Be Everywhere - Hear Everything (BEE): Audio Scene Reconstruction by Sparse Audio-Visual Samples

Mingfei Chen [1] Kun Su [1] Eli Shlizerman [1, 2*]
[1] Department of Electrical & Computer Engineering, University of Washington
[2] Department of Applied Mathematics, University of Washington

Figure 1: **Be Everywhere - Hear Everything (BEE)**: Audio reconstruction of a scene with dynamic emitters at arbitrary listener locations, leveraging inputs from sparse A/V receivers. Please see supplementary materials for sample videos.

Abstract

Fully immersive and interactive audio-visual scenes are dynamic such that the listeners and the sound emitters move and interact with each other. Reconstruction of an immersive sound experience, as it happens in the scene, requires detailed reconstruction of the audio perceived by the listener at an arbitrary location. The audio at the listener location is a complex outcome of sound propagation through the scene geometry and interacting with surfaces and also the locations of the emitters and the sounds they emit. Due to these aspects, detailed audio reconstruction requires extensive sampling of audio at any potential listener location. This is usually difficult to implement in realistic real-time dynamic scenes. In this work, we propose to circumvent the need for extensive sensors by leveraging audio and visual samples from only a handful of A/V receivers placed in the scene. In particular, we introduce a novel method and end-to-end integrated rendering pipeline which allows the listener to be everywhere and hear everything (BEE) in a dynamic scene in real-time. BEE reconstructs the audio with two main modules, Joint Audio-Visual Representation, and Integrated Rendering Head. The first module extracts the informative audio-visual features of the scene from sparse A/V reference samples, while the second module integrates the audio samples with learned time-frequency transformations to obtain the target sound. Our experiments indicate that BEE outperforms existing methods by a large margin in terms of quality of sound reconstruction, can generalize to scenes not seen in training and runs in real-time speed.

*Corresponding author: shlizee@uw.edu

1. Introduction

It is Friday night, and your favorite jazz band is performing at the Birdland Jazz Club in New York. Your friends will be attending but you cannot attend in person. Imagine that instead, it would be possible to join them virtually, as in the scenario illustrated in Figure 1. Enabling such an immersive experience requires high-fidelity real-time spatial audio reconstruction of the scene and could unlock novel experiences in applications of virtual reality, mixed reality, and immersive live-streaming.

While the dynamic aspects of such scenes are the ones that make them immersive, these same aspects make audio reconstruction a challenging problem. In particular, for these scenes, sound reconstruction is an outcome of (i) scene properties related to sound propagation, such as

room geometry, surface materials, etc., and (ii) actions of the emitters, such as their positions and emitted sounds at every time step. Due to these, a possible direct approach for audio reconstruction at arbitrary listener locations could be to place a microphone at each such location, e.g., designing a dense mesh of microphones. Such a solution is usually impractical in realistic scenes. An alternative approach to deal with multiple moving emitters at each time could be to track the moving emitters and to design high-end equipment to collect clean emitter sounds of each emitter so that they can be integrated and synthesized for each possible location of the listener. With known emitter locations, this approach can utilize common audio reconstruction techniques to render the Room Impulse Response (RIR) for each emitter-listener pair, then perform convolution of the emitter sound with the corresponding RIR, and integrate the outcomes to obtain the reconstructed audio at arbitrary listener location [13, 27, 12, 21, 15, 29, 4, 22, 23, 24, 16, 17]. While this approach is plausible, beyond the requirement to design novel equipment, it also implies an extensive computational cost of the integration, which is expected to increase with the number of emitters.

Due to the above described challenges, both approaches are generally impractical and warrant the development of methods that synthesize the audio for an arbitrary listener location from a sparse set of fixed sensors. Indeed, recently developed neural sound synthesis methods have been shown to synthesize audio through generative neural network modeling conditioned on A/V samples from sensors for a single listener location [28, 25, 8, 9, 6, 10, 30]. While promising, they appear to depend on the training samples on which the network has been trained on, since scene properties that determine sound propagation are not explicitly captured during training. This restricts the reconstruction accuracy and the generalization to various dynamic scenes.

To address these limitations, in this work, we propose a novel neural sound reconstruction and synthesis system that leverages samples from a sparse set of fixed A/V sensors that sample the audio (waveforms captured by microphones) and in addition *the visual information* (egocentric images captured by cameras) at any given time. Our proposed end-to-end integrated audio rendering pipeline is capable to render high-quality audio and generalize the audio reconstruction to arbitrary listener locations, effectively allowing the listener to be everywhere and hear everything (BEE). BEE contains two modules, namely, the Joint Audio-Visual Representation module (JAVR) and Integrated Rendering Head (IRH). JAVR learns and represents the properties of the scene by projecting visual samples into a world coordinate system and obtains a 3D visual feature volume for the acoustic propagation space of the scene. Through this space the A/V receivers and the listeners are associated and then correlated by injecting audio features into the 3D visual representation and employing cross-attention. This constitutes the audio-visual representation of the scene. The target listener sound is synthesized by the IRH module within BEE, which learns time-frequency (TF) transformations after integrating the audio-visual features generated from JAVR and the received audio samples on different levels through two decoupled branches.

In summary, our main contributions in this work are as follows: 1) We develop an end-to-end integrated rendering pipeline, named BEE, to address audio reconstruction at arbitrary listener locations for dynamic scenes by *sparse audio-visual* samples. 2) BEE constructs an effective Joint Audio-Visual Representation module that can learn an audio-visual representation of the scene. Such representation along with an Integrated Rendering Head module implicitly untangles the contribution of each emitter to the sound at an arbitrary listener location. 3) Experiments on the SoundSpaces dataset [7] with Replica and Matterport3D scenes demonstrate that BEE outperforms existing methods by a large margin in quality, ability to generalize to various scenes, and runs in real-time.

2. Related Work

Audio Scene Reconstruction. Most traditional audio scene reconstruction methods reconstruct the sound at arbitrary listener locations by convolving the sound waveform of each emitter with the corresponding Room Impulse Response (RIR) and then summing the outcomes of each emitter. Conventional RIR modeling can be divided into two categories: 1) Wave-based methods which aim to solve the acoustic wave equation using numerical techniques [13, 27, 12, 21]. 2) Geometry-based methods [1, 3, 15, 29, 4, 20] which treat sound propagation as optic rays and determine the path of sound propagation according to energy attenuation. Recent methods utilize deep learning approaches to generate spatial RIRs for arbitrary emitter-receiver pairs. Methods such as IR-GAN [22] and fast-RIR [23] learn a deep generative model, while other methods such as IR-MLP [24] and NAF [16] learn an implicit neural function to represent RIR. Few-shotRIR [17] introduced a transformer-based method to infer RIRs based on a sparse set of images and echoes observed by receiver sensors and showed generalization to unseen scenes. These RIR-based methods require explicit location and sound waveform of each emitter and the computational cost typically increases linearly with the number of emitters. The location and the source sound waveform of each emitter might be difficult to obtain without special scene setup. Our approach, in costrast, utilizes the sparse audio-visual samples to build the audio-visual representation of the scene, and then learns audio reconstruction at listener locations. As such it can handle unknown numbers, locations, and source sound waveforms of emitters, and is generalizable over unseen scenes.

Audio-Visual Sound Synthesis. Our task can also be considered as an audio-visual guided sound synthesis at arbitrary listener locations. Popular sound synthesis methods such as WaveNet [28] build an autoregressive model to synthesize the target sound conditioned on all previous sounds. WarpNet [25] learns an end-to-end neural synthesis approach to synthesize high-quality binaural audio from mono audio. Additional works propose to use generative adversarial networks to synthesize the target sound [8, 9]. While these sound synthesis methods succeed in directly generating the target sound, they do not incorporate aspects related to sound propagation in the scene, i.e., the geometry structure, the materials of the scene, and the correlation of the given audio which indicates the condition of emitters. The synthesis quality of these methods is thus dependent on provided training samples and typically difficult to generalize to unseen scenes. Recently, audio-visual guided methods [6, 10, 30] demonstrated advanced quality of sound synthesis. Given visual and audio input, the methods aim to associate the visual with audio features to reverberate the input sound [6], generate stereophonic audio [10, 30] or perform audio separation tasks [30]. These methods rely on visual and audio samples from a single sensor and overlook associating the audio and visual features in the 3D scene space. As a result, such approaches do not perform well on audio reconstruction tasks in scenes with complex layouts.

3. Methods

3.1. Overview

In this paper, we introduce the novel task of audio reconstruction at arbitrary listener locations based on audio-visual samples by A/V sensors sparsely placed in the 3D scene. Specifically, given a scene, N sparsely distributed A/V reference receivers $\{M_i | i = 1, 2, ..., N\}$ (e.g., $N = 4$) are placed in fixed locations. Each receiver M_i can continuously obtain one binaural sound waveform S_i and one egocentric RGB-D view I_i at location (x_i, y_i, h_i) with orientation θ_i. The view I_i is captured by a pre-calibrated camera with parameters $C_i = ([\mathbf{R}_i|\mathbf{t}_i], \mathbf{K}_i)$, where $[\mathbf{R}_i|\mathbf{t}_i]$ is the extrinsic matrix and $\mathbf{K}_i$ is the intrinsic matrix. Listeners L_k can be positioned at arbitrary location (x_k, y_k, h_k) with orientation θ_k.

Our proposed framework, BEE, synthesizes the target binaural sound S_k for L_k, based on the audio-visual samples from the N receivers $\{M_i = (I_i, S_i, C_i, (x_i, y_i, h_i), \theta_i) | i = 1, 2, ..., N\}$. As illustrated in Figure 2, BEE consists of two modules: 1) Joint Audio-Visual Representation (JAVR) module (Section 3.2) and 2) Integrated Rendering Head (IRH) (Section 3.3). JAVR is applied first and learns a joint audio-visual feature embedding $\tilde{\mathbf{Q}}_{ik}$ for each receiver-listener pair (M_i, L_k) w.r.t their pose information (locations and orientation angles). Based on the audio-visual feature embeddings $\tilde{\mathbf{Q}}_{ik}$. IRH follows JAVR and learns to transform and integrate $\{S_i | i = 1, 2, ..., N\}$ from the reference receivers to synthesize the output sound S_k for L_k. We describe the training details of BEE in Section 3.4.

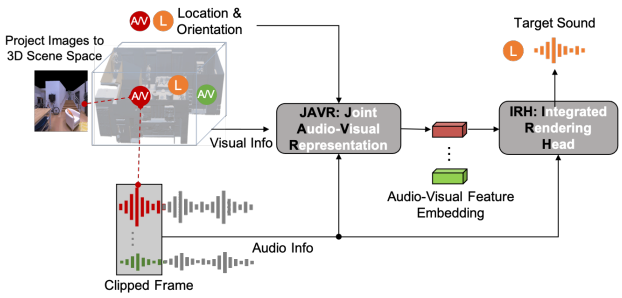

Figure 2: **Overview of BEE framework** which contains two modules, Joint Audio-Visual Representation (JAVR) module and Integrated Rendering Head (IRH). JAVR projects all images observed from reference A/V receivers to 3D scene space and integrates the obtained visual feature volume with received audio clips with respect to locations and orientations of receivers and target listener L. IRH renders the sound at L by integrating the audio-visual feature embeddings of all receiver-listener pairs from JAVR.

3.2. Joint Audio-Visual Representation (JAVR)

JAVR extracts the audio-visual features for each (M_i, L_k) pair in four steps: 1) Deployment of 3D visual encoder to build a visual feature volume from $\{(I_i, C_i) | i = 1, 2, ..., N\}$ for the acoustic propagation space $\mathbf{P}$ of the scene; 2) Enhancement of visual features of $\mathbf{P}$ using spatial locations of M_i and L_k; 3) Encoding audio samples $\{S_i | i = 1, 2, ..., N\}$ to binaural audio features through the Binaural Audio Encoder; 4) Learning the audio-visual features on $\mathbf{P}$ and extract the feature embedding $\tilde{\mathbf{Q}}_{ik}$ w.r.t each (M_i, L_k) pair. These steps are illustrated in Figure 3 (a) and we describe them further below.

3D Visual Encoder. The 3D *Visual* Encoder projects pixels of given RGB-D views $\{I_i | i = 1, 2, ..., N\}$ to a world coordinate system to form 3D visual features volume for the acoustic propagation space $\mathbf{P}$ of the scene. As a first step, a CNN backbone is applied I_i to extract image features $\mathbf{F}_i \in \mathbb{R}^{W \times H \times C}$. Compared to I_i, $\mathbf{F}_i$ contains fewer pixels and aggregates contextual visual information at each pixel. Therefore, the encoder effectively projects pixel locations in $\mathbf{F}_i$ instead of I_i to the world coordinate system. Specifically, we associate each pixel location $z \in \mathbb{R}^{W \times H}$ in $\mathbf{F}_i$ to the corresponding pixel location $\hat{z}$ in I_i and then multiply $\hat{z}$ with the inverse of the intrinsic matrix $\mathbf{K}_i$ to transform $\hat{z}$ to the coordinate system of C_i camera. We then multiply the obtained coordinate with the inverse of the camera pose $[\mathbf{R}_i|\mathbf{t}_i]$ to project it to the 3D world coordinate system, de-

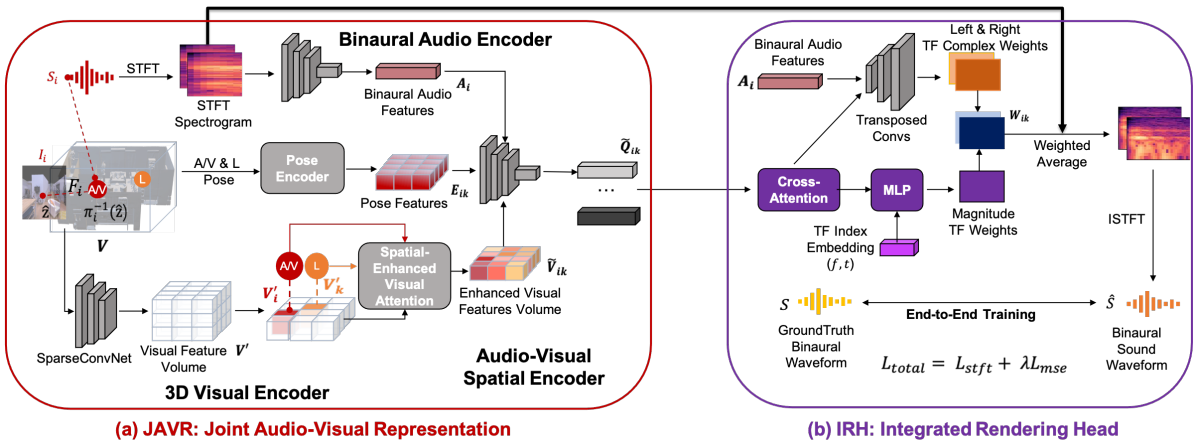

Figure 3: **Structure of BEE modules.** (a) JAVR includes four sub-modules that extract audio-visual features: 3D Visual Encoder, Spatial Enhanced Visual Representation, Binaural Audio Encoder and Audio-Visual Spatial Enocoder. (b) IRH transforms S_i with the learned time-frequency (TF) weights for the synthesis of the target sound $\hat{S}$. For training, $\hat{S}$ is supervised by S directly.

noted as $\pi_i^{-1}(\hat{z}) \in \mathbb{R}^3$. In the next step, each pixel-aligned feature in $\mathbf{F}_i$ is associated with a point in the world coordinate system. Altogether the features form a 3D visual feature volume $\mathbf{V}$, where $\mathbf{V}(\pi_i^{-1}(\hat{z})) = \mathbf{F}_i(z)$. Due to the sparse distribution of the receivers $\mathbf{V}$ is sparse as well and hence we propose to fill it with SparseConvNet [11] which obtains a denser visual feature volume $\mathbf{V}'$.

Spatial Enhanced Visual Representation. The acoustic propagation path depends on the locations of the emitter and the listener. Therefore, to learn a reliable transformation from the reference sound S_i to the target sound S_k, it is necessary to focus on different regions of the space $\mathbf{P}$ that correspond to the locations of the receiver-listener pair (M_i, L_k). To attain this, we introduce the Spatial Enhanced *Visual* Attention module. In this module, we first obtain the location-aligned visual feature $\mathbf{V}'_i$ and $\mathbf{V}'_k$ for M_i and L_k from $\mathbf{V}'$ respectively. Bilinear interpolation is applied if the location is fractional. Then, for each point $\mathbf{P}_j$ within $\mathbf{P}$, we take $\mathbf{V}'_i$ and $\mathbf{V}'_k$ as query embeddings to compute the correspondence scores s_{ij} and s_{kj} w.r.t. $\mathbf{V}'_j$ according to

$$s_{lj} = \frac{(\boldsymbol{W}_{1l}\mathbf{V}'_l + b_{1l})(\boldsymbol{W}_{2j}\mathbf{V}'_j + b_{2j})^\top}{\sqrt{C}}, \quad (1)$$

where $l = \{i, k\}$, C is the channel dimension of $\mathbf{V}'$ and $\boldsymbol{W}$ represents linear projection layers. We multiply $\mathbf{V}'$ with the point-wise scores s_{ij} and s_{kj} and sum the two new weighted feature volumes and $\mathbf{V}'$ to obtain the Spatial Enhanced visual feature volume $\tilde{\mathbf{V}}_{ik}$ for the pair (M_i, L_k).

Binaural Audio Encoder. Since the waveform information includes both emitter locations and the corresponding emitted sounds, we include *audio* features in the framework. Given the binaural sound waveform set $\{S_i | i = 1, 2, ..., N\}$ obtained from N reference receivers, we build a Binaural Audio Encoder to extract the binaural audio features from S_i. Specifically, we first transform the original S_i to complex spectrogram $\hat{S}_i \in \mathbb{C}^{2\times F\times T}$ by applying the Short-Time Fourier Transform (STFT), where F is the number of frequency bins and T is the number of overlapping time windows. We then utilize a 2D convolutional layer sequence followed by linear layers to extract high-level audio features $\mathbf{A}_i \in \mathbb{R}^{2\times C'}$ from both channels of $\hat{S}_i$, where C' is the feature channel dimension.

Audio-Visual Spatial Encoder. We integrate the obtained *visual* and *audio* representation on the acoustic propagation space $\mathbf{P}$, and aggregate the audio-visual feature embedding for each receiver-listener (M_i, L_k) pair. The integration is done by first generating pose feature embedding for (M_i, L_k) w.r.t. each point $\mathbf{P}_j$ within the space $\mathbf{P}$. Specifically, we compute the location vector from $\mathbf{P}_j$ to M_i and L_k, denoted as $\vec{d}_{ji}$ and $\vec{d}_{jk}$ respectively, and concatenate them with the orientation angles of M_i and L_k (θ_i and θ_k). By mapping the combined pose vector $(\vec{d}_{ji}, \vec{d}_{jk}, \theta_i, \theta_k)$ to a higher dimension through the sinusoidal encoding followed by a linear layer, we can obtain the pose feature embedding $\mathbf{E}(i, j, k)$. After this, $\mathbf{E}(i, j, k)$ is concatenated with the corresponding visual feature $\tilde{\mathbf{V}}_{ik}(\mathbf{P}_j)$ and the binaural audio feature $\mathbf{A}_i$ to form the audio-visual feature as $\mathbf{Q}_{ik}(\mathbf{P}_j) = \{\tilde{\mathbf{V}}_{ik}(\mathbf{P}_j), \mathbf{A}_i, \mathbf{E}(i, j, k)\}$ at the point $\mathbf{P}_j$. By gathering $\mathbf{Q}_{ik}(\mathbf{P}_j)$ of all the points $\mathbf{P}_j$ within the space $\mathbf{P}$, we build an audio-visual feature volume $\mathbf{Q}_{ik}$. As a final step, a stack of convolution layers followed by linear layers is learned to aggregate an audio-visual feature embedding $\tilde{\mathbf{Q}}_{ik} \in \mathbb{R}^d$ out of $\mathbf{Q}_{ik}$. $\tilde{\mathbf{Q}}_{ik}$ is taken as the audio-visual representation of the space P w.r.t the pair (M_i, L_k).

3.3. Integrated Rendering Head

Given a set of audio-visual feature embeddings $\{\tilde{\mathbf{Q}}_{ik} | i = 1, 2, ..., N\}$ generated by the JAVR module, the Integrated

Rendering Head (IRH) module renders the target sound S_k for the listener L_k according to three steps, as described below and illustrated in Figure 3 (b).

In the first step, more comprehensive features are aggregated by a Cross-Attention module which enhances the audio-visual feature of each receiver-listener pair with other pairs. Specifically, each $\tilde{\mathbf{Q}}_{ik}$ is taken as a query embedding to calculate the attention score w.r.t. $\{\tilde{\mathbf{Q}}_{mk}|m = 1, 2, ..., N; m \neq i\}$, similar to the Equation 1. The enhanced feature $\hat{\mathbf{Q}}_{ik}$ is the weighted sum of $\tilde{\mathbf{Q}}_{mk}$ with the calculated attention scores.

In the second step, two decoupled branches are deployed to predict the transformation weights from each S_i to S_k. For higher fidelity, we convert the waveform of S_i to a spectrogram through Short-Time Fourier Transform (STFT), and predict the time-frequency (TF) transformation weights for the frequency band f and time window t. Since binaural audio features contain information on emitters and source sounds, we first inject high-level audio-visual representation feature $\hat{\mathbf{Q}}_{ik}$ to the audio features $\mathbf{A}_i$, and then decode the complex binaural TF weights out of the joint features through a transposed convolutional network. The second branch utilizes MLP network to predict TF magnitude transform weights, that depends on the visual and spatial relationship than the detailed observed sounds. Specifically, we encode f and t with the sinusoidal encoding and concatenate the encoded embeddings with $\hat{\mathbf{Q}}_{ik}$ as input of MLP. The final complex binaural TF weight matrix $\mathbf{W}_{ik} \in \mathbb{R}^{4\times F\times T}$ for each pair (M_i, L_k) is the product of the binaural weights and magnitude weights, where we transform each binaural complex feature channel of $\mathbf{W}_{ik}$ into a 2-channel real matrix.

In the third step, the spectrogram of each input sound $\tilde{S}_i$ is multiplied with the corresponding spectrogram transformation weights $_{ik}$. The average of the obtained newly transformed spectrograms is the output as the final predicted spectrogram for the target listener L_k. The inverse STFT operation then outputs the rendered target sound waveform $\hat{S}_k$ represented as

$$\hat{S}_k = \text{ISTFT}(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{W}_{ik}\tilde{S}_i)), \mathbf{W}_{ik} = \Psi(\hat{\mathbf{Q}}_{ik}, \mathbf{A}_i, f, t). \quad (2)$$

3.4. Training

We use an end-to-end training strategy to train our proposed framework, where all the components and modules are optimized jointly by minimizing the discrepancy between the rendered sound $\hat{S}_k$ for the target listener L_k and the corresponding ground truth sound S_k. The Mean Squared Error (MSE) loss is combined with the STFT loss [2] to regulate our model in both time and frequency domains. Specifically, the MSE loss $\mathcal{L}_{\text{mse}}$ is defined as $\|S_k - \hat{S}_k\|_2^2$, and STFT loss includes conversion of the sound waveform into the frequency-time domain, and then is computed as the sum of two terms: the spectral convergence loss $\mathcal{L}_{\text{sc}} = \frac{\||S_k|-|\hat{S}_k|\|_2}{\||S_k|\|_2}$ and the magnitude loss $\mathcal{L}_{\text{mag}} = \||S_k| - |\hat{S}_k|\|_1$ The total loss is formulated as

$$\mathcal{L}_{\text{stft}} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{mag}}, \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{stft}} + \lambda\mathcal{L}_{\text{mse}}. \quad (3)$$

4. Experiments

4.1. Datasets and Metrics

Datasets Setting. We evaluate our model on SoundSpaces [7], a realistic acoustic simulation platform for audio-visual embodied AI research. In this dataset, we use AI-Habitat simulator [19] with SoundSpaces audio on Replica scenes [26] and Matterport3D scenes [5]. The scenes consist of real-world indoor spaces, e.g., apartments and offices. For each scene, SoundSpaces computes an axis-aligned 3D bounding box of the scene, samples location points from a 2D square grid that slices the bounding box in the horizontal plane, and provides dense pairs of binaural RIRs for different head orientations generated by geometric sound propagation methods.

Metrics. We use three main standard metrics to evaluate audio reconstruction performance (the lower, the better): 1) L1 distance of STFT spectrogram (**STFT**) of the left and right channels; 2) Deep Perceptual Audio Metric (**DPAM**) [18]: deep learning based perceptual quality metric that is well-calibrated with human judgments; 3) Energy Envelope Error (**ENV**) [10]: the envelope of the signals which measures the Euclidean distance between the envelopes of the ground-truth left and right channels and the predictions.

4.2. Implementation Details

The visual-audio information from the scene is obtained by four A/V reference receivers on the midpoints of four edges of the smallest rectangle containing the room floor plane, with orientation to the interior. At each reference receiver a 128×128 egocentric RGB-D image is rendered, the received sound S_i is constructed and the target listener sound S_k is constructed by convolution of emitted sound clips with the corresponding binaural RIRs and their summation. Since there is typically noise in real setting, we add simulated binaural isotropic ambient noise with a $0-5$ dB signal-to-noise ratio (SNR). Since our model requires continuous sound streams, we use a sliding window with a length of 598ms to clip the sound waveforms. For each time step, the window is moved forward by $1/24$s, i.e., 24 fps and the coefficient λ is set to 20 to balance the values of the loss functions. Adam optimizer [14] is used for optimization with exponentially decaying rate, starting from $5e^{-5}$ for 15 epochs.

Method	Visual	Transform	Seen Scenes STFT ↓	Seen Scenes DPAM ↓	Seen Scenes ENV ↓	Unseen Scenes STFT ↓	Unseen Scenes DPAM ↓	Unseen Scenes ENV ↓
Replica: 12 seen scenes, 6 unseen scenes								
Nearest	✗	✓	1.614	0.992	0.257	1.686	0.993	0.277
Mean	✗	✓	1.600	1.039	0.265	1.618	1.036	0.275
Interpolation	✗	✓	1.575	1.039	0.256	1.614	1.033	0.267
AViTAR [6]	✓	✗	0.181	0.334	0.163	0.199	0.327	0.184
Few-shotRIR [17]	✓	✗	0.233	0.449	0.227	0.245	0.436	0.239
Mono2Binaural [10]	✓	✓	0.194	0.376	0.156	0.236	0.364	0.177
APNet [30]	✓	✓	0.164	0.263	0.154	0.185	0.253	0.176
BEE (Ours)	✓	✓	**0.151**	**0.215**	**0.133**	**0.177**	**0.221**	**0.160**
Matterport3D: 54 seen scenes, 25 unseen scenes								
Nearest	✗	✓	4.851	1.047	0.837	5.029	1.064	0.874
Mean	✗	✓	3.174	1.068	0.611	3.456	1.078	0.650
Interpolation	✗	✓	3.475	1.066	0.658	3.521	1.081	0.669
AViTAR [6]	✓	✗	0.516	0.610	0.595	0.509	0.625	0.548
Few-shotRIR [17]	✓	✗	0.597	0.476	0.731	0.591	0.500	0.694
Mono2Binaural [10]	✓	✓	0.533	0.440	0.545	0.582	0.492	0.529
APNet [30]	✓	✓	0.500	0.352	0.537	0.515	0.393	0.528
BEE (Ours)	✓	✓	**0.425**	**0.274**	**0.455**	**0.438**	**0.348**	**0.458**

Table 1: **Testing results comparison on the SoundSpaces Dataset with Replica and Matterport3D scenes.** *Visual* and *Transform* indicate using visual information and learning transformation of input sounds to render target sounds respectively.

4.3. Baselines

Non-learning baselines. 1) *Nearest*: The output sound of the reference receiver closest to the location of the target listener; 2) *Mean*: The average of all received sound waveforms from reference receivers $\{S_i|i = 1, 2, ...N\}$; 3) *Interpolation*: Linear interpolation of $\{S_i|i = 1, 2, ...N\}$ by merging N sounds with the location-relevant weights. The weight of each receiver is the inverse proportion of the distance to the listener and normalization through Softmax.

Existing audio-visual sound synthesis solutions. 1) Visual Acoustic Matching (*AViTAR*) [6]: AViTAR re-synthesizes the audio to match the target room acoustics given an image and one audio clip as input; 2) *Mono2Binaural* [10]: Mono2Binaural learns to decode the monoaural soundtrack into its binaural counterpart by injecting visual information about object and scene configurations; 3) Associative Pyramid Network (*APNet*) [30]: APNet associates the visual features and the audio features to boost the performance on stereophonic audio generation and audio source separation tasks; 4) *Few-shotRIR* [17]: a transformer-based method that infers RIRs based on a sparse set of observed images and echoes. Among these methods, APNet and Mono2Binaural learn complex masks to transform the input audio spectrograms (*Transform*). To adapt these methods to our task, we inject the pose information of the receiver-listener pairs into the audio-visual features, generate binaural spectrograms for target listeners, and train with the same strategies as BEE.

4.4. Main Results Comparison

We compare BEE with baseline approaches on Replica and Matterport3D scenes and report the results in Table 1. For Replica scenes, we randomly select 12 scenes ranging from $9.5m^2$ to $141.5m^2$ for training and the remaining 6 scenes as unseen scenes. For Matterport3D, we select 54 scenes with an average size over $100m^2$ as training scenes and 25 other scenes as unseen scenes. Each Matterport3D scene contains multiple individual rooms and complex layouts. We sample emitter-receiver-listener pairs at each training scene for training, and test both seen and unseen scenes respectively on new emitter-receiver-listener pairs and source sound clips. *Transform* indicates transforming the input audio to the target audio.

Comparison with non-learning baselines. As Table 1 shows these three baselines perform significantly less accurate than the learning-based methods on sparse A/V sensors-based scene audio reconstruction task.

Comparison with existing audio-visual based methods. While these methods are more accurate than non-learning methods, we observe a significant accuracy gain for BEE vs. other learning methods, especially on Matterport3D scenes that are of larger scale and of more complex layouts. Compared with *AViTAR*, which directly generates target sound based on audio-visual features without *Transform*, BEE outperforms *AViTAR* by 13.9%, 44.3%, 16.4% on STFT, DPAM, ENV metrics respectively. While *Transform* assists *Mono2Binaural* and *APNet* to achieve better accuracy on both datasets, BEE achieves higher accuracy than these methods. Compared with *APNet*, the second-best method in Table 1, BEE enhances the accuracy by achieving STFT, DPAM, and ENV metrics better by 14.95%, 12.6% and 9.1% respectively on Matterport3D.

Variants	Module	STFT ↓	DPAM ↓	ENV ↓
w/o. 3DAV Enc.	JAVR	0.465	0.373	0.512
w/o. Vis Atten.	JAVR	0.454	0.363	0.478
w/o. Integrate	IRH	0.509	0.400	0.554
w/o. AV UpConv	IRH	0.490	0.424	0.470
w/o. Mag Branch	IRH	0.626	0.393	0.518
Full Model	-	**0.438**	**0.348**	**0.458**

Table 2: **Ablations on Matterport3D unseen scenes**.

4.5. Ablation Study

Joint Audio-Visual Representation. To verify the contribution of the Joint Audio-Visual Representation module in Section 3.2, we implement two variants that remove Audio-Visual Spatial Encoder (*w/o. 3DAV Enc.*) and the Spatial Enhanced Visual Representation module (*w/o. Vis Attn.*) respectively. As observed from the first two rows of the ablations in Table 2, BEE boosts the accuracy of STFT, DPAM and ENV by integrating the obtained visual and audio representation on the acoustic propagation feature space with respect to each spatial point within the space. Particularly notable is 10.5% improvement on ENV metric. *w/o. Vis Atten.* underperforms BEE on all the metrics, demonstrating the contribution of The Spatial Enhanced Visual Representation module to the audio reconstruction accuracy by enhancing visual features for each receiver-listener pair with their corresponding spatial locations.

Integrated Rendering. Considering N reference receivers at different locations, Integrated Rendering Head (IRH) in Section 3.3 first utilizes a Cross-Attention module at the beginning to learn an integrated audio-visual feature embedding w.r.t each reference receiver and performs a weighted average over the input sounds at the end to fully incorporate the observed information from the A/V sensors. Ablations in Table 2, removing the Cross-Attention module and directly supervising each rendered sound without averaging them (*w/o. Integrate*), indicate the necessity of integrated rendering strategies by significantly inaccurate metrics. With these strategies, BEE boosts the accuracy of STFT, DPAM and ENV by 13.9%, 13% and 17.3%.

Decoupled Spectrogram Weights Branches. IRH component of BEE introduces two decoupled branches to generate the final spectrogram transformation weights, utilizing different audio-visual feature combinations. To further investigate the effectiveness of these two branches, we introduce two variants: *w/o. AV UpConv* and *w/o. Mag Branch*, which removes the Transposed Convolutional Layer branch and the MLP-based Magnitude TF weights prediction branch respectively. As shown in Table 2, *UpConv* branch is advantageous for perceptual quality (DPAM) than the *Magnitude* branch, while the *Magnitude* branch achieves higher accuracy on STFT and ENV. This can be interpreted as follows. *UpConv* branch starts from the high-level low-resolution features of input audio to gradually generate high-resolution target spectrogram weight masks. This improves the quality of the generated results by incorporating dependence on the detailed input audio contents. However, in addition, dependence on the comprehensive audio-visual scene representation and spatial relationship is needed to infer the time delay and sound energy changes from sensors to listeners. BEE handles both of these aspects by deploying decoupled *UpConv* and *Magnitude* branches.

Components	3D Vis Enc	JAVR	IRH	Total
Speed (ms/sample)	16.00	18.40	11.94	30.34

Table 3: **Speed Analysis**. BEE can run at a real-time speed.

4.6. Speed Analysis

To define real-time operation, we focus on the ability to render audio output within specified time constraints of exceeding the input sound frame rate of 24fps (frame per second). We use one GeForce RTX 2080 Ti for speed testing and report the results in Table 3, with values averaged for 500 samples. For each sample, BEE takes 30.34ms (around 33fps) for rendering, achieving real-time for inference. Among the 30.34ms, the 3D Visual Encoder in JAVR takes 16ms, while other components take 14.24ms in total. Since BEE does not require emitter information as input, as long as the count of reference A/V sensors remains fixed, rendering time complexity remains consistent.

4.7. Qualitative Results

In Figure 4 we visualize some of BEE results for qualitative interpretation. We show: 1) loudness maps, and 2) generated waveforms & spectrograms. In comparison with other audio-visual sound synthesis methods, BEE generates more reliable loudness maps consistent with scenes layouts and better reconstructs the continuous change of the loudness in the scene while navigating through it for both seen and unseen scenes. This is observed by plotting the cumulated loudness error of all the spatial points in a scene. BEE achieves a significantly lower accumulated loudness map error for both seen and unseen scenes. Furthermore, other methods have significant gaps in waveform generation and additional noise in the generated spectrograms. Notably, compared to the second-best method APNet, BEE achieves a reduction of 0.14 DPAM and 0.03 ENV error, which enhanced waveform accuracy significantly, as depicted in the waveform plots.

4.8. Human Subject Study

To further evaluate the perceptual quality of the generated sound, we conducted a human subject study with 250 responses from more than 50 different people evaluating generated sound on SoundSpaces Dataset. For each sample, the listener is navigated through a scene with random emitters and emitted sounds. Participants are shown

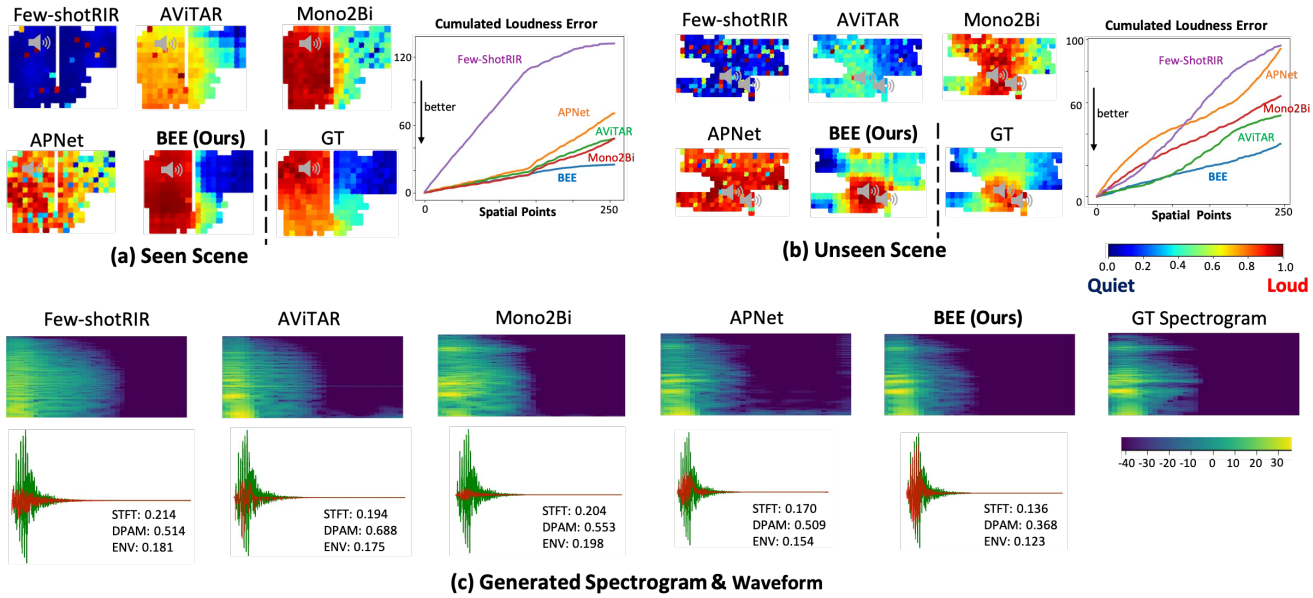

Figure 4: **Qualitative Examples Comparison.** Top (a,b): Loudness maps of seen and unseen scenes respectively. Grey speakers represent the emitters. The curve plots represent accumulated loudness error for all navigatable points in the scenes (lower is better - zero curve is GT). Bottom(c): visualizes generated waveforms and spectrograms and compares BEE with previous methods (Green: GT; Red: Prediction).

Methods	Mono2Binaural [10]	APNet [30]	BEE (Ours)
Votes	29.6%	26.4%	**44%**

Table 4: **Human Subject Study**. BEE is preferred over the other two methods by a large margin.

the visual observation at each listener point, the navigation routes on the scene layout, and the ground truth target sound. Three generated audio samples by BEE, *APNet* and *Mono2Binaural* respectively are provided for participants to choose the sample that sounded most similar to the ground truth. As Table 4 indicates, BEE is preferred by participants over the other two methods achieving 44% preference ratio, higher by 14.4% than the second-best method.

4.9. Discussion

Non-learning baselines such as *Nearest*, *Mean*, and *Interpolation* are ineffective in handling random noise in audio samples or in warping audio samples based on scene properties. Learning baselines such as *AViTAR*, *Mono2Binaural*, and *APNet* integrate audio and visual feature maps with pose information to synthesize target sound, but designed to infer spatial relationships, emitter actions, and scene properties from a single A/V sensor. *FewshotRIR* integrates audio-visual information from sensors, but it does not use the observations from all sensors in full.

In contrast, BEE incorporates JAVR module to explicitly model the 3D visual volume and to integrate audio-visual features with respect to sensor and target listener poses, allowing for scene property capture and representative audio-visual feature learning. The IRH module enhances high-level audio-visual representation for each receiver-listener pair with other pairs, utilizing two branches to generate reliable binaural and magnitude transformation weights based on different levels of audio-visual features. Moreover, BEE does not rely on explicit input of emitter locations or emitted sound waveforms which can be challenging to obtain without special scene setup in real practice. In this case, receivers are not directly associated with emitters, allowing handling of unknown numbers, locations, and source sounds of emitters in dynamic scenes.

In summary, here we propose a real-time generalizable end-to-end integrated rendering pipeline (BEE), which reconstructs the audio of a scene at an arbitrary location of the listener according to inputs from A/V receivers sparsely placed in the scene. To reconstruct the audio, BEE utilizes a Joint Audio-Visual Representation module to extract the informative audio-visual features of the scene, and then integrates audio samples with the learned time-frequency transformations using the Integrated Rendering Head. Our experiments indicate that BEE outperforms existing methods in all metrics and the results appear to be generalizable to unseen scenes.

Acknowledgement We acknowledge the partial support of National Science Foundation grant OAC-2117997 and the departments of Applied Mathematics and Electrical and Computer Engineering at the University of Washington.

References

[1] Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979. 2

[2] Sercan Ö. Arık, Heewoo Jun, and Gregory Diamos. Fast spectrogram inversion using multi-head convolutional neural networks. *IEEE Signal Processing Letters*, 26(1):94–98, 2019. 5

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

[4] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 5

[6] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 2, 3, 6

[7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision*, pages 17–36. Springer, 2020. 2, 5

[8] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations*, 2019. 2, 3

[9] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. GANSynth: Adversarial neural audio synthesis. In *International Conference on Learning Representations*, 2019. 2, 3

[10] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2, 3, 5, 6, 8

[11] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. *CVPR*, 2018. 4

[12] Nail A Gumerov and Ramani Duraiswami. A broadband fast multipole accelerated boundary element method for the three dimensional helmholtz equation. *The Journal of the Acoustical Society of America*, 125(1):191–205, 2009. 2

[13] Brian Hamilton and Stefan Bilbao. Fdtd methods for 3-d room acoustics simulation with high-order accuracy in space and time. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2112–2124, 2017. 2

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[15] Asbjørn Krokstad, Staffan Strom, and Svein Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968. 2

[16] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *arXiv preprint arXiv:*, 2021. 2

[17] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics, 2022. 2, 6

[18] Pranay Manocha, Adam Finkelstein, Richard Zhang, Nicholas J. Bryan, Gautham J. Mysore, and Zeyu Jin. A differentiable perceptual audio metric learned from just noticeable differences. In *Interspeech*, Oct. 2020. 5

[19] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 5

[20] Eva-Marie Nosal, Murray Hodgson, and Ian Ashdown. Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms. *The Journal of the Acoustical Society of America*, 116(2):970–980, 2004. 2

[21] Nikunj Raghuvanshi, Rahul Narain, and Ming C Lin. Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):789–801, 2009. 2

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

[23] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. *arXiv preprint arXiv:2110.04057*, 2021. 2

[24] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. *arXiv preprint arXiv:2202.03416*, 2022. 2

[25] Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021. 2, 3

[26] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[27] Lonny L Thompson. A review of finite-element methods for time-harmonic acoustics. *The Journal of the Acoustical Society of America*, 119(3):1315–1330, 2006. 2

[28] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *Arxiv*, 2016. 2, 3

[29] Michael Vorländer. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America*, 86(1):172–178, 1989. 2

[30] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer, 2020. 2, 3, 6, 8