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

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

Abstract

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 demonstrate the advantage of our model in real-world audio-visual scenes.

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. Recognizing this, the machine perception research community has seen a spectrum of works [2, 3, 5, 6, 11–13, 15, 16] 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.

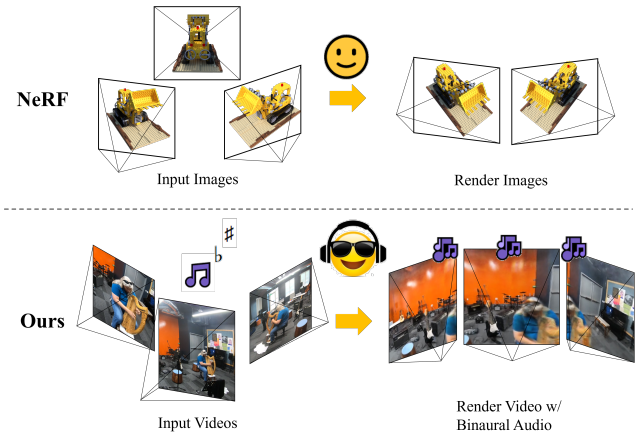


Figure 1. NeRF learns to render visual scenes at novel poses. Beyond visual rendering, we present AV-NeRF for learning to synthesize audio-visual consistent scenes including video frames at a novel view and the corresponding binaural audio in the environment. Consistent sight and sound can provide users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

Although there are some similar works [4, 9], these methods have some constraints that limit their usage in solving our problem. Luo *et al.* [9] 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. Du *et al.* [4] propose a manifold learning method that maps latent vectors to (image, audio) pairs. However, 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. Briefly, we (1) introduce a novel acoustic-aware audio generation method to encode our prior knowledge of sound propagation; (2) propose a coordinate transformation mechanism for effective direction expression; (3) introduce a binaural audio augmentation method.

2. 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.

2.1. V-NeRF

NeRF [10] 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:

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

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 σ 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 σ and a feature vector; then NeRF maps the feature vector and 2D direction (θ, ϕ) to a color $\mathbf{c}$. This process is illustrated in Fig. 3a.

NeRF then uses the volume rendering method [8] 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:

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

where $T(t) = \exp(-\int_{t_n}^{t}(\mathbf{r}(s))ds)$ and $\mathbf{d} = (\theta, \phi)$.

2.2. A-NeRF

The target of A-NeRF is to learn a neural acoustic representation that can map 5D coordinates (x, y, z, θ, ϕ) 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 (θ, ϕ), and F is the number of frequency bins:

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

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 (θ, ϕ) and pass it to the rest part of A-NeRF to generate a difference mask $\mathbf{m}_d$. Given a sound of interest, we can use $\mathbf{m}_m$ and $\mathbf{m}_d$ to synthesize new binaural audio.

2.3. AV-Bridge

Given the fact that 3D geometry partially determines the sound propagation in an environment, we propose an acoustic geometry-aware audio generation method. 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 [7] to utilize this geometric information for acoustic-aware audio generation. We design a Hypernetwork ψ to convert the environment vector v into parameters $\mathbf{W}_A$ of A-NeRF inspired by [4]:

$$\psi : v \rightarrow \mathbf{W}_A \ . \tag{4}$$

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:

$$\mathbf{W}_i \leftarrow \mathbf{W}_i \odot M \ , \tag{5}$$

where $\odot$ is Hadamard product.

2.4. Coordinate Transformation

Because the human perception of the sound direction is based on the relative orientation to the sound source instead of the absolute direction, we propose expressing viewing direction (θ, ϕ) 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 $\angle(\mathbf{V}_1, \mathbf{V}_2)$.

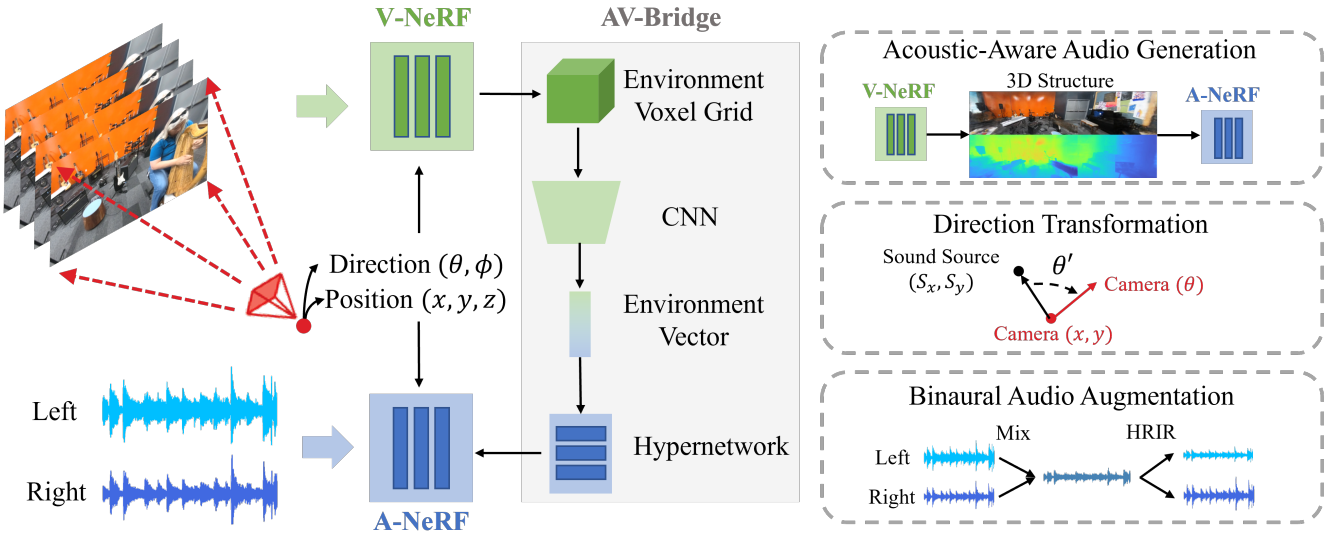
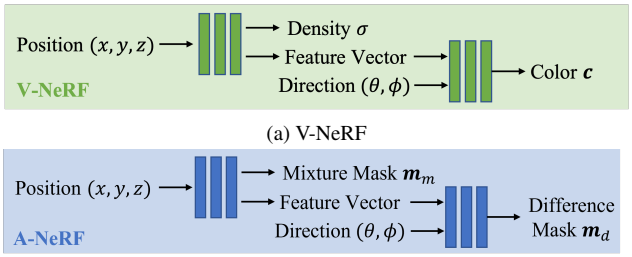


Figure 2. The pipeline of our method. Given the position (x, y, z) and viewing direction (θ, ϕ) of a listener, our method can render an image the listener would see and the corresponding binaural audio the listener would hear. Our model consists of V-NeRF, A-NeRF, and AV-Bridge. V-NeRF learns to generate 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.



(a) V-NeRF

(b) A-NeRF

Figure 3. Architecture of V-NeRF and A-NeRF.

Figure 4. Recording devices and two representative indoor scenes.

2.5. Binaural Audio Augmentation

To provide our model with more high-quality binaural audio for training, we apply head-related impulse response (HRIR) to the stereo audio to generate binaural audio following Xu *et al.* [14]. We exploit an open-sourced HRIR database [1] for binaural audio augmentation.

2.6. Learning Objective

We refer to the combination of V-NeRF (Sec. 2.1) and A-NeRF (Sec. 2.2) as the baseline method. We integrate AV-Bridge (Sec. 2.3), coordinate transformation module (Sec.2.4), and data augmentation mechanism (Sec. 2.5) 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 [10]:

$$\mathcal{L}_V = ||C(\mathbf{r}) - \hat{C}(\mathbf{r})||^2 \ , \tag{6}$$

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 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 predicted spectrums:

$$\begin{aligned} \mathcal{L}_A &= ||s_m - \hat{s}_m||^2 + ||s_d - \hat{s}_d||^2 \\ &= ||s_m - s_s * \mathbf{m}_m||^2 + ||s_d - s_s * \mathbf{m}_m * \mathbf{m}_d||^2. \end{aligned} \tag{7}$$

3. Experiments

3.1. Experimental Settings

We collect two representative indoor scenes in a medium and large room. The recording device and the indoor scenes are shown in Fig. 4.

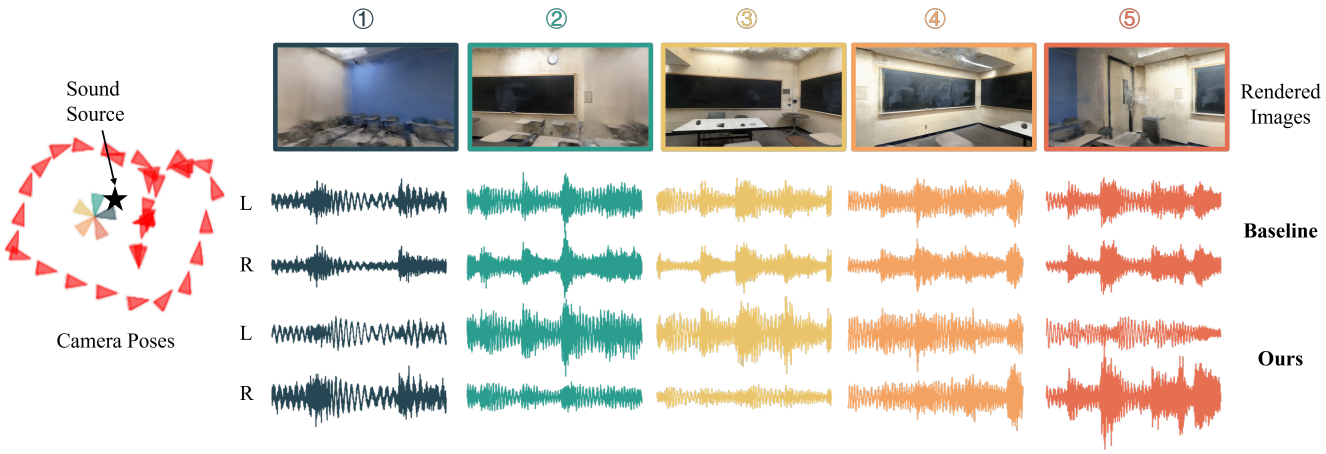


Figure 5. Results in Real-World Audio-Visual Scenes. We synthesize audio-visual scenes at novel camera poses that have no spatial overlap with the training camera poses. We visualize both the rendered visual frames and binaural audio. The first column of both figures is the camera poses, including training 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.

3.2. Experimental Results

We show the rendering results of the medium room in Fig. 5. We rotate the camera to generate novel 360-degree audio-visual scenes with the camera position fixed. As shown in the figure, our AV-NeRF can render binaural audio consistent with the camera orientations.

4. Conclusion

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. We demonstrate the effectiveness of our method in real-world indoor scenes.

References

[1] V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano. The cipic hrtf 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

[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. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 1

[4] Yilun Du, M. Katherine Collins, B. Joshua Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. In *NeurIPS*, 2021. 1, 2

[5] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 1

[6] 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

[7] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *ICLR*, 2017. 2

[8] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *SIGGRAPH*, 18(3):165–174, 1984. 2

[9] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *NeurIPS*, 2022. 1

[10] 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

[11] 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

[12] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, pages 631–648, 2018. 1

[13] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 1

[14] 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

[15] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, pages 570–586, 2018. 1

[16] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, pages 52–69. Springer, 2020. 1