

# AV-NeRF: Learning Neural Fields for Real-World Audio-Visual Scene Synthesis

Susan Liang  
University of Rochester  
sliang22@ur.rochester.edu

Chao Huang  
University of Rochester  
chuang65@cs.rochester.edu

Yapeng Tian  
University of Rochester  
yapeng.tian@utdallas.edu

Anurag Kumar  
Meta Reality Labs Research  
anuragkr90@meta.com  
Homepage: <https://liangsusan-git.github.io/project/avnerf/>

## Abstract

*Human perception of the complex world relies on a comprehensive analysis of multi-modal signals, and the co-occurrences of audio and video signals provide humans with rich cues. This paper focuses on novel audio-visual scene synthesis in the real world. Given a video recording of an audio-visual scene, the task is to synthesize new videos with spatial audios along arbitrary novel camera trajectories in that audio-visual scene. Directly using a NeRF-based model for audio synthesis is insufficient due to its lack of prior knowledge and acoustic supervision. To tackle the challenges, we first propose an acoustic-aware audio generation module that integrates our prior knowledge of audio propagation into NeRF, in which we associate audio generation with the 3D geometry of the visual environment. In addition, we propose a coordinate transformation module that expresses a viewing direction relative to the sound source. Such a direction transformation helps the model learn sound source-centric acoustic fields. Moreover, we utilize a head-related impulse response function to synthesize pseudo binaural audio for data augmentation that strengthens training. We qualitatively and quantitatively demonstrate the advantage of our model on real-world audio-visual scenes. We refer interested readers to view our video results for convincing comparisons.*

## 1. Introduction

Vision and sound play essential roles in human perception of the surrounding scene. These two modalities contain not only semantic information (*e.g.*, the class of objects and the content of speech) but also spatial information (*e.g.*, the position of sound sources). Our brain can analyze and integrate different modalities to thoroughly understand the surrounding environment. Naturally, the absence of either modality hinders our sense of the physical world. Recog-

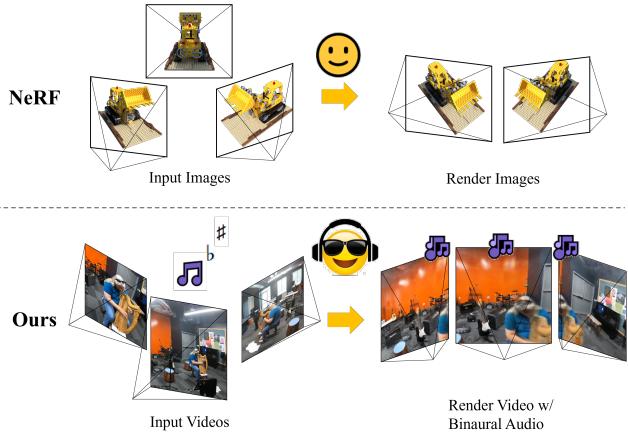


Figure 1. NeRF learns to render visual scenes at novel poses. Beyond visual rendering, we present AV-NeRF for learning to synthesize audio-visual consistent scenes including video frames at a novel view and the corresponding binaural audio in the environment. Consistent sight and sound can provide users with an immersive and realistic perceptual experience.

nizing this, the machine perception research community has seen a spectrum of works [2, 3, 8, 9, 17, 20, 24, 25, 27, 30, 31] proposed to learn and model auditory and visual signals jointly.

Different from past audio-visual learning works, this paper focuses on the synthesis of novel audio-visual scenes in the real world. We define novel audio-visual scene synthesis as a task to synthesize a target video, including visual frames and the corresponding spatial audio, along an arbitrary camera trajectory from given source videos and trajectories. Learning from source video in a real-world environment with binaural audio, the generated target spatial audio and video frames are expected to be consistent with the given camera trajectory visually as well as acoustically to ensure perceptual realism and immersion.

Although there are some similar works [6, 15], these

methods have some constraints that limit their usage in solving our problem. Luo *et al.* [15] propose neural acoustic fields to model sound propagation in a room. However, their model works in a simulation environment and relies on ground-truth acoustic labels that are difficult to obtain in a real-world scene. Du *et al.* [6] propose a manifold learning method that can map a latent vector to a pair of image and audio. However, the latent vectors are uninterpretable and the manifold cannot support the controllable generation of audio-visual pairs.

In this paper, we propose a novel NeRF-based method for synthesizing real-world audio-visual scenes, dubbed AV-NeRF. NeRF [16] implicitly and continuously represents the visual scene using neural fields and the neural fields can be used to render novel views. Although NeRF has achieved impressive success in the vision domain, directly using NeRF in our task does not yield the best performance because of the lack of prior knowledge and acoustic supervision. NeRF uses the classical volume rendering method [11] to render the color of a ray based on the physical property of light propagation. However, there is no prior knowledge for NeRF-based audio synthesis. Moreover, the recording mechanisms of these two signals are different. At a given camera pose, we can capture an image with millions of pixels but can only record one acoustic signal (assuming the acoustic property of the environment is time-invariant). Each pixel can serve as visual supervision while the whole acoustic signal is the auditory supervision. The imbalance between the amounts of visual data and audio data hinders the optimization of NeRF for audio synthesis. To tackle these challenges, we propose an acoustic-aware audio generation module that incorporates our prior knowledge of audio propagation into NeRF, a coordinate transformation mechanism that encourages NeRF learning sound source-centric neural fields, and a binaural audio augmentation method to improve acoustic supervision. To this end, our AV-NeRF is capable of generating audio-visual consistent scenes using the same amount of data.

Considering that 3D geometry and material property of an environment determine the sound propagation (*e.g.*, the existence of a barrier between the sound source and the sound receiver indicates the decay of sound energy), we propose an acoustic-aware audio generation module, named as AV-Bridge, that correlates the spatial effects of the sound with the 3D visual geometry. Visual NeRF can generate not only the color  $c$  of one point in 3D space but also its density  $\sigma$  [16]. The density function of an environment serves as an important indicator of 3D geometry. Our AV-Bridge extracts density information from the visual NeRF to obtain a voxel grid and learns its latent representation. We feed this latent representation to audio NeRF and enforce audio NeRF learning acoustic-aware audio generation.

In vanilla NeRF [16], viewing direction  $(\theta, \phi)$  is ex-

pressed in an absolute coordinate system. This coordinate expression is a natural practice in visual space such that the same orientation at different spatial positions is expressed equally. However, this expression method in audio space is sub-optimal. Human perception of the sound direction is based on the relative direction to the sound source instead of the absolute direction. For example, when a person walks around an omnidirectional sound source and keeps facing the sound source, his sensation of the sound source position is constant but his absolute orientation varies. Considering this fact, we propose a coordinate transform mechanism that expresses the viewing direction relative to the sound source to encourage our model to learn a sound source-centric acoustic field.

Unlike simulation environments [3,4] where the ground-truth acoustic information is easily accessible, samples in real-world scenes are captured sparsely and often can not cover the whole environment. To render a realistic video at novel poses, we utilize a head-related impulse response function (HRIR) to augment training data. HRIR serves as a sound filter that represents changes in sound with respect to the position of the sound source. This augmentation strategy can improve acoustic supervision and enhance the spatial effects of rendered audio at novel poses.

In summary, our contributions include: (1) proposing a novel method of synthesizing real-world audio-visual scenes at novel positions and directions; (2) introducing a novel acoustic-aware audio generation method to encode our prior knowledge of sound propagation; (3) proposing a coordinate transformation mechanism for effective direction expression; (4) introducing a binaural audio augmentation method to improve the acoustic supervision; (5) quantitatively and qualitatively demonstrating advantages of our method. In the paper, we target modeling acoustic fields and establishing the correlation between visual and acoustic worlds. The rendering of the purely visual world has been well studied [16, 19, 22, 29] and is not the main focus of our work.

## 2. Related Work

**Neural Fields.** Our method is based on the neural fields, especially Neural Radiance Fields (NeRF) [16]. NeRF learns an implicit and continuous representation of visual scenes using a neural network and synthesizes novel views based on the trained neural network. After the presence of NeRF, several works extend NeRF to broader domains, such as video (silent video) [7, 13, 14], audio [15], and audio-visual [6]. Luo *et al.* [15] propose neural acoustic fields (NAF) to capture how sound propagates in an environment. Although NAF achieves convincing results, its reliance on synthetic environments and ground truth impulse response functions hinder its application in real-world scenes. Their discretization of position and direction (the

listener can only move in a fixed x-y grid and rotate 0°, 90°, 180°, and 270°) restricts the camera motion while our model can handle continuous position and direction input. Du *et al.* [6] propose a manifold learning method that maps vectors in the latent space to audio and image space. While the learned manifold can be used for audio-visual interpolation, the model does not support controllable audio-visual generation because the latent space does not contain interpretable spatial and temporal information. Our method instead learns implicit neural representations that are explicitly conditioned on spatial coordinates enabling controllable real-world scene generation.

**Visually Informed Spatial Audio Generation.** Spatial audio can not only help people localize the sound source but also establish an immersive experience in 3D environments. Recently, many visually informed audio spatialization approaches [8, 18, 28, 32] have been proposed. Among them, Gao *et al.* [8] focus on normal field-of-view video and binaural audio. Zhou *et al.* [32] propose a unified framework to solve the sound separation and stereo sound generation at the same time. Xu *et al.* [28] propose a PseudoBinaural pipeline that correlates spatial locations with binaural audio people receive using HRIR function [1]. Although these methods generate good binaural audio, they assume the ground-truth visual frames are available, which does not hold when the model generates audio at a novel pose. Our method does not rely on ground-truth images for audio spatialization because our AV-NeRF learns an acoustic field that is only conditioned on the pose. Moreover, our model uses two orders of magnitude less data for training than their methods.

**Geometry and Material Based Acoustic Simulation.** Another line of work [3, 4, 12, 26] focuses on acoustic simulation by modeling an environment’s 3D geometry and material property. Chen *et al.* [3] propose a SoundSpaces audio platform where the room impulse response function of given discrete listener and sound source positions can be calculated by modeling the geometry and material property of an environment. Extending the work, Chen *et al.* [4] further propose SoundSpaces 2.0, which supports realistic acoustic generation for arbitrary sounds captured from arbitrary microphone locations in 3D environments. Li *et al.* [12] propose blending the early reverberation part modeled by geometric acoustic simulation and frequency modulation with a late reverberation tail extracted from recorded impulse response. Tang *et al.* [26] propose estimating reverberation time ( $T_{60}$ ) and equalization (EQ) using a learning-based acoustic analysis method. The predicted  $T_{60}$  and EQ are then used for material optimization enabling audio rendering. Unlike these virtual rendering approaches, our method learns from and synthesizes real-world audio-visual scenes conditioned on the learned 3D geometry of an environment.

### 3. Method

Our method learns neural fields for synthesizing real-world audio-visual scenes at novel poses. When training AV-NeRF, we feed the model with several video clips (with binaural audio) and corresponding camera trajectories when capturing these video clips. We encourage AV-NeRF learning a mapping from camera trajectories to video clips. At inference time, we feed AV-NeRF with an arbitrary camera trajectory and expect the model to output a target video that is consistent with the input camera trajectory visually and acoustically. The whole pipeline is illustrated in Fig. 2. Our model consists of three trainable modules: V-NeRF, A-NeRF and AV-Bridge. V-NeRF learns to generate acoustic masks, A-NeRF learns to generate visual frames and AV-Bridge is optimized to extract geometric information from V-NeRF and integrate this information into A-NeRF.

#### 3.1. V-NeRF

NeRF [16] uses a Multi-Layer Perceptron (MLP) to represent a visual scene implicitly and continuously. It learns a mapping from camera poses to colors and densities:

$$\text{NeRF} : (x, y, z, \theta, \phi) \rightarrow (\mathbf{c}, \sigma) , \quad (1)$$

where  $\mathbf{X} = (x, y, z)$  is the 3D position,  $\mathbf{d} = (\theta, \phi)$  is the direction,  $\mathbf{c} = (r, g, b)$  is the color, and  $\sigma$  is the density. To render view-dependent color  $\mathbf{c}$  and ensure multiview consistency, NeRF first maps a 3D coordinate  $(x, y, z)$  (we apply positional encoding to all input coordinates, unless otherwise noted) to density  $\sigma$  and a feature vector; then NeRF maps the feature vector and 2D direction  $(\theta, \phi)$  to a color  $\mathbf{c}$ . This process is illustrated in Fig. 3a.

NeRF then uses the volume rendering method [11] to generate the color of any ray  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  marching through the visual scene with near and far bounds  $t_n$  and  $t_f$ :

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt , \quad (2)$$

where  $T(t) = \exp(-\int_{t_n}^t (\mathbf{r}(s))ds)$  and  $\mathbf{d} = (\theta, \phi)$  (expressed as a 3D Cartesian unit vector). This continuous integral is estimated by quadrature in practice. Considering that our model learns a neural representation of real-world scenes, we do not apply normalized device coordinate (NDC) transformation to 3D coordinates.

#### 3.2. A-NeRF

The target of A-NeRF is to learn a neural acoustic representation that can map 5D coordinates  $(x, y, z, \theta, \phi)$  to corresponding acoustic masks  $\mathbf{m}_m, \mathbf{m}_d \in \mathcal{R}^{2 \times F}$ , where  $\mathbf{m}_m$  means the change of magnitude and phase of sound w.r.t. the position  $(x, y, z)$  while  $\mathbf{m}_d$  means the change of magnitude and phase of sound w.r.t. the direction  $(\theta, \phi)$ , and  $F$  is

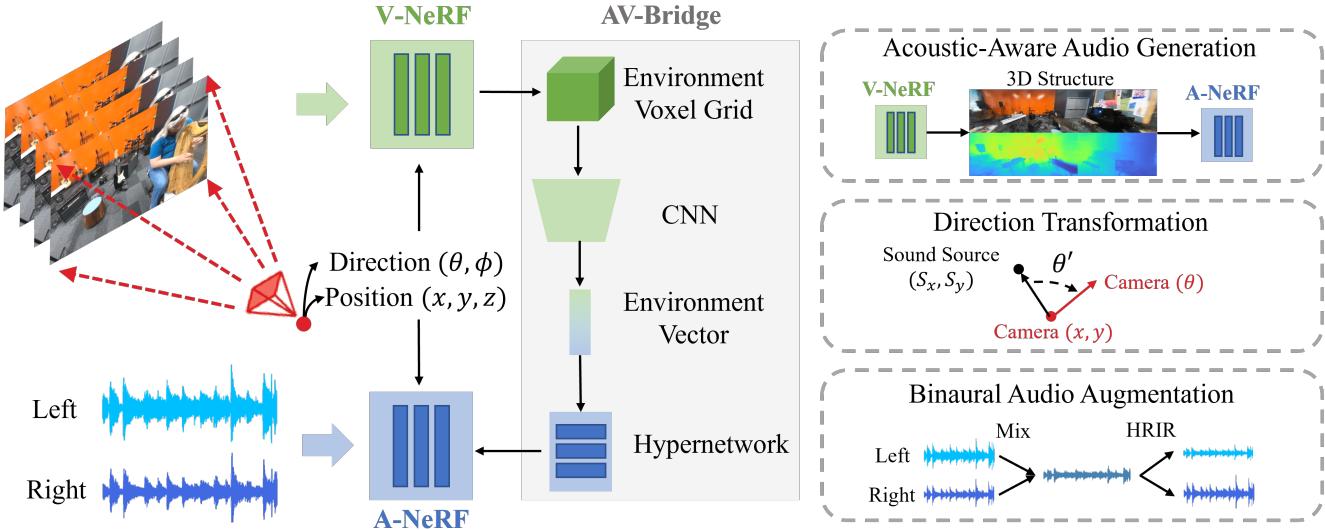


Figure 2. The pipeline of our method. Given the position  $(x, y, z)$  and viewing direction  $(\theta, \phi)$  of a listener, our method can render an image the listener would see and the corresponding binaural audio the listener would hear. Our model consists of V-NeRF, A-NeRF, and AV-Bridge. V-NeRF learns to generate acoustic masks, A-NeRF learns to generate visual frames and AV-Bridge is optimized to extract geometric information from V-NeRF and incorporate this information into A-NeRF.

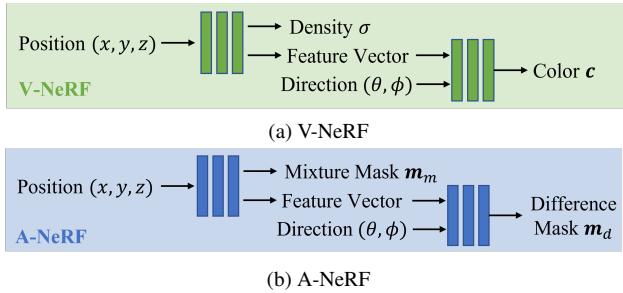


Figure 3. Architecture of V-NeRF and A-NeRF.  
the number of frequency bins:

$$\text{NeRF} : (x, y, z, \theta, \phi) \rightarrow (\mathbf{m}_m, \mathbf{m}_d) . \quad (3)$$

In practice, as shown in Fig. 3b, we feed A-NeRF with 3D position  $(x, y, z)$  to obtain a mixture mask  $\mathbf{m}_m$  and a feature vector. Then we concatenate this feature vector with the input direction  $(\theta, \phi)$  and pass it to the rest part of A-NeRF to generate a difference mask  $\mathbf{m}_d$ .

For the sake of clarity, we use  $\hat{a}$  to represent predicted audio and  $a$  to represent ground-truth audio in the rest of paper. Given a sound  $a_s$ , we first utilize short-time Fourier transform (STFT) to convert input audio from the time domain to the time-frequency domain  $\text{STFT}(a_s) = s_s \in \mathcal{R}^{2 \times F \times W}$ , where  $W$  is the number of time frames. Then, we multiply  $s_s$  and  $\mathbf{m}_m$  using complex multiplication to predict the spectrum of the mix sound  $\hat{s}_m \in \mathcal{R}^{2 \times F \times W}$  at target position  $(x, y, z)$ . To enable binaural audio generation, we further multiply the predicted spectrum  $\hat{s}_m$  and  $\mathbf{m}_d$  to calculate the difference between left and right channels  $\hat{s}_d \in \mathcal{R}^{2 \times F \times W}$ . Finally, we use inverse short-

time Fourier transform (ISTFT) to generate binaural audio  $\hat{a}_l, \hat{a}_r$ . We define  $\hat{a}_m = \hat{a}_l + \hat{a}_r$  and  $\hat{a}_d = \hat{a}_l - \hat{a}_r$ , so  $(\hat{a}_m + \hat{a}_d)/2 = (\hat{a}_l + \hat{a}_r + \hat{a}_l - \hat{a}_r)/2 = \hat{a}_l$  and  $(\hat{a}_m - \hat{a}_d)/2 = (\hat{a}_l + \hat{a}_r - \hat{a}_l + \hat{a}_r)/2 = \hat{a}_r$ . This process can be formulated mathematically in Eq. 4. Instead of estimating the left and right channels of the target audio directly, we predict the mix of target audio  $\hat{a}_m$  and the difference between two channels of the target audio  $\hat{a}_d$ . Gao and Grauman [8] indicate that directly predicting two channels makes audio spatialization network learn some shortcuts.

$$\begin{aligned} \hat{a}_m &= \text{ISTFT}(\hat{s}_m) \\ &= \text{ISTFT}(s_s * \mathbf{m}_m) , \\ \hat{a}_d &= \text{ISTFT}(\hat{s}_d) \\ &= \text{ISTFT}(\hat{s}_m * \mathbf{m}_d) \\ &= \text{ISTFT}(s_s * \mathbf{m}_m * \mathbf{m}_d) , \\ \hat{a}_l &= (\hat{a}_m + \hat{a}_d)/2 , \\ \hat{a}_r &= (\hat{a}_m - \hat{a}_d)/2 . \end{aligned} \quad (4)$$

### 3.3. AV-Bridge

Given the fact that 3D geometry and material property determine the sound propagation in an environment, we propose an acoustic geometry-aware audio generation method by integrating 3D visual geometry information learned by V-NeRF with A-NeRF. When V-NeRF learns to represent visual scenes, thanks to the multiview consistency constraint, it can capture the density function  $\sigma$  of the environment. Such density function represents the existence of objects within the environment. So we extract geometric in-

formation from V-NeRF and encode it into a feature vector for acoustic-aware audio generation.

Specifically, we query V-NeRF with discrete 3D points that are uniformly scattered in the environment. We compose the output volume density into an environment voxel grid, which represents the 3D structure of the scene. We then use a convolutional neural network to encode this voxel grid into a compact environment vector. After obtaining the environment vector, we propose a Hypernetwork [10] to utilize this geometric information for acoustic-aware audio generation. We design a Hypernetwork  $\psi$  to convert the environment vector  $v$  into parameters  $\mathbf{W}_A$  of A-NeRF inspired by [6]:

$$\psi : v \rightarrow \mathbf{W}_A . \quad (5)$$

For each learnable linear layer  $\mathbf{W}_i \in \mathcal{R}^{m \times n}$  in A-NeRF, we train a three-layer MLP to output a weight matrix  $M$  of the same shape as  $\mathbf{W}_i$ . The input of each MLP is the environment vector  $v$ . The matrix  $M$  is fused with the parameters  $\mathbf{W}_i$  to generate new parameters for guiding audio generation:

$$\mathbf{W}_i \leftarrow \mathbf{W}_i \odot M , \quad (6)$$

where  $\odot$  is Hadamard product. Directly predicting high-dimensional matrix  $M$  is not only computation-consuming but also difficult to optimize. So we decompose the high-dimensional matrix  $M \in \mathcal{R}^{m \times n}$  into two low-dimensional matrices  $A \in \mathcal{R}^{m \times k}$  and  $B \in \mathcal{R}^{k \times n}$  [23], where  $k < m$  and  $k < n$ , and express matrix  $M$  as  $M = \sigma(A \times B)$ . We optimize MLP to generate these two low-dimensional matrices instead of the original matrix.

### 3.4. Coordinate Transformation

Viewing direction  $(\theta, \phi)$  in V-NeRF is expressed in an absolute coordinate system. This is a natural practice in visual space such that the same orientation at different spatial positions is expressed equally. As discussed in the introduction, however, this expression method in audio space is sub-optimal because the human perception of the sound direction is based on the relative direction to the sound source instead of the absolute direction. To overcome this shortage, we propose expressing viewing direction relative to the sound source. This coordinate transformation encourages A-NeRF learning a sound source-centric acoustic field.

Given the 3D position of the sound source  $\mathbf{X}_s = (x_s, y_s, z_s)$  and camera pose  $(\mathbf{X}, \mathbf{d}) = (x, y, z, \theta, \phi)$ , we obtain two direction vectors:  $\mathbf{V}_1 = \mathbf{X}_s - \mathbf{X} = (x_s - x, y_s - y, z_s - z)$  and  $\mathbf{V}_2 = (\sin(\theta) \cos(\phi), \sin(\theta) \sin(\phi), \cos(\theta))$ . We calculate the angle between  $\mathbf{V}_1$  and  $\mathbf{V}_2$  as the relative direction coordinates  $\angle(\mathbf{V}_1, \mathbf{V}_2)$ . This angle  $\angle(\mathbf{V}_1, \mathbf{V}_2)$  represents the rotation angle relative to the sound source. With this coordinate transformation, different camera poses can share the same view direction encoding if they face the sound source at the same angle. We will express this angle as a 2D Cartesian unit vector when feeding it to A-NeRF.

### 3.5. Binaural Audio Augmentation

Capturing high-quality binaural audio requires a professional recording system such as a binaural microphone and dummy head that mimics the sound people can hear [8]. This requirement restricts the application of our method in videos captured by commodity devices such as iPhones and GoPro. Although several commodity devices support recording stereo sound, the microphone arrangement is arbitrary and can not imitate the sound effects caused by human heads. So we apply head-related impulse response (HRIR) to the stereo audio to generate binaural audio with rich spatial information following Xu *et al.* [28]. HRIR is a response function that characterizes the influence of the human head and sound position on sound propagation. We exploit an open-sourced HRIR database [1] for binaural audio augmentation. Specifically, we retrieve a pair of HRIR functions  $h_l$  and  $h_r$  of a given angle from the database and convolve  $h_l$  and  $h_r$  with the mixed ground-truth stereo audio  $(a_l + a_r)/2$  to generate a binaural audio (Eq. 7). Please note that the mix of audio for data augmentation is different from that for modeling training. We devide the mix of two channels by 2 to get the average signal for data augmentation.

$$a_l \leftarrow h_l \circledast (a_l + a_r)/2 , \\ a_r \leftarrow h_r \circledast (a_l + a_r)/2 , \quad (7)$$

where  $\circledast$  is the convolution operator. By augmenting training audio, we encourage our model to correlate acoustic effects with the position of the sound source.

### 3.6. Learning Objective

We refer to the combination of V-NeRF (Sec. 3.1) and A-NeRF (Sec. 3.2) as the baseline method. We integrate AV-Bridge (Sec. 3.3), coordinate transformation module (Sec. 3.4), and data augmentation mechanism (Sec. 3.5) into the baseline method to assemble our AV-NeRF model. Because AV-Bridge is optimized together with A-NeRF and the coordinate transformation module and data augmentation mechanism do not contain learnable parameters, the baseline method and AV-NeRF are optimized using the same learning objective.

The loss function of V-NeRF is the same as [16]:

$$\mathcal{L}_V = ||C(\mathbf{r}) - \hat{C}(\mathbf{r})||^2 , \quad (8)$$

where  $C(\mathbf{r})$  is the ground-truth color along the ray  $\mathbf{r}$  and  $\hat{C}(\mathbf{r})$  is the color rendered by V-NeRF.

We use the L2 loss function to supervise A-NeRF. Given a mono source audio  $a_s$  and a binaural target audio  $a_t$ , we calculate the mix audio  $a_m = a_{t(l)} + a_{t(r)}$ , the difference audio  $a_d = a_{t(l)} - a_{t(r)}$ , and spectrums of  $a_s$ ,  $a_m$ , and  $a_d$ , which are  $s_s$ ,  $s_m$ , and  $s_d$ , respectively. Then we minimize the distance between calculated spectrums and pre-



Figure 4. Recording devices and two representative indoor scenes.



Figure 5. Example scene re-collected from FAIR-PLAY dataset. We collect video clips belonging to the same scene but recorded at different camera poses.

dicted spectrums:

$$\begin{aligned} \mathcal{L}_A &= \|s_m - \hat{s}_m\|^2 + \|s_d - \hat{s}_d\|^2 \\ &= \|s_m - s_s * \mathbf{m}_m\|^2 + \|s_d - s_s * \mathbf{m}_m * \mathbf{m}_d\|^2. \end{aligned} \quad (9)$$

The first term of  $\mathcal{L}_A$  encourages A-NeRF predicting masks that represent spatial effects caused by distance, and the second term encourages A-NeRF generating masks that capture the difference between two channels.

## 4. Experiments

### 4.1. Experimental Settings

**Datasets.** To the best of our knowledge, our method is the first NeRF-based system capable of synthesizing real-world videos with perceptually spatial binaural audio at arbitrary poses. No publicly available dataset satisfies our experimental needs. So we validate our method on two representative real-world indoor scenes collected by ourselves. Since our model can generate binaural audio, we also train our AV-NeRF on FAIR-PLAY dataset [8] for objective comparisons.

**(1) Real-World Indoor Scenes.** We collect two representative indoor scenes in rooms of different sizes. The medium room is  $7 \times 7 m^2$  ( $23 \times 23 ft^2$ ) and the large room is  $40 \times 20 m^2$  ( $130 \times 64 ft^2$ ). As shown in Fig. 4, we mount a Zoom H3-VR Ambisonic microphone to a GoPro Max camera to assemble our recording system. The audio is recorded as binaural audio at 48kHz. The video is captured at 30 fps in GoPro’s linear Field of View mode to prevent image distortion. We move this system around in the environment

to record both auditory and visual signals at different positions and from different viewing directions. We use a loudspeaker playing music as a sound source. For each scene, we collect 2 minutes of data.

**(2) FAIR-PLAY** FAIR-PLAY dataset [8] is collected in a music room consisting of 1,871 10s video clips of people playing instruments. Each video is accompanied by binaural audio recorded by a professional binaural microphone. We re-organize this dataset by selecting video clips that belong to the same scene. Such video clips can capture people playing instruments at different camera poses, which provides our model with rich spatial information. Fig. 5 shows an example scene composed of several video clips capturing a person playing the harp. We collect four representative scenes from FAIR-PLAY for training and evaluation (harp, cello, drum, and guitar). Please note that we only use solo videos. We use “split1” provided by FAIR-PLAY authors to split training and test samples.

**Architectures.** We instantiate A-NeRF and V-NeRF as Multilayer Perceptron (MLP). A-NeRF has 6 fully-connected layers with 256 neurons per layer. V-NeRF has 8 fully-connected layers with 256 neurons per layer. Given a 3D environment voxel grid, we use a CNN to extract geometric information. CNN has 4 2D convolution layers with ReLU activations and 2D max pooling layers between two consecutive layers. Hypernetwork is a stack of 3-layer MLPs. For each fully-connected layer of A-NeRF, we use a 3-layer MLP to generate a low-rank matrix  $M$ . Additional details are provided in the Appendix.

**Training Details.** Given an audio-visual scene (e.g. several video clips), we first train V-NeRF to learn the visual world. We extract visual frames from input videos and use COLMAP structure-from motion library [21] to estimate the intrinsic and extrinsic of the camera. We set the initial learning rate as  $1e-3$  and optimize V-NeRF for 200K iterations. For easy optimization, we only use one NeRF model instead of a pair of coarse and fine models proposed by Mildenhall et al. [16]. After training V-NeRF, we freeze the parameters of V-NeRF and train the rest of our framework. We randomly sample 0.63s audio clips from videos and average poses of each clip. We set the batch size as 8, the learning rate as  $1e-3$  and train A-NeRF for 50K iterations. We apply binaural audio augmentation to real-world indoor scenes but do not use it on the FAIR-PLAY dataset because the head-related impulse response function open-sourced by Algazi *et al.* [1] is inconsistent with the head model of the recording system used by Gao and Grauman [8]. Additional details are provided in the Appendix.

### 4.2. Results on Real-World Audio-Visual Scenes

We validate our method qualitatively on real-world audio-visual scenes. We compose A-NeRF and V-NeRF as our baseline, which does not contain AV-Bridge (Sec. 3.3),

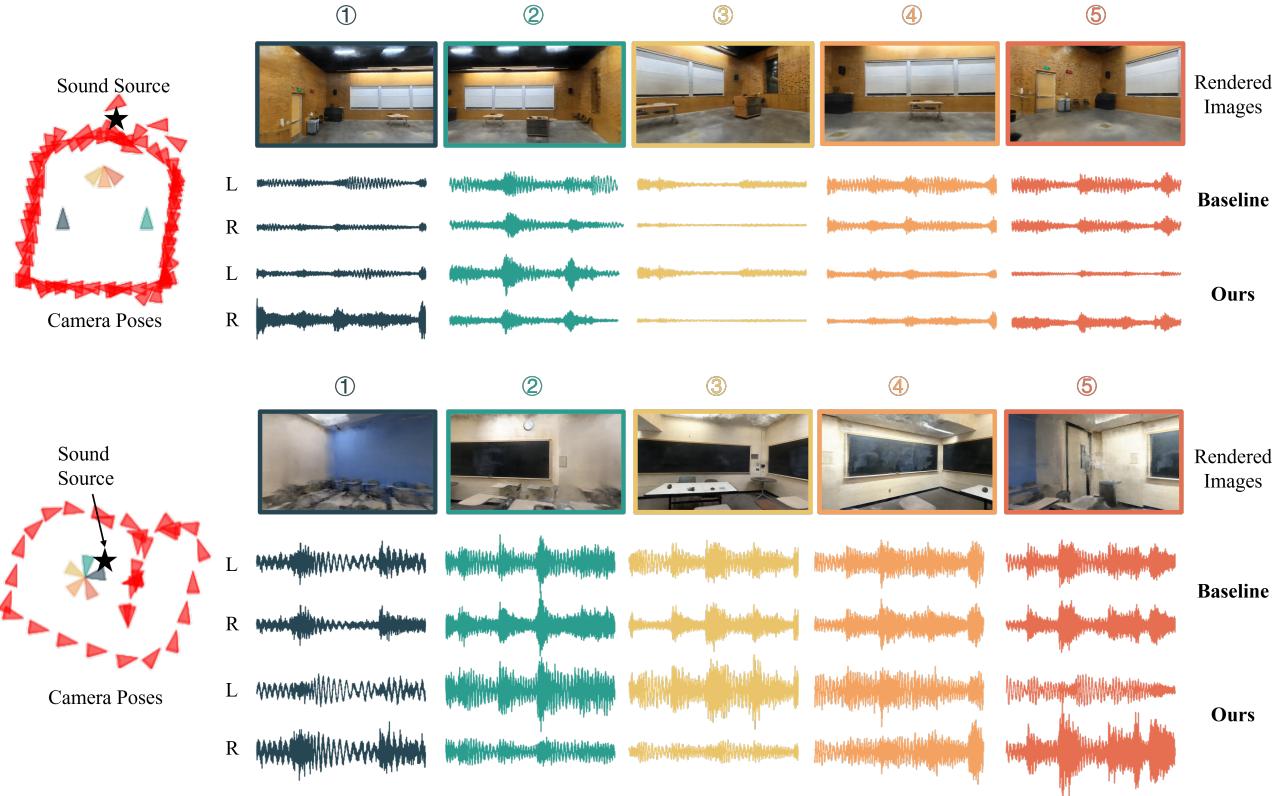


Figure 6. Results on Real-World Audio-Visual Scenes. We synthesize audio-visual scenes at novel camera poses that have no spatial overlap with the training camera poses. We visualize both the rendered visual frames and binaural audio. The first column of both figures is the camera poses, including training poses (colored red) and novel poses (colored otherwise). We mark the sound source as a black pentagram. Starting from the second column, we show rendered images and rendered binaural audio. The color of rendered results corresponds to that of the camera pose.

Methods	Harp		Cello		Drum		Guitar		Overall	
	STFT	ENV								
M2B [8] w/ GT image	0.551	0.119	0.630	0.131	0.103	0.048	0.839	0.142	0.531	0.110
P2B [28] w/ GT image	0.498	0.118	0.819	0.140	0.093	0.045	0.818	0.138	0.557	0.110
Mono-Mono	0.939	0.149	0.702	0.133	0.155	0.063	0.966	0.150	0.691	0.124
Right-Left	3.728	0.241	1.915	0.187	0.375	0.073	3.068	0.218	2.272	0.180
M2B [8] w/ retrieved image	0.872	0.168	<u>0.644</u>	<u>0.131</u>	0.145	<u>0.058</u>	0.873	0.143	<u>0.634</u>	0.125
P2B [28] w/ retrieved image	<u>0.661</u>	0.139	0.920	<u>0.131</u>	<u>0.135</u>	<u>0.058</u>	<b>0.822</b>	<b>0.138</b>	0.635	<u>0.117</u>
Ours	<b>0.638</b>	<b>0.121</b>	<b>0.635</b>	<b>0.124</b>	<b>0.123</b>	<b>0.053</b>	0.865	0.141	<b>0.565</b>	<b>0.110</b>

Table 1. We compare our AV-NeRF with other methods of binaural audio generation. M2B is an abbreviation of MONO2BINAURAL [8] and P2B is an abbreviation of PSEUDO2BINAURAL [28]. We feed M2B and P2B models with two kinds of images: ground-truth images, and retrieved images with the nearest camera pose. We highlight the top two results in each column.

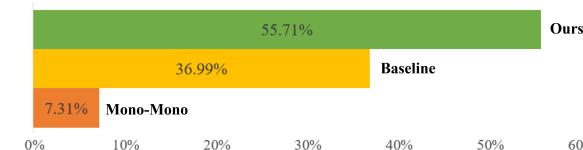


Figure 7. User study of our method. We ask participants to select videos with the best consistency between rendered audio and video. We compare our method with Mono-Mono and baseline. This figure exhibits the distinct advantage of our method.

coordinate transformation (Sec. 3.4), or binaural audio augmentation (Sec. 3.5). The top half of Fig. 6 shows the rendering results of the large room. The first column shows the camera poses used for training and novel view synthesis from a bird’s eye view. The direction of a triangle’s sharp angle is the camera’s orientation, and the position of the triangle’s sharp angle is the camera’s position. Red triangles represent camera poses of training samples. Triangles of other colors are camera poses used for novel audio-visual scene synthesis. There is no spatial overlap between train-

ing poses and novel poses. The position of the sound source is marked with a black pentagram. We show audio and visual rendering results at different camera poses starting from the second column. The first row is the rendered image. The second and third rows are binaural audio rendered by our baseline method. The fourth and fifth rows show binaural audio rendered by our AV-NeRF. Our AV-NeRF can render binaural audio with rich spatial information compared to the baseline method. When the sound source is on the right side of the camera view (camera pose ① and ⑤), AV-NeRF can generate two channels with a distinct difference – the amplitude of the right channel is clearly larger than that of the left channel. When the sound source is on the left side of the camera view (camera pose ② and ③), the amplitude of the left channel is clearly larger than that of the right channel. AV-NeRF can also generate balanced binaural audio when the camera faces the sound source directly.

The bottom half of Fig. 6 shows the rendering results of the medium room. Similarly, we show the training camera poses, novel camera poses, and sound source on the left side. We rotate the camera to generate novel 360-degree audio-visual scenes. Please note that the camera position is fixed when we change the camera orientation. As shown in the figure, our AV-NeRF can render binaural audio consistent with the camera orientations. When the camera is rotated gradually from pose ① to ⑤, the amplitude ratio (energy ratio) of the left channel to the right channel will first increase (from pose ① to pose ③) and then decrease (from pose ③ to pose ⑤).

**Perceptual User Studies.** We also conduct user studies to validate the fidelity of our generated audio-visual scenes. For each room, we generate 4 audio-visual scenes with different camera motion trajectories. We compare scenes generated by AV-NeRF with those generated by the baseline. We also include a Mono-Mono method that directly copies an input mono audio twice to generate fake binaural audio. We totally recruited 22 participants with normal hearing. We ask participants to watch videos generated by three methods and select the video with the spatial consistency between rendered audio and video. Considering different methods may achieve comparable performance, we allow participants to select multiple options. As shown in Fig. 7, our method receives the most preference and exhibits a distinct advantage over other methods.

#### 4.3. Results on FAIR-PLAY

For numerical comparisons, we evaluate our AV-NeRF on FAIR-PLAY dataset. We use two metrics to evaluate the audio generation quality of a model: STFT distance and envelope (ENV) distance. For the definitions of these two metrics, please refer to [8].

Table 1 shows the quantitative results of our AV-NeRF on four scenes. We compare our model with other methods

Methods	Harp		Cello		Drum		Guitar	
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Baseline	0.702	0.135	0.776	0.130	0.154	0.063	0.877	0.138
w/ AVB	0.635	0.129	0.642	0.126	0.149	0.060	0.877	0.141
w/ AVB & CT	0.638	0.121	0.635	0.124	0.123	0.053	0.865	0.141

Table 2. Ablation study on different components of our method. We report results on four FAIR-PLAY scenes. **AVB:** AV-Bridge, **CT:** Coordinate Transformation.

of binaural audio generation. We select the original method MONO2BINAURAL (M2B) proposed in [8] and a recently proposed method PSEUDO2BINAURAL (P2B) [28] for comparison. We also use Mono-Mono and Right-Left as two baseline methods. Mono-Mono copies the input mixed audio twice to generate fake binaural audio and Right-Left switches two channels to generate binaural audio. We use Right-Left as the upper bound of STFT and ENV distance. There is no surprise that our model does not surpass M2B and P2B because these two models are trained on the entire dataset with nearly 1,500 training samples and the visual encoder is pre-trained on ImageNet [5] while our model is trained on less than 10 samples per scene from scratch. Even with much less training data, our model achieves competitive results to M2B and P2B. M2B and P2B models rely on ground-truth images for audio spatialization. However, our model does not have access to ground-truth images. For a fair comparison, we adapt M2B and P2B to generate audio at novel poses. Instead of feeding M2B and P2B with ground-truth images, we input them with retrieved images with the nearest camera poses. After the absence of ground-truth images, their models have a clear performance degradation and fall behind our AV-NeRF. AV-NeRF outperforms M2B and P2B with distinct advantages.

**Ablation Studies.** We conduct ablation studies on FAIR-PLAY to further validate the proposed pipeline. We decompose AV-NeRF into the baseline model, AV-Bridge, and coordinate transformation module. As shown in Fig. 2, adding AV-Bridge and coordinate transformation module to AV-NeRF can boost our model performance.

## 5. Discussion

In this work, we propose a first-of-its-kind NeRF system that is capable of synthesizing real-world audio-visual scenes accompanied by binaural audio. Our model can generate audio with rich spatial information at novel camera poses. We demonstrate the effectiveness of our method on FAIR-PLAY and real-world indoor scenes.

There is a range of promising future directions. First, our method learns and models the spatial audio effects caused by distance and orientation. The reverberation that exists in environments is not modeled. It would be interesting to future extend our work to address sound propagation. Second, we focus on static scenes with a fixed sound source in our study. However, there could be multiple sounding objects in a single environment, and these sound sources

may move. How to learn implicit neural representations for audio-visual scenes with multiple dynamic sound sources is still an open question. Third, there are power audio-visual simulation platforms (e.g., SoundSpaces 2.0 [4]) that can render realistic audio and visual scenes in virtual environments. An interesting direction is to explore how to integrate the AV-NeRF with virtual simulation to strengthen learning robustness and generalization ability.

## References

- [1] V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano. The cipic hrf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pages 99–102, 2001. [3](#), [5](#), [6](#)
- [2] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, pages 15516–15525, 2021. [1](#)
- [3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspace: Audio-visual navigation in 3d environments. In *ECCV*, 2020. [1](#), [2](#), [3](#)
- [4] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspace 2.0: A simulation platform for visual-acoustic learning. *arXiv*, 2022. [2](#), [3](#), [9](#)
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. [8](#)
- [6] Yilun Du, M. Katherine Collins, B. Joshua Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. In *NeurIPS*, 2021. [1](#), [2](#), [3](#), [5](#)
- [7] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. [2](#)
- [8] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. [1](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [9] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, pages 5784–5794, 2021. [1](#)
- [10] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *ICLR*, 2017. [5](#)
- [11] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. [2](#), [3](#)
- [12] Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng. Scene-aware audio for 360° videos. *ACM TOG*, 37(4), 2018. [3](#)
- [13] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard A. Newcombe, and ZhaoYang Lv. Neural 3d video synthesis from multi-view video. In *CVPR*, pages 5511–5521, 2022. [2](#)
- [14] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. [2](#)
- [15] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *NeurIPS*, 2022. [1](#), [2](#)
- [16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. [2](#), [3](#), [5](#), [6](#)
- [17] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, volume 13697, pages 218–234, 2022. [1](#)
- [18] Pedro Morgado, Nuno Vasconcelos, Timothy R. Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. In *NeurIPS*, pages 360–370, 2018. [3](#)
- [19] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, pages 3504–3515, 2020. [2](#)
- [20] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. [1](#)
- [21] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. [6](#)
- [22] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 32, 2019. [2](#)
- [23] Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhosiny. Adversarial generation of continuous images. In *CVPR*, pages 10753–10764, 2021. [5](#)
- [24] Luchuan Song, Bin Liu, Guojun Yin, Xiaoyi Dong, Yufei Zhang, and Jia-Xuan Bai. Tacr-net: editing on deep video and voice portraits. In *ACM MM*, pages 478–486, 2021. [1](#)
- [25] Luchuan Song, Guojun Yin, Bin Liu, Yuhui Zhang, and Nenghai Yu. Fsft-net: face transfer video generation with few-shot views. In *ICIP*, pages 3582–3586. IEEE, 2021. [1](#)
- [26] Zhenyu Tang, Nicholas J Bryan, Dingzeyu Li, Timothy R Langlois, and Dinesh Manocha. Scene-aware audio rendering via deep acoustic analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2020. [3](#)
- [27] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. [1](#)
- [28] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *CVPR*, 2021. [3](#), [5](#), [7](#), [8](#)
- [29] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *NeurIPS*, 33:2492–2502, 2020. [2](#)

- [30] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, pages 570–586, 2018. <sup>1</sup>
- [31] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision*, pages 52–69. Springer, 2020. <sup>1</sup>
- [32] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020. <sup>3</sup>