SOAF: Scene Occlusion-aware Neural Acoustic Field

Supplementary Material

1. Evaluation Metrics

RWAVS Dataset. Following [5], we evaluate our method using metrics, such as the magnitude distance (MAG) [11] and envelope distance (ENV) [8] on the RWAVS dataset.

MAG measures audio quality in the time-frequency domain and is defined as

$$\mathrm{MAG}(\mathbf{m}_{\mathrm{prd}}, \mathbf{m}_{\mathrm{gt}}) = ||\mathbf{m}_{\mathrm{prd}} - \mathbf{m}_{\mathrm{gt}}||^2, \tag{1}$$

where $\mathbf{m}_{\mathrm{prd}}$ and $\mathbf{m}_{\mathrm{gt}}$ are the predicted and ground truth magnitude, respectively.

ENV describes audio quality in the time domain and is defined as

$$\mathrm{ENV}(a_{\mathrm{prd}}, a_{\mathrm{gt}}) = ||\mathrm{hilbert}(a_{\mathrm{prd}}) - \mathrm{hilbert}(a_{\mathrm{gt}})||^2, \tag{2}$$

where a_{prd} and a_{gt} are the predicted and ground truth audio, and $\mathrm{hilbert}$ denotes Hilbert transformation function [9].

SoundSpaces Dataset. Following [5, 10], we utilize the reverberation time (T60), acoustic parameter clarity (C50), and early decay time (EDT) as evaluation metrics on the SoundSpaces dataset.

T60 is a crucial acoustic parameter that characterizes the reverberation effect of a room, defined as the time needed for a sound to decay by 60 decibels (dB). We calculate the percentage error of T60 by

$$\mathrm{T60}(a_{\mathrm{prd}}, a_{\mathrm{gt}}) = \frac{|\mathrm{T60}(a_{\mathrm{prd}}) - \mathrm{T60}(a_{\mathrm{gt}})|}{\mathrm{T60}(a_{\mathrm{gt}})}, \tag{3}$$

where a_{prd} and a_{gt} are the predicted and ground truth impulse response.

C50 is a measurement of audio clarity that quantifies the energy ratio between early reflections and late reverberation. The C50 distance is formatted as

$$\mathrm{C50}(a_{\mathrm{prd}}, a_{\mathrm{gt}}) = |\mathrm{C50}(a_{\mathrm{prd}}) - \mathrm{C50}(a_{\mathrm{gt}})|. \tag{4}$$

EDT reflects people's perception of reverberation by focusing on the early reflections of impulse responses. The EDT distance is formatted as

$$\mathrm{EDT}(a_{\mathrm{prd}}, a_{\mathrm{gt}}) = |\mathrm{EDT}(a_{\mathrm{prd}}) - \mathrm{EDT}(a_{\mathrm{gt}})|. \tag{5}$$

2. Additional Experiments

Contribution of Occlusion Number n. We split the receivers in a multi-room scene according to their number of occlusions n to the sound source, and evaluate them separately. As shown in Table 1, our full model achieves improved performance by integrating the occlusion number n to the sound source in the scene.

	Region $n = 1$		Region $n = 2$	
Methods	MAG↓	ENV↓	MAG↓	ENV↓
Ours - w/o n	1.912	0.144	0.390	0.076
Ours - *full*	**1.896**	**0.143**	**0.375**	**0.075**

Table 1. Ablation study of the occlusion number n.

Methods	MAG	LRE	ENV	RTE	DPAM
DSP [4]	1.016	3.468	0.274	0.119	0.588
VAM [1]	0.390	0.996	0.156	0.079	0.459
ViGAS [2]	0.370	1.089	0.147	0.094	0.357
NACF [6]	0.459	1.364	0.176	0.138	0.506
INRAS [10]	0.455	1.503	0.179	0.148	0.485
NAF [7]	0.448	1.204	0.522	0.138	0.353
AV-NeRF [5]	0.370	1.013	0.145	0.098	0.381
AV-Cloud [3]	0.351	0.936	0.145	0.074	0.276
AV-Cloud + Our Priors	**0.340**	**0.921**	**0.144**	**0.063**	**0.259**

Table 2. Results on RWAVS in AV-Cloud's evaluation protocol.

Methods	T60 (%) ↓	C50 (dB) ↓	EDT (sec) ↓
NACF [6]	2.33	0.55	0.0162
NACF w/o Acoustic Context	2.96	0.75	0.0183
NACF w/o **Energy Decay Loss**	5.15	0.98	0.0255

Table 3. Ablation of energy decay loss in NACF on Soundsapces.

Representation agnosticism. We integrate our priors into the official implementation of AV-Cloud [3], and present results in Table 2. The improved performance demonstrates that our contributions are orthogonal to previous works and agnostic to the underlying scene representation.

3. Room Impulse Response Prediction

On the SoundSpaces dataset, following [5, 10], we adapt our framework to predict room impulse responses in the time domain instead of estimating acoustic masks. Specifically, we discard the mixture mask $\mathbf{m}_m$ and replace the output layers of MLPs to synthesize RIR signals $\mathbf{rir}_l$, $\mathbf{rir}_r$, rather than $\mathbf{m}_d^l$, $\mathbf{m}_d^r$, while leaving other components unchanged. To align with the baselines, we do not apply any additional optimization strategies, despite their potential to further improve performance (*e.g.*, energy decay loss; see Table 3). For fairness, the network is supervised only by the L_2 distance between the predicted and ground-truth magnitudes computed via STFT with FFT $= 512$, hop length $= 128$, window length $= 512$, and a Hamming window. Given the impulse responses for the left and right channels, $\mathbf{rir}_l$ and $\mathbf{rir}_r$, we can generate the binaural audio a_t^* based on the source audio a_s^* with temporal convolution:

$$a_t^* = [a_s^* \otimes \mathbf{rir}_l,\ \ a_s^* \otimes \mathbf{rir}_r], \tag{6}$$

where $\otimes$ denotes the convolution operation.

4. Additional Visualizations

We provide additional qualitative comparisons of the synthesized audio within different scenes in Figure 1 and Figure 2. For each scene, we present several example images of the input video, alongside the scene geometry that we reconstructed from the video frames. As shown in these figures, compared with AV-NeRF [5], our method demonstrates a superior ability to generate binaural audio with enhanced scene distance-aware (Figure 1c), occlusion-aware (Figure 1d, 2c) and direction-aware (Figure 2d) effects from novel views in various scenarios.

References

[1] 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. 1

[2] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6409–6419, 2023. 1

[3] Mingfei Chen and Eli Shlizerman. Av-cloud: Spatial audio rendering through audio-visual cloud splatting. *Advances in Neural Information Processing Systems*, 37:141021–141044, 2024. 1

[4] Corey I Cheng and Gregory H Wakefield. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *journal of the Audio Engineering Society*, 49(4):231–249, 2001. 1

[5] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 2

[6] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arXiv preprint arXiv:2309.15977*, 2023. 1

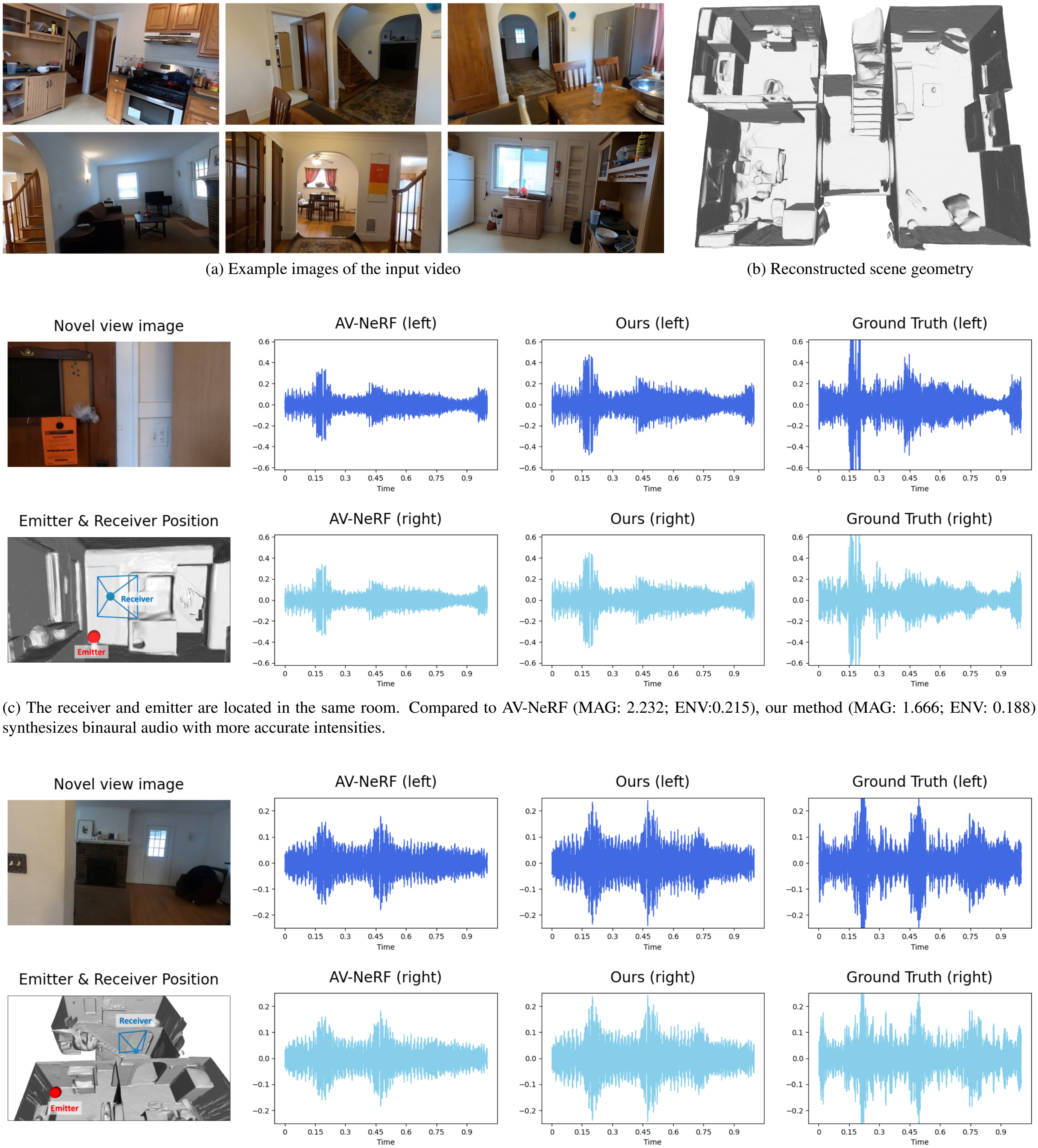
[7] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022. 1

[8] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018. 1

[9] Julius O Smith. *Mathematics of the discrete Fourier transform (DFT): with audio applications*. Julius Smith, 2008. 1

[10] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022. 1

[11] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 1



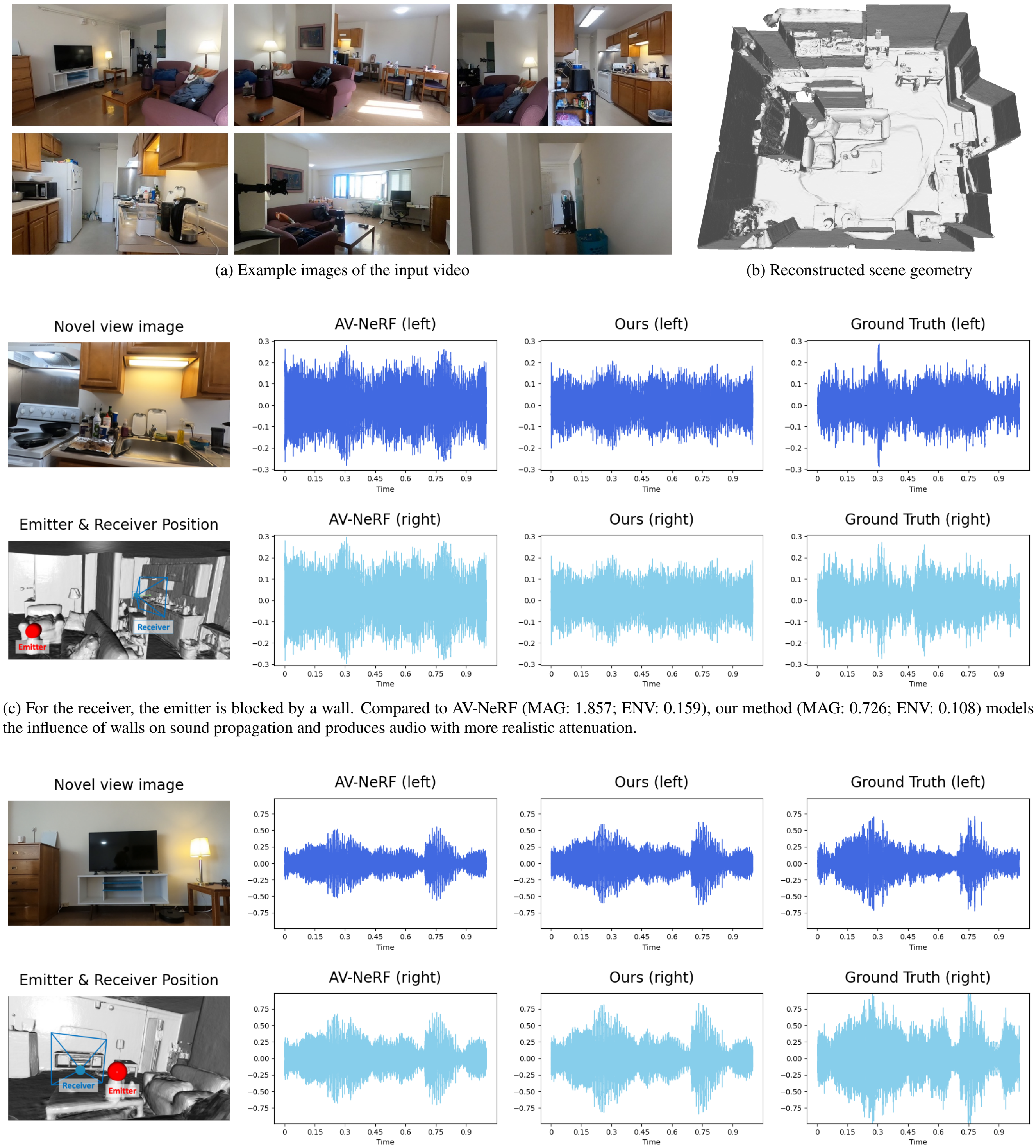
(a) Example images of the input video

(b) Reconstructed scene geometry

(c) The receiver and emitter are located in the same room. Compared to AV-NeRF (MAG: 2.232; ENV:0.215), our method (MAG: 1.666; ENV: 0.188) synthesizes binaural audio with more accurate intensities.

(d) The receiver and emitter are located in different rooms. Compared to AV-NeRF (MAG: 0.573; ENV: 0.102), our method (MAG: 0.537; ENV: 0.098) generates audio with enhanced occlusion-aware attenuation.

Figure 1. Novel-view audio synthesis within a *house* from the RWAVS dataset. (a) Example video images. (b) Scene geometry extracted from the input video. (c, d) Audio synthesis from two different receiver poses. Our approach generates superior binaural audio with (c) distance-aware and (d) occlusion-aware effects than AV-NeRF.



(a) Example images of the input video

(b) Reconstructed scene geometry

(c) For the receiver, the emitter is blocked by a wall. Compared to AV-NeRF (MAG: 1.857; ENV: 0.159), our method (MAG: 0.726; ENV: 0.108) models the influence of walls on sound propagation and produces audio with more realistic attenuation.

(d) The receiver is to the left of the emitter. Compared to AV-NeRF (MAG: 4.636; ENV: 0.339), our method (MAG: 4.412; ENV: 0.334) improves spatial effects by synthesizing distinct audio in the left and right channels.

Figure 2. Novel-view audio synthesis within an *apartment* from the RWAVS dataset. (a) Example video images. (b) Scene geometry extracted from the input video. (c, d) Audio synthesis from two different receiver poses. Our approach achieves better audio generation performance with (c) occlusion-aware and (d) direction-aware effects than AV-NeRF.