen locations. Additional ablation studies verify the importance of utilizing local geometric features to enable test time generation fidelity. Next, we demonstrate that learning acoustic fields could facilitate improved visual representations when training images are sparse. Finally we show that the learned NAF can be used to infer scene structure.

4.1. Setup

For evaluating the learned acoustic fields, we use the Soundspaces dataset (Chen et al., 2020). This dataset consists of R_i probe points for each scene, with each probe capable of representing an emitter or listener location for up to R_i^2 emitter and listener pairs. The emitters are represented as omnidirectional, while the listener acts as a stereo receiver that can have one of four different orientations. The listeners and emitters are at fixed height. For each scene, we holdout 10% of the RIRs randomly as a test set. Our NAFs are trained on 6 representative scenes, selected such that 2 consist of multi-room layouts, 2 consist a single room with a non-rectangular walls, and 2 consist of a single room with rectangular walls as in Figure 4. Each scene is trained for 200 epochs, which takes around 12 hours for the largest scenes on four `Nvidia V100s`. In each batch, we sample 20 impulse responses, and randomly select $2,000$ frequency & time pairs within each spectrogram. An initial learning rate of 5×10^{-4} is used for the network and the grid features. We add a small amount of noise sampled from $\mathcal{N}(0, 0.1)$ to each coordinate during training to prevent degenerate solutions. We visualize the six scenes and the results as inferred by our NAFs in Figure 4.

4.2. Architecture Details

The Soundspaces dataset lacks the full parameterization of an acoustic field described in Equation 4, so we train NAF with a restricted parameterization that is available in the dataset. This allows for two degrees of freedom along the $x - y$ plane for the listener locations $q \in \mathbb{R}^2$ and the emitter location $q' \in \mathbb{R}^2$. The listener can assume four possible orientations $\theta \in \{0, 90, 180, 270\}$, while the emitter is omnidirectional. In particular, we utilize a parameterization of Ω_{grid} which maps an input tuple $[x, y, x', y', f, t] \in \mathbb{R}^6 \times \{0, 90, 180, 270\} \times \{0, 1\}$ to a single scalar value that represents the intensity for a given time and frequency in the STFT:

$$\Omega_{\text{grid}}(x, y, \theta, k, x', y', t, f) \Rightarrow \boldsymbol{v}_{\text{STFT}}(t, f) \tag{6}$$

To encode the rotation θ, as there are only 4 possible discrete rotations in the dataset, we directly query into a learnable embedding matrix of shape $\mathbb{R}^{4\times k}$, returning a $\mathbb{R}^{1\times k}$ vector. Similarily, to encode the left and right ear, we similarly query into a learnable embedding matrix of shape $\mathbb{R}^{2\times k}$, returning a $\mathbb{R}^{1\times k}$ vector. The f, t tuple representing the frequency and time respectively are scaled to $(-1, 1)$ and processed with sinusoidal encoding using 10 frequencies of $\sin$ and $\cos$. To obtain local geometric features for either a emitter or listener in a scene, we assume that our scene is contained within a set of pixels $\mathcal{P} = \{P_1 ... P_k\}$ which form a grid over the scene. For a given position tuple (x, y) as query location, we then interpolate the local features. Where $\mathcal{L}(\cdot)$ is the interpolation function. $(p_1^* \dots p_k^*)$ are the set of all pixel that form the grid, and $\tilde{f}(\cdot)$ represents the features stored at a given pixel:

$$(x, y) \Rightarrow \mathcal{L}(x, y; \tilde{f}(p_1^*), \dots \tilde{f}(p_k^*)) \tag{7}$$

$$= \sum_{i=0}^{k} w_i \tilde{f}(p_i^*) \tag{8}$$

w_i is determined by a Nadaraya-Watson estimator with a gaussian weighting kernel applied to the distance between

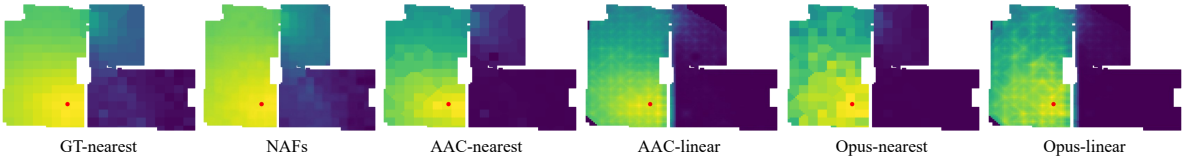

Figure 5. **Comparison of the acoustic fields.** From left to right, we visualize the loudness maps generated by the full ground truth, our NAFs, and by AAC or Opus coding combined with linear and nearest neighbor interpolation on the training set. Emitter location shown in red. Our method can faithfully reproduce the loudness map present in the ground truth.

	Large 1		Large 2		Medium 1		Medium 2		Small 1		Small 2		Mean	
Model	Spectral↓	T60↓	Spectral↓	T60↓	Spectral↓	T60↓	Spectral↓	T60↓	Spectral↓	T60↓	Spectral↓	T60↓	Spectral↓	T60↓
AAC-nearest	1.913	9.996	1.989	13.31	2.111	6.148	2.122	6.051	2.296	9.798	2.509	5.809	2.156	8.519
AAC-linear	1.904	8.847	1.964	11.63	2.105	4.585	2.116	4.422	2.299	8.253	2.521	6.021	2.151	7.293
Opus-nearest	1.740	12.20	1.817	15.15	1.887	7.875	1.898	7.897	2.058	10.68	2.238	7.564	1.940	10.23
Opus-linear	1.780	11.30	1.827	13.55	1.922	6.710	1.934	6.917	2.097	9.116	2.284	6.981	1.974	9.096
NAF (Dual)	0.415	3.286	0.422	4.001	0.386	2.729	0.387	**2.446**	0.364	**2.758**	0.371	2.578	0.391	2.966
NAF (Shared)	**0.406**	**2.872**	**0.413**	**3.351**	**0.382**	**2.472**	**0.383**	2.541	**0.354**	2.854	**0.341**	**2.240**	**0.380**	**2.722**

Table 1. **Quantitative Results on Test Set Accuracy.** We report the spectral loss between generated and ground truth log spectrograms across methods, as well as the percentage (%) difference for the T60 reverberation time. The best method for each room is **bolded**. For the nearest and linear baselines, we perform interpolation in the time domain using samples from the training set.

query and grid coordinates:

$$w_i = K((x,y),(x_i,y_i)) / \sum_{j=1}^{k} K((x,y),(x_k,y_k)) \quad (9)$$

$$K(\boldsymbol{x},\boldsymbol{x}') = \exp(-\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2/2\sigma^2) \quad (10)$$

Because this interpolation function is differentiable, we jointly learn the grid features during training. These queried features are combined with the coordinates processed with sinusoidal encoding using 10 frequencies of sin and cos functions. We process both the listener and emitter position tuples this way. We combine the grid based features with the sinusoidal embeddings and the discrete indexed embeddings as the input to our multilayer perceptron f_ϕ. Please refer to Figure 2 for a visualization of our model, and Appendix B for further details. We compare using a shared local geometric feature with the emitter and listener, as well as using have the emitter and listener query their own individual grids.

4.3. Evaluation on neural acoustic fields

We first validate that we can capture environmental acoustics at unseen emitter-listener positions.

Baselines. We compare our model against two widely used high performance audio coding methods: Advanced Audio Coding (AAC) and Xiph Opus. For each method, we apply both linear and nearest neighbor interpolation to the coded acoustic fields. Both linear and nearest neighbor approaches are widely used (Savioja et al., 1999; Raghuvanshi et al., 2010; Pörschmann et al., 2020) in modeling of spatial audio. We also compare sharing and using individual local geometric features in our NAFs. Each method is provided with the same train-test split. We visualize the acoustic fields produced by each method in Figure 5. For additional details please see section E of the appendix.

Method	Storage (MiB)
AAC	346.74
Opus	181.37
NAF (Shared)	8.41

Table 2. **Average space consumption across six scenes.** Our NAF can compactly encode the spatial acoustic field at a fraction of the storage size.

Results. We evaluate the results of our synthesis by measuring the spectral loss (Défossez et al., 2018) between the generated and the ground truth log-spectrograms, as well as measuring the percentage error between the T60 reverberation time in the time domain. In this case, lower spectral loss and T60-error values indicates a better result. As shown in Table 1, our NAFs achieve significantly higher quality on the modeling of unseen impulse responses compared to strong interpolation baselines. A comparison of using shared and dual local geometric features indicates that despite having fewer learnable parameters, we achieve better performance by sharing the local geometric features. Examples of individual impulse responses generated by our model are shown in Figure 3. Figures 4 shows the different scenes and the loudness change predicted by our NAFs. The size of a spatial acoustic field is important for real life applications. In Table 2 we show that on average our method uses a magnitude less space than the baseline methods. Our model is capable of predicting smoothly varying acoustic fields that are affected by the physical surroundings.

Generalization through Geometric Conditioning. We

	Large Room 1						Large Room 2					
	PSNR ↑			MSE ↓			PSNR ↑			MSE ↓		
Training Images	75	100	150	75	100	150	75	100	150	75	100	150
NeRF	25.41	27.36	29.85	6.618	3.506	1.740	25.70	27.74	29.34	6.921	3.905	2.185
NeRF + NAF	**26.19**	**27.59**	**29.90**	**5.209**	**2.983**	**1.625**	**26.24**	**28.22**	**29.45**	**5.641**	**3.075**	**2.034**

Table 3. **Quantitative Results on Cross-Modal Image Learning.** Quantitative results on joint training of NeRF and NAF jointly conditioned on a single local grid. We use very sparse training images in highly complex scenes. When evaluated on 50 test images, we observe that cross-modal learning helps improve PSNR when the visual training data is more sparse. MSE results are multiplied by 10^3.

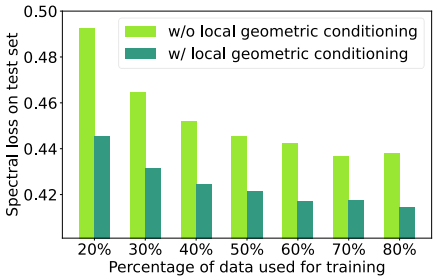

Figure 6. **Local Geometric Conditioning.** Comparison of NAF with and without local geometric conditioning trained with different amounts of data.

next assess the impact of utilizing local geometric conditioning as a means to generalize to novel combinations of emitter-listener positions. On the "Large 1" room, in Figure 6 we evaluate test set spectral error when NAF is trained with a limited percentage of the training data either with or without local geometric conditioning. We find that such geometric conditioning enables better test set reconstruction error, with the performance gap increasing with less data.

4.4. Cross-modal learning

In this experiment, we explore the effect of jointly learning acoustics and visual information when we are given sparse visual information. Recall that our NAF includes a local geometric feature grid $\mathcal{P}$ that covers the entire scene. For our cross-modal learning experiment, we jointly learn this feature grid with a NeRF network modified to accept both local features along with the traditional sinusoidal embedding. In the acoustics branch, we query the grid using emitter and listener positions. In the NeRF branch, we use point samples along the ray projected on the grid plane to query the features. In both cases, the process is fully differentiable. We use a standard implementation of NeRF with a coarse and fine network. In both the cross-modal and RGB only experiments we augment the fine network with a learnable local feature grid. In the NeRF only setting, we minimize color C reconstruction loss for a ray r over a batch of rays $\mathcal{R}$: $\mathcal{L}_{\text{RGB}} = \sum_{r \in \mathcal{R}} ||\hat{C}(r) - C(r)||_2^2$. In contrast, in the NAF + NeRF experiment, we jointly minimize $\mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{NAF}}$, where $\mathcal{L}_{\text{NAF}}$ is defined in equation 5. We utilize 64 coarse samples and 128 fine samples for each ray, and sample 1024 rays per batch.

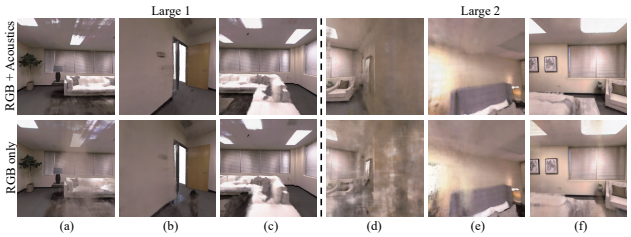

Figure 7. **Qualitative Visualization of Cross-Modal Image Learning.** Qualitative comparison between NeRF learned jointly with a NAF with RGB and acoustic supervision, and NeRF learned with only RGB supervision. We observe fewer floating artifacts when jointly training with audio. **(a)-(c)** Three views from "Large 1". **(d)-(f)** Three views from "Large 2".

Results. We train on the two large rooms in our training set. For each room $75, 100, 150$ images are used for training, while the same 50 images of novel views are used for all configurations during testing. In Table 3 we observe that training with acoustic information helps improve the PSNR and MSE of the visual output. This effect is more significant when the training images are very sparse, the NAF network helps less when there is sufficient visual information. Qualitative results are shown in Figure 7, we see there is a reduction of floaters in free space.

4.5. Inferring scene structure

Given a reverberant sound, humans are able to build a mental representation of the surrounding room and make a judgement about the distance of nearby obstacles (Kolarik et al., 2016). We investigate the intermediate representations constructed by our neural network in the process of learning an acoustic field, and examine if these representations can be used to decode the scene structure.

Setup. The intermediate representation of the NAF depends on both listener locations q and emitter location q', the rotation angle θ, the ear k, the time t and frequency f. For consistency, at a given location (x^*, y^*) in the scene, we extract the NAF latent by setting the emitter location $q'_i = (x^*, y^*)_i$. For the listener location, we iterate over five randomly selected points in the scene $q \in [q_1, \ldots, q_5]$, which we keep constant for all q'_i. The rotation angle is fixed to $\theta = 0$, and we compute the representation average over all possible (k, t, f), and concatenate latents for the selected q.

	Explained variance						
Features	Large 1	Large 2	Medium 1	Medium 2	Small 1	Small 2	Mean
MFCC	0.501	0.458	0.614	0.642	0.820	0.723	0.626
NAF latents	**0.908**	**0.891**	**0.900**	**0.923**	**0.936**	**0.916**	**0.913**

Table 4. **Quantitative Results on scene structure decoding.** We measure the explained variance scores of the predicted wall distance against the ground truth wall distance at test time locations after linear decoding. NAF latents consistently achieve higher explained variance scores than MFCC features.

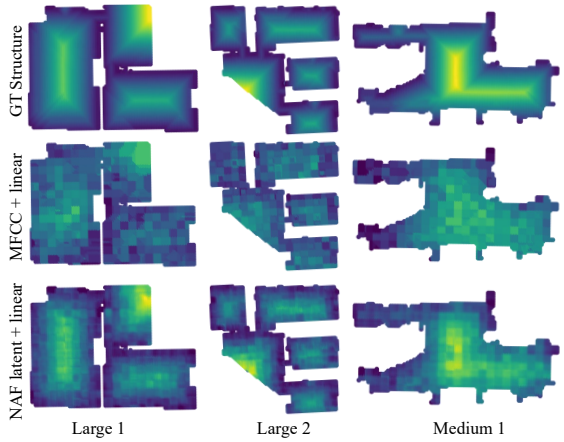

Figure 8. **Qualitative Visualization of scene structure decoding with a linear layer. Top:** The ground truth scene structure map, at each position we visualize the distance to the nearest wall. **Middle:** Linear decoding results using MFCC features. **Bottom:** Linear decoding results using NAF features.

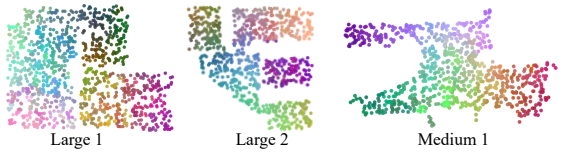

Figure 9. **Visualization of NAF latents.** We apply TSNE to reduce the dimensionality of the NAF latents. The latents learned by our NAF exhibit clear structure.

For our NAFs, latents are extracted from the last layer prior to the output which includes 512 neurons. As a comparison to our learned representation, we extract Mel-frequency cepstral coefficients (MFCCs) from the ground truth impulse response provided by a nearest neighbor interpolator. We use a similar setup as above, for a given location we set this to be q_i', and iterate over the same five listener locations $q_{1\ldots5}$. We average the MFCCs over the left and right ear, and concatenate for the selected q. After flattening, the MFCC features are approximately 500 dimensional for any given room.

We fit a single linear layer to NAF and MFCC features respectively. For testing and visualization of the linear decoding results, we sample a regular grid of points with $0.1m$ distance between each point. For fitting the linear decoder, we randomly sample points within the scene such that the number of training points are 10% as many as the testing points. For each location in the scene, we extract the distance to the nearest wall as the decoding target.

Results. We visualize the results of our linear decoding in Figure 8. As shown in Figure 9, the intermediate representation of our NAFs reveals an underlying structure that is both smooth and semantically meaningful. In the multiroom scenes, the latent is well separated for each room. We are able to successfully decode the scene structure with a linear layer when using our NAFs, but decoding fails when using MFCC features. In Table 4, we show the amount of explained variance of our decoding results on the test set. Our learned features are able to consistently achieve much higher scores than those using MFCC features.

5. Limitations and Future Work

In this work, we provide the first exploration into learning an implicit function that represents the underlying spatial acoustic field. Our model can generalize to continuous locations after training, and provides learned representations that are useful for audio-visual tasks and understanding scene structure.

However our model still has limitations. In line with prior spatial acoustic field coding work, our approach does not model the phase. While a magnitude only approximation may still model plausible spatial acoustic effects in a compact and continuous fashion, such a representation may not be sufficient for tasks that depend on the phase (E.g. phase based microphone array direction-of-arrival estimation). Progress on learned approaches for waveform recovery (Oord et al., 2016; Kalchbrenner et al., 2018), offer some promise for joint modeling of magnitude and phase. These approaches are orthogonal to our work, and we leave this exploration to a future study. Also similar to prior acoustic field work, our NAFs requires a precomputed acoustic field. While this is not a limitation for many applications, the ability to generalize from extremely sparse training samples could open up new potential use cases. Finally, like other prior work that utilize implicit neural representations, our NAFs are fit to a specific scene. The ability to predict the acoustic field of novel scenes remains an open question.

6. Conclusion

In summary, this paper introduces Neural Acoustic Fields (NAFs), a compact, continuous, and differentiable acoustic

representation which can represent the underlying reverberation of different audio sources in a scene. By conditioning NAFs locally on the underlying scene geometry, we demonstrate that our approach enables the prediction of plausible environmental reverberations even at unseen locations in the scene. Furthermore, we demonstrate that the acoustic representations learned through NAFs are powerful, and may be utilized to facilitate audio-visual cross-modal learning, as well as to infer the structure of scenes.

References

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

Aytar, Y., Vondrick, C., and Torralba, A. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016.

Breebaart, J., Herre, J., Faller, C., Rödén, J., Myburg, F., Disch, S., Purnhagen, H., Hotho, G., Neusinger, M., Kjörling, C., et al. Mpeg spatial audio coding/mpeg surround: Overview and current status. *Preprint 119th Conv. Aud. Eng. Soc.*, (CONF), 2005.

Chaitanya, C. R. A., Raghuvanshi, N., Godin, K. W., Zhang, Z., Nowrouzezahrai, D., and Snyder, J. M. Directional sources and listeners in interactive sound propagation using reciprocal wave field coding. *ACM Transactions on Graphics (TOG)*, 39(4):44–1, 2020.

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 *ECCV*, 2020.

Chen, Z. and Zhang, H. Learning implicit fields for generative shape modeling. In *Proc. CVPR*, pp. 5939–5948, 2019.

Défossez, A., Zeghidour, N., Usunier, N., Bottou, L., and Bach, F. Sing: Symbol-to-instrument neural generator. *arXiv preprint arXiv:1810.09785*, 2018.

DeVries, T., Bautista, M. A., Srivastava, N., Taylor, G. W., and Susskind, J. M. Unconstrained scene generation with locally conditioned radiance fields. *arXiv preprint arXiv:2104.00670*, 2021.

Du, Y., Collins, M. K., Tenenbaum, B. J., and Sitzmann, V. Learning signal-agnostic manifolds of neural fields. In *Advances in Neural Information Processing Systems*, 2021.

Gan, C., Huang, D., Zhao, H., Tenenbaum, J. B., and Torralba, A. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10478–10487, 2020a.

Gan, C., Zhang, Y., Wu, J., Gong, B., and Tenenbaum, J. B. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, pp. 9701–9707, 2020b.

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., et al. Threedworld: A platform for interactive multi-modal physical simulation. *NeurIPS*, 2021.

Gerzon, M. A. Periphony: With-height sound reproduction. *Journal of the audio engineering society*, 21(1):2–10, 1973.

Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., and Funkhouser, T. Local implicit grid representations for 3d scenes. In *Proc. CVPR*, pp. 6001–6010, 2020.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S., and Kavukcuoglu, K. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, 2018.

Kolarik, A. J., Moore, B. C., Zahorik, P., Cirstea, S., and Pardhan, S. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, 78(2):373–395, 2016.

Mehra, R., Antani, L., Kim, S., and Manocha, D. Source and listener directivity for interactive wave-based sound propagation. *IEEE transactions on visualization and computer graphics*, 20(4):495–503, 2014.

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 *Proc. ECCV*, 2020.

Niemeyer, M., Mescheder, L., Oechsle, M., and Geiger, A. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proc. ICCV*, 2019.

Niemeyer, M., Mescheder, L., Oechsle, M., and Geiger, A. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. CVPR*, 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. *arXiv preprint arXiv:1609.03499*, 2016.

Paasonen, J., Karapetyan, A., Plogsties, J., and Pulkki, V. Proximity of surfaces—acoustic and perceptual effects. *Journal of the Audio Engineering Society*, 65(12):997–1004, 2017.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, 2019.

Pierce, A. D. *Acoustics: an introduction to its physical principles and applications*. Springer, 2019.

Pörschmann, C., Arend, J. M., Bau, D., and Lübeck, T. Comparison of spherical harmonics and nearest-neighbor based interpolation of head-related transfer functions. In *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.

Pulkki, V. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6): 503–516, 2007.

Raghuvanshi, N. and Snyder, J. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014.

Raghuvanshi, N. and Snyder, J. Parametric directional coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M., and Govindaraju, N. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. In *ACM SIGGRAPH 2010 papers*, pp. 1–11. 2010.

Ratnarajah, A., Zhang, S.-X., Yu, M., Tang, Z., Manocha, D., and Yu, D. Fast-rir: Fast neural diffuse room impulse response generator. *arXiv preprint arXiv:2110.04057*, 2021.

Richard, A., Markovic, D., Gebru, I. D., Krenn, S., Butler, G., de la Torre, F., and Sheikh, Y. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, pp. 2304–2314, 2019.

Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705, 1999.

Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., and Kweon, I. S. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4358–4366, 2018.

Singh, N., Mentch, J., Ng, J., Beveridge, M., and Drori, I. Image2reverb: Cross-model reverb impulse response synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

Sitzmann, V., Zollhöfer, M., and Wetzstein, G. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proc. NeurIPS 2019*, 2019.

Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., and Lipman, Y. Multiview neural surface reconstruction by disentangling geometry and appearance. *Proc. NeurIPS*, 2020.

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 570–586, 2018.

Zhao, H., Gan, C., Ma, W.-C., and Torralba, A. The sound of motions. In *ICCV*, pp. 1735–1744, 2019.

Appendix

A. Additional Visualization of Rooms

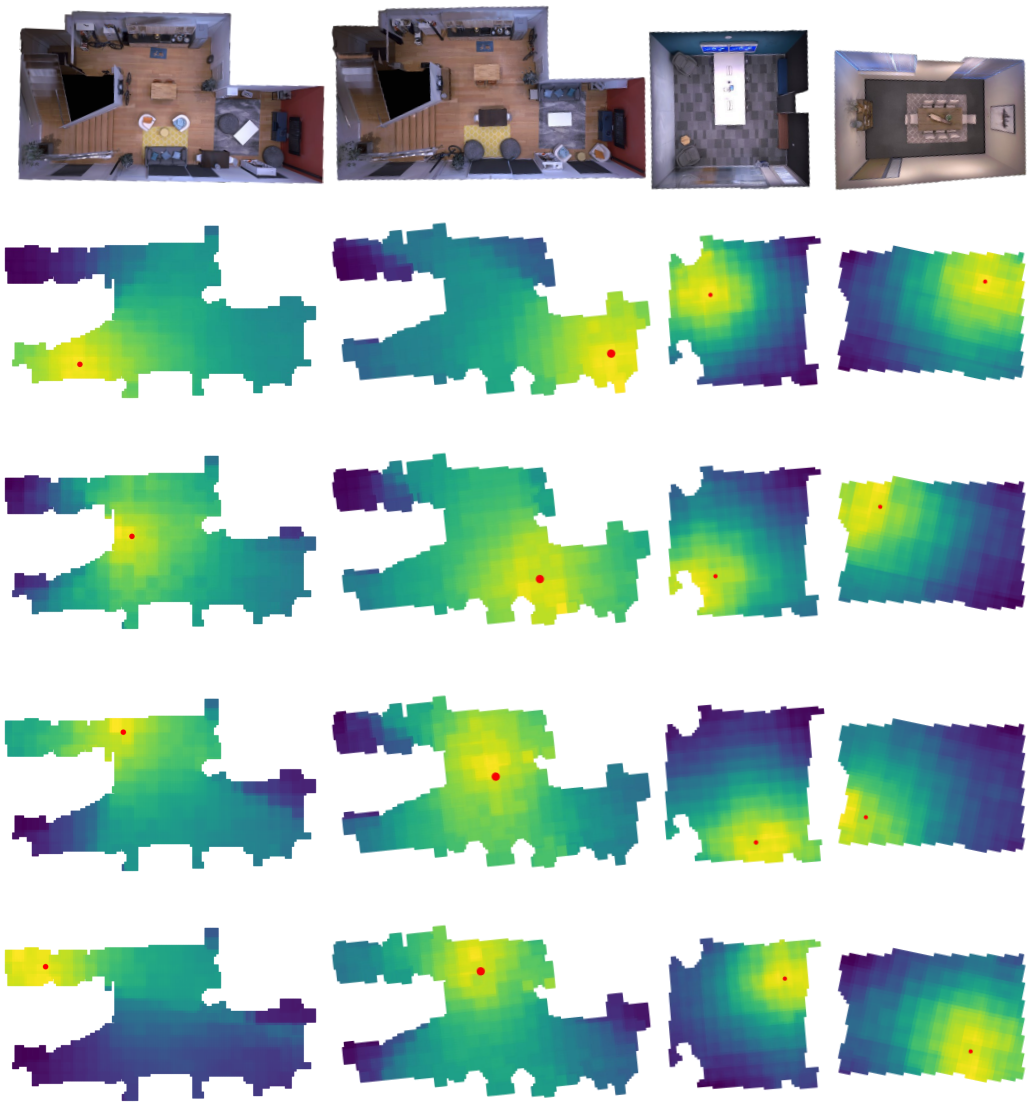

Figure A1. **Additional Qualitative Predictions of NAF.** Qualitative visualization of the loudness map as predicted by NAF across four different rooms.

We show additional NAF predictions of loudness as we move an emitter inside different rooms in Figure A1. For each room, note how the sound is affected by the geometry. In wide open spaces the sound is highly dispersed. While in thin structures the sound tends to concentrate locally. As we move farther from the source, the loudness of the sound decreases.

B. Architecture and Training Details

We visualize the two alternative models that we experiment with, in Figure A2 is a network that uses different local feature grids for the emitter and receiver. The network uses the emitter and listener positions to sample from the two different grids.

In Figure A3 we show a model that does not utilize any kind of local geometry conditioning. The listener, emitter, phase, and time input are transformed using sinusoidal embedding, while the orientation and left/right are retrieved. All transformed inputs are directly fed to the network.

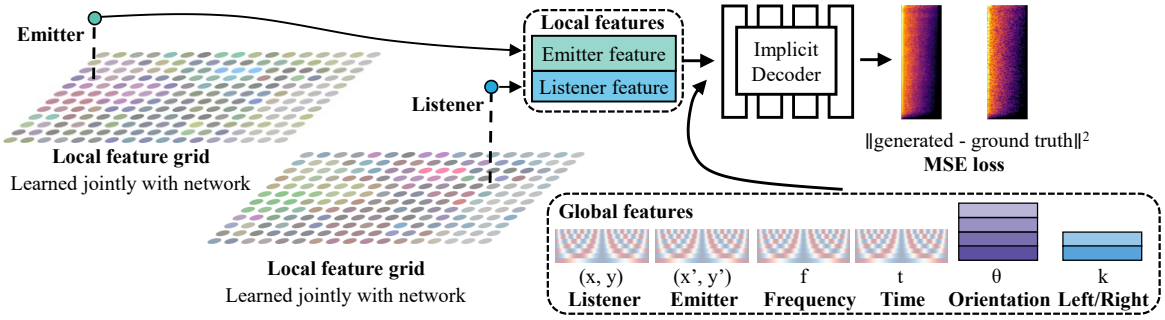

Figure A2. Architecture of the model that uses emitter and listener specific local geometry conditioning.

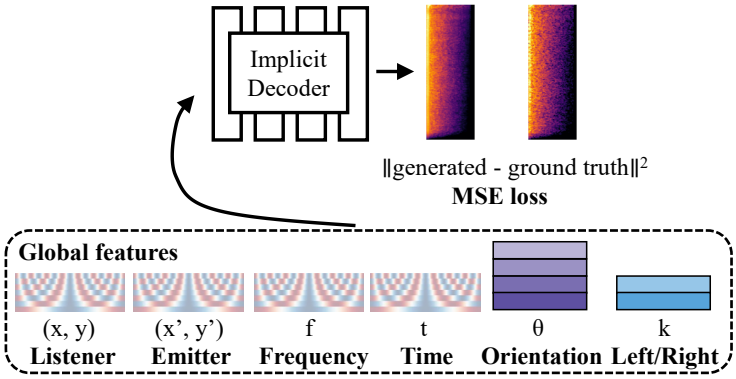

Figure A3. Architecture of the model that uses no local geometry conditioning.

Each network consists of 8 fully connected layers in a feedforward fashion, as well as a skip connection consisting of two fully connected layers. The skip connection takes the input and adds its output to that for the fourth intermediate layer. We utilize an intermediate feature size of 512, and Leaky ReLU with a slope of 0.1 as the activation function. The grid is initialized to stretch the bounding box of a scene. Each point is located at a distance of 0.25m from the nearest neighbor. 64 features are used for each point. Each element of the grid is initialized i.i.d. from $\mathcal{N}(0, \frac{1}{\sqrt{64}})$. We initialize the bandwidth for each point at $\sigma = 0.25$, and jointly train the bandwidth as part of the network. For the network and the grid, we utilize an initial learning rate of $5e-4$. The *Adam* optimizer is used when training our network. We utilize a orientation embedding of shape $\mathcal{R}^{7\times4\times512}$ where 7 is the number of intermediate outputs, 4 is the number of orientations, and 512 is the feature dimension. For the left-right embedding, we use a shape of $\mathcal{R}^{7\times2\times512}$. We perform additive conditioning by adding a $\mathcal{R}^{512}$ vector to each intermediate output for both the orientation and the left/right.

For each scene, to generate a log-spectrogram for each impulse response, we compute the mean and standard deviation $\mu_{(f,t)}, \sigma_{(f,t)}$ for each frequency/time index in the log-spectrogram, and normalize the data prior to training:

$$v_{(f,t)} = \frac{v_{(f,t)} - \mu_{(f,t)}}{3.0 \times \sigma_{(f,t)}}$$

For the sinusoidal embedding, we utilize both $\cos$ and $\sin$ with 10 frequencies each for encoding position, phase, and time. For encoding position we utilize a max frequency of 2^7Hz, while for encoding time and frequency we utilize a max frequency of 2^{10}Hz.

Since we do not know beforehand the time duration of an impulse response at an unseen location, we compute the maximum impulse length for each scene and use this length to zero pad the training impulse responses. Because the padded regions do not contain useful information, we want the network to focus modeling efforts on the early regions of the impulse response. We achieve this by stochastically padding the impulse response to maximum impulse length with 0.1 probability. Because the implicit function is trained on individual (t, f) coordinates within a given v_{STFT}, training samples do not need to be of the same length. During test time, we perform inference up to the maximum duration of scene impulse response.

C. Dataset Visualization

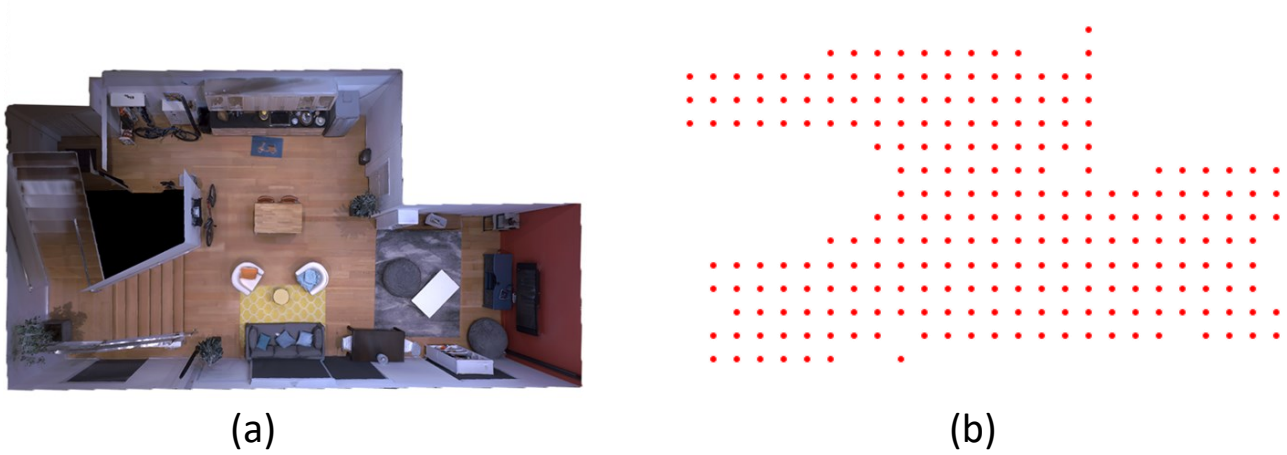

Figure A4. A room the emitter-listener probes. **(a)** The 3D structure of a room. **b** The probes marking the location of emitters/listeners.

In Figure A4, we visualize both the room and underlying set of probe positions in the training data. Due to occlusion and the geometry, even slightly moving the emitter or listener position can result in different results. As we demonstrated in Table 7, both nearest neighbor and linear interpolation perform poorly compared to our learned solution. In contrast, recovered acoustic fields from NAF trained on these probe positions is substantially denser (Figure A1).

D. Storage Comparison

	Storage (MiB)						
Method	Large 1	Large 2	Medium 1	Medium 2	Small 1	Small 2	Mean
AAC	495.97	478.55	483.42	451.14	116.75	54.64	346.74
Opus	258.51	257.08	245.65	231.06	66.15	29.75	181.37
NAF (Dual)	8.78	8.87	8.87	8.92	8.45	8.37	8.71
NAF (Shared)	8.44	8.49	8.49	8.51	8.28	8.23	8.41

Table A1. **Storage cost of different methods.** We average the amount of data required for different methods of inference for the six scenes. Our NAFs are able to compactly represent the scene while maintaining higher quality.

We compare the averaged on disk storage cost of the different methods for inferring the spatial audio using a precomputed training set in Table A1. Both linear and nearest interpolation methods require access to the entire training set, while our NAF based approaches compactly encode the acoustic scene.

E. Details of the compression baselines

If uncompressed, the precomputed spatial acoustic field can reach gigabyte or terabyte sizes depending on probe density, scene size, and bandwidth of the impulse. When applied to gaming and virtual reality applications, minimizing the space taken up by these acoustic representations is critical and have been widely studied.

We utilize two state-of-the-art lossy coding methods applied to the audio. They are respectively Advanced Audio Coding (AAC-LC) and Xiph Opus. These two methods were chosen because they are in widespread usage for media encoding, are among the best coding methods for a given bitrate, and have high quality open-source implementations available.

We describe the parameters and additional details for these two coding methods.

E.1. AAC baseline

We utilize `ffmpeg 5.0`, and select the open source "aac" implementation. We set the combined stereo bitrate to 24 kBit/s in constant bit rate mode, as we found that there are occasional encode/decode failures below this bitrate.

E.2. Opus baseline

We utilize `opustools 0.2` backed by `libopus 1.3.1`. The encoder is set to 12kBit/s for stereo (6kBit/s per channel) in constrained variable bitrate mode. Complexity it set to the maximum of 10, and music mode is set (as opposed to speech tuning mode).

F. Alternative Neural Representations

Representation	Spectral loss ↓	T60↓
Time domain	2.046	49.72
Magnitude + phase	0.427	5.694
Magnitude only	0.406	2.872

Table A2. **Learning different representations** We compare learning magnitude only, jointly learning magnitude and phase, as well as directly learning in the time domain. For the magnitude + phase, we allow the network to fit the instantaneous frequency, a representation that is believed to be easier for networks to learn.

Our current method follows prior work in learning in the log-magnitude STFT domain. In this section, we investigate two possible alternatives: learning phase + log-magnitude, and directly learning in the time domain. The MSE and T60 error percentage is presented in Table A2. We observe that jointly modeling phase + log-magnitude degrades the performance slightly compared to modeling just the log-magnitude, while modeling in the time domain performs poorly.

G. L_2 regularized grid in NeRF

	Large 1		Large 2	
	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓
NeRF + grid + L_2	22.69	6.956	24.86	7.128
NeRF + grid	**25.41**	**6.618**	**25.70**	**6.921**

Table A3. **Regularizing the grid.** In this experiment, we compare learning NeRF with a grid without regularization, and with L_2 regularization.

In Table A3 we compare NeRF that utilizes a grid and trained using image reconstruction loss, against a variant where a L_2 penalty with weight $1e-5$ to ensure a smooth latent space is added to the image reconstruction loss. There are 75 images used in the training set. We observe degraded performance when we apply this penalty. This indicates that our NAFs are providing more information than simple regularization to ensure a smooth latent grid.