AV-RIR: Audio-Visual Room Impulse Response Estimation

Anton Ratnarajah Sreyan Ghosh Sonal Kumar Purva Chiniya Dinesh Manocha
University of Maryland, College Park
{jeran,sreyang,sonalkum,pchiniya,dmanocha}@umd.edu

Abstract

Accurate estimation of Room Impulse Response (RIR), which captures an environment's acoustic properties, is important for speech processing and AR/VR applications. We propose AV-RIR, a novel multi-modal multi-task learning approach to accurately estimate the RIR from a given reverberant speech signal and the visual cues of its corresponding environment. AV-RIR builds on a novel neural codec-based architecture that effectively captures environment geometry and materials properties and solves speech dereverberation as an auxiliary task by using multi-task learning. We also propose Geo-Mat features that augment material information into visual cues and CRIP that improves late reverberation components in the estimated RIR via image-to-RIR retrieval by 86%. Empirical results show that AV-RIR quantitatively outperforms previous audio-only and visual-only approaches by achieving 36% - 63% improvement across various acoustic metrics in RIR estimation. Additionally, it also achieves higher preference scores in human evaluation. As an auxiliary benefit, dereverbed speech from AV-RIR shows competitive performance with the state-of-the-art in various spoken language processing tasks and outperforms reverberation time error score in the real-world AVSpeech dataset. Qualitative examples of both synthesized reverberant speech and enhanced speech are available online[1].

1. Introduction

Reverberation, caused by sound reflecting off surrounding surfaces, transforms how a listener perceives the sound once it is released from a sound source. The transformation is influenced by specific properties of the surrounding area, like spatial geometry, the composition and material properties of surfaces and objects within the environment, and the positioning of various sound sources in proximity. For example, someone speaking or playing music in a large auditorium is perceptually significantly different from someone speaking in a small classroom [6, 73]. The environmental effect that any sound goes through because of the transformation can be quantitatively described by the room impulse response (RIR). RIR is a fundamental concept that characterizes how an acoustic space affects sound, essentially representing the transfer function between a sound source and a receiver, encapsulating all the direct and reflective paths that sound can travel within any indoor or outdoor environment.

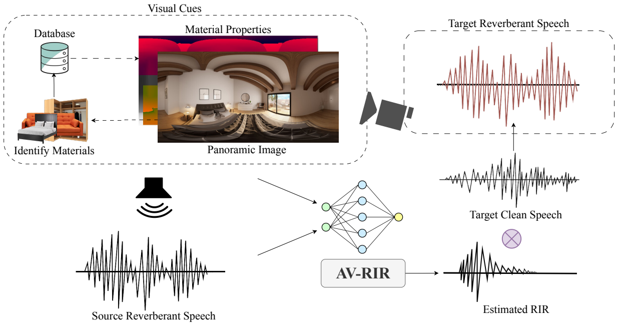

Figure 1. Overview of **AV-RIR**: Given a source reverberant speech in any environment, AV-RIR estimates the RIR from the reverberant speech using additional visual cues. The estimated RIR can be used to transform any target clean speech as if it is spoken in that environment.

RIR estimation, defined as estimating the RIR component from a given reveberant speech signal (see Eq. 1), finds its major application in augmented reality (AR) and virtual reality (VR) [2–4, 50]. Usually, when sound effects do not align acoustically with the visual scene, it can disrupt the audio-visual human perception. In AR and VR settings, discrepancies between the acoustics of the real environment and the virtually simulated space lead to cognitive dissonance. This phenomenon, known as the "room divergence effect", can significantly detract from the user experience [63, 84]. RIR estimation from real-world speech can help overcome these problems.

Prior work on RIR estimation mainly deals with recorded audio signals and does not take into account visual cues. Directly estimating the RIR from source reverberant speech has been extensively studied using traditional signal processing methods [18, 25, 35, 48]. However, these ap-

[1] https://anton-jeran.github.io/AVRIR/

proaches may not work well in some real-world applications, mainly because they are based on the assumption that the source is a modulated Gaussian pulse, not actual speech, [25, 35] or they require pre-knowledge of the specific attributes about the speaker or the microphone used for recording [18, 48]. Recently, neural learning-based RIR estimation techniques have been proposed to estimate RIR from reverberant speech [69, 81]. These techniques are capable of estimating early components (i.e., the direct response and early reflections of RIR) and are not very effective in estimating late components (i.e., the late reverberation of RIR) because the early components of the RIR have impulsive sparse components, while the late components have a noise-like structure with significantly lower magnitude compared to early components. Therefore audio-only approach approximates the late components using a sum of decaying filtered noise signal [1, 81].

Main Contributions. We propose AV-RIR, a novel multi-modal multi-task learning approach for RIR estimation. AV-RIR employs a novel neural codec-based multi-modal architecture that takes as input a reverberant speech uttered in a source environment, the panoramic image of the environment, and a novel Geo-Mat feature that incorporates information about room geometry and the materials of surfaces and objects. The multi-modal architecture consists of carefully designed encoders, decoders, and a Residual Vector Quantizer that learns rich task-specific (RIR estimation and speech deverberation) features while discarding the noise in training data [91]. Additionally, AV-RIR incorporates a dual-branch structure for multi-task learning, where we solve an auxiliary speech dereverberation task alongside the primary RIR estimation task. This approach effectively redefines the ultimate learning objective as decomposing reverberated speech into its constituent anechoic speech and RIR components. Furthermore, we propose Contrastive RIR-Image Pre-training (CRIP) to improve late reverberation of the estimated RIR during inference time using image-to-RIR retrieval. To summarize, our main contributions are as follows:

1. We propose AV-RIR, a novel multi-modal multitask learning approach for RIR estimation.
2. AV-RIR employs a neural codec-based multi-modal architecture that takes as input audio, visual cues and a novel Geo-Mat feature. We also propose CRIP to improve late reverberation effects using retrieval.
3. During training, AV-RIR solves an auxiliary speech dereverberation task for learning RIR estimation. Through this, AV-RIR essentially learns to separate anechoic speech and RIR.
4. We perform extensive experiments to prove the effectiveness of AV-RIR. AV-RIR outperforms prior works by significant margins both quantitatively and qualitatively. We achieve 36% - 63% on RIR estimation on the SoundSpaces dataset [10], and 56% - 79% people find that AV-RIR is closer to the ground-truth in the visual acoustic matching task over our baselines. Additionally, the dereverbed speech predicted by AV-RIR improves performance across various spoken language processing (SLP) tasks. We also perform extensive ablation experiments to demonstrate the critical role of each modules within the AV-RIR framework.

2. Background and Related Work

Room Impulse Response. The reverberation effects in a recorded speech can be characterized by a transfer function known as room impulse response (RIR) [66]. We can breakdown the speech content ($\mathcal{S}_C$) and the RIR ($\mathcal{RIR}$) corresponding to the given reverberant speech ($\mathcal{S}_R$) using a convolution operation ($\circledast$) as follows:

$$\mathcal{S}_{\mathrm{R}} = \mathcal{S}_{\mathrm{C}} \circledast \mathcal{RIR}, \tag{1}$$

where RIR represents the intensity and time of arrival of direct sound, early reflections, and late reverberation. The RIR can either be measured in a controlled environment [5, 16, 24, 77] or simulated using physics-based simulators [86, 92]. Measuring RIR requires sophisticated hardware and human labor. At the same time, acoustic simulators require a 3D mesh representation of the environment [85] and complete knowledge of room materials [76]. Thus, these simulators cannot capture all the acoustic effects of an RIR at an interactive rate [85]. Prior works also show that accurately generating RIRs without these cues is infeasible (e.g., only from RGB images [53, 79]). Alternatively, estimating RIR from reverberant speech has seen better success [81]. Reverberant speech can be easily recorded using household devices (e.g., mobile phone, Amazon Echo, etc.). Prior work explores audio-only techniques [44, 69, 81], and to the best of our knowledge, AV-RIR is the first audio-visual method to solve this task.

Room Acoustic Estimation and Matching. Interactive applications (i.e., AR / VR, computer games, etc.) demand accurate RIRs to generate realistic sound effects. Many physics-based solutions have been proposed to generate RIRs for synthetic scenes [45, 55, 64, 86, 92]. Alternatively, several machine-learning algorithms are being proposed to estimate RIRs for the given environment [15, 47, 52, 53, 65, 67, 68]. Pioneering learning-based work on generating RIR from a single RGB image of physical environments includes Image2Reverb [79], which is based on a conditional GAN-based architecture. Generating RIR from a single RGB image might not be the most effective as it does not have enough information, like 3D geometry, information about the material properties of objects in the environment, speaker position, etc. The existing physics-based and machine-learning solutions to generate accurate RIRs

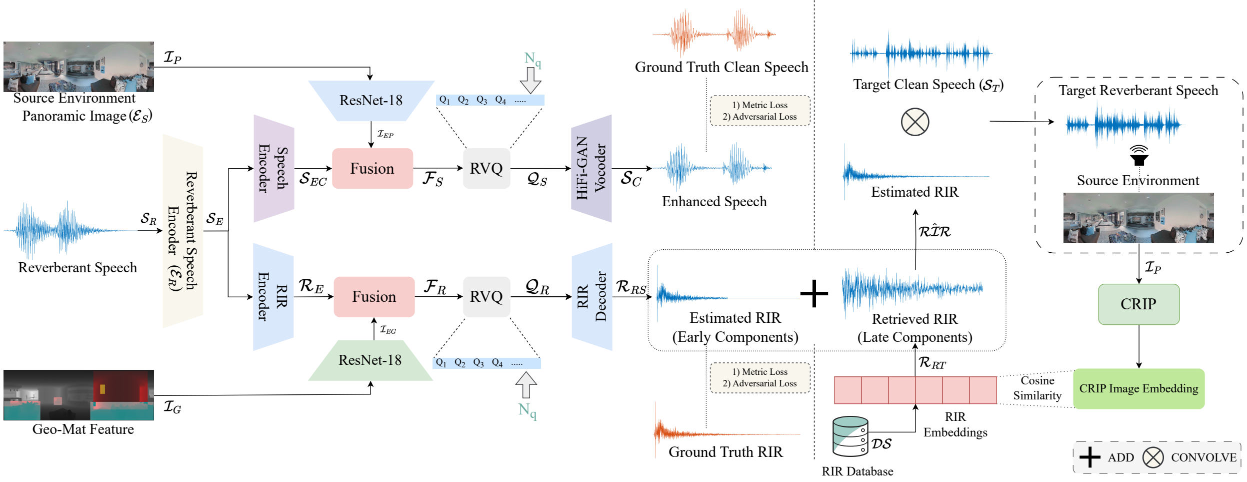

Figure 2. Overview of our AV-RIR learning method: Given the input reverberant speech $\mathcal{S}_R$ from any source environment $\mathcal{E}_S$, the primary task of AV-RIR is to estimate the room impulse response $\mathcal{RIR}$ by separating it from the clean speech $\mathcal{S}_C$ (see Eq. 1). The input $\mathcal{S}_R$ is first encoded using a Reverberant Speech Encoder $\mathcal{E}_R$. The latent output $\mathcal{S}_E$ is then passed to two different encoders in two different branches. While one of these branches solves the RIR estimation task, the other solves the speech dereverberation task by estimating $\mathcal{S}_C$. Outputs from both the Speech Dereverberation Encoder $\mathcal{S}_{EC}$ and RIR Encoder $\mathcal{R}_E$ are fused with ResNet-18 encodings from the panoramic image $\mathcal{I}_P$ and Geo-Mat feature $\mathcal{I}_G$ respectively. The output latent multi-modal encodings $\mathcal{I}_{EP}$ and $\mathcal{I}_{EG}$ are then passed to a trainable Residual Vector Quantization module (RVQ), which quantizes $\mathcal{F}_S$ to latent codes $\mathcal{Q}_S$, and $\mathcal{F}_R$ to latent codes $\mathcal{Q}_R$. Finally, the HiFi-GAN vocoder decodes the enhanced speech $\mathcal{S}_C$ from $\mathcal{Q}_S$ and the RIR decoder decodes estimated early components of RIR $\mathcal{R}_{\mathcal{RS}}$ from $\mathcal{Q}_R$ which are used to calculate losses for training. At inference time, our CRIP retrieves an RIR from a database $\mathcal{DS}$ and is used to improve late reverberation in the estimated RIR. Finally, post addition, the final estimated RIR is convolved with any $\mathcal{S}_C$ to make it sound like it was uttered in $\mathcal{E}_S$.

for dynamically moving listeners and sources use a 3D geometric representation of the environment, the locations of the speaker and listener, and a few measured RIRs from the same real environment. In many real-world scenarios, we do not have the 3D representation of the environment and measured RIR from the same environment.

Recent works demonstrate the feasibility of estimating early RIR components from reverberant speech [69, 81], and predicting late reverberation from the image of the environment [79]. Furthermore, recent advances involve neural algorithms for converting reverberant sounds between different environments [11, 13, 46, 80], diverging from traditional methods reliant on pre-computed RIRs as described in Eq. 1. However, these end-to-end approaches often incur substantial latency due to deep model inference, making them less viable for real-time mobile applications.

Speech Dereverberation. The human auditory cortex's ability to adaptively filter out reverberation in various acoustic environments is well-documented [32]. Inspired by this, researchers have developed speech enhancement systems capable of transforming reverberant speech to anechoic speech [37, 56, 58]. Initially focused on multi-microphone inputs [19, 33, 54], recent deep learning techniques have shown promise with single-channel inputs [23, 42, 60, 89]. While visual cues have been explored for speech enhancement, most studies have concentrated on near-field ASR using visible lip movements of the speaker [27, 30, 91]. Recent works use panoramic room images for speech dereverberation [14, 17].

3. Methodology

3.1. Overview: AV-RIR

Fig. 2 gives an overview of our approach. Given a reverberant speech $\mathcal{S}_R$, the task of AV-RIR is to learn to estimate the RIR from $\mathcal{S}_R$. To achieve this, we propose a novel multi-modal neural architecture and solve two parallel tasks for learning accurate RIR estimation. As input, together with the $\mathcal{S}_R$, AV-RIR also receives the RGB panoramic image $\mathcal{I}_P$ of the source environment and our proposed Geo-Mat feature map $\mathcal{I}_G$. The construction of $\mathcal{I}_G$ is illustrated in Fig. 3. The $\mathcal{S}_R$ is first passed through a reverberant speech encoder, after which AV-RIR breaks into two branches that solve two different tasks. The bottom branch, which also receives ResNet-18 encoded features of $\mathcal{I}_G$, solves our primary RIR estimation task. The other branch, which receives the encoded features of $\mathcal{I}_P$, solves an auxiliary speech dereverberation task to predict enhanced speech $\mathcal{S}_C$. After the multi-modal feature fusion step, both branches employ a Residual Vector Quantizer module. During inference, to synthesize any target speech as if spoken in the source environment, we convolve the estimated RIR with the target clean speech. Additionally, we propose CRIP (Fig. 4) to retrieve an RIR from a datastore $\mathcal{DS}$, conditioned on the

$\mathcal{I}_P$ of the source environment, to improve late components. Next, we will describe each module in detail.

3.2. AV-RIR Architecture

Reverberant Speech Encoder ($\mathcal{E}_R$). Our $\mathcal{E}_R$ consists of a simple CNN-based architecture with a single 1-D CNN layer and a single input and output channel. As speech dereverberation and RIR estimation are similar learning problems and based on convolution operations (the latter learns deconvolution, and the former learns the inverse), the latent output $\mathcal{S}_E$ from the encoder serves as efficient representations for both tasks AV-RIR solves.

Room Impulse Response Encoder ($\mathcal{E}_{IR}$). $\mathcal{E}_{IR}$ is adapted and modified from the S2IR-GAN encoder [69]. Our three-layer $\mathcal{E}_{IR}$ has 256, 512, and 1024 output channels, 14401, 41, and 41 kernel lengths, and 225, 2, and 2 strides, respectively. The large kernel length in the first layer encodes the RIR features efficiently. We significantly reduce the input dimension by a factor of 900. We process reverberant speech $\mathcal{S}_R$ segments of $\mathbb{R}^{1\times 14400}$ samples. Therefore, every reverberant speech sample is encoded into $\mathbb{R}^{1024\times 16}$ RIR temporal features.

Vision Encoders ($\mathcal{E}_P$, $\mathcal{E}_{GM}$). Prior work on audio-visual speech dereverberation and localization has shown that ResNet-18 [29] is capable of extracting strong cues from image $\mathcal{I}_P$ and depth maps [13, 14, 53]. Therefore, we employ two separate ResNet-18-based feature encoders $\mathcal{E}_P$ and $\mathcal{E}_{GM}$ to encode the $\mathcal{I}_P$ and the Geo-Mat feature $\mathcal{I}_G$, respectively, and reshape the features to $\mathbb{R}^{1024\times 4}$.

Multi-modal Fusion Modules ($\mathcal{M}$). Similar to previous neural audio codec architectures [91], we fuse the visual features with the audio stream along the temporal axis. The combined audio-visual encoded representation is projected into the designed multi-dimensional space [90] and passed to the next stage to quantize into codes.

Residual Vector Quantizer (RVQ). RVQ are used in neural audio codecs to compress the encoder output into a discrete set of code vectors to transmit the data at a fixed target bitrate $\mathcal{R}$ (bits/second). For AV-RIR, we modify the RVQ proposed in SoundStream [93]. SoundStream proposes a VQ with trainable codebooks that are trained together with the model end-to-end. SoundStream [93] adapts the VQ proposed in [71, 87] and improves the codebook with the multi-stage VQ [88]. Our primary modification involves relaxing the constraints (i.e., the target bitrate) to improve the speech dereverberation performance. SoundStream is designed for real-time transmissions and streaming compressed audio at 3-18 Kbps. For our task, we relaxed the compression to ≈59 Kbps. We observed that relaxing beyond 59 Kbps did not significantly improve the performance. Audio codecs have shown increased performance by increasing the birates in subjective tests [20]. Our RVQ cascades $N_q = 64$ layers of VQ and uses a large codebook size $\mathcal{N} = 8192$. Having a larger $\mathcal{N}/\mathcal{N}_q$ ratio has been shown to achieve higher coding efficiency [93].

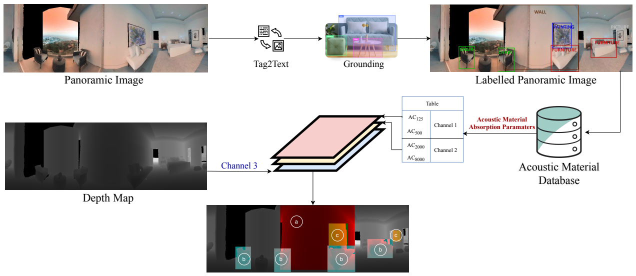

Figure 3. The computation pipeline of **Geo-Mat feature map**. The first two channels of the Geo-Mat feature ($\mathcal{I}_G$) comprise the absorption coefficients ($\mathcal{AC}$) of each acoustic material. The third channel comprises the depth map. We illustrate objects in the environment having similar $\mathcal{AC}$ with similar colors: chairs and furniture with similar materials are represented in light blue, painting, and wall pictures with similar materials are represented in yellow, and the rest in grey. More details on the method to obtain $\mathcal{AC}$ is described in Section 3.3.

Decoders ($\mathcal{D}_{IR}$, $\mathcal{D}_S$). As the two branches of AV-RIR are responsible for decoding separate outputs from the compressed codes (i.e., enhanced speech and RIR), we use two different decoders for this task. We use a HiFi-GAN vocoder for the speech dereverberation branch to decode enhanced speech from the compressed code. HiFi-GAN [40] has shown impressive performance in generating high-fidelity speech, especially with audio codecs [90]. For the RIR estimation branch, we use a modified SoundStream decoder [93]. We modify the decoder to have 6 transposed convolutional (Conv) blocks with output channels of (256, 128, 64, 32, 32, 16) and strides of (5, 5, 2, 2, 1, 1). The output from the last transposed Conv block is passed to a final 1D Conv layer with kernel size 1 and stride 1 to project the code to the waveform domain.

3.3. Geo-Mat Features

The Geo-Mat feature ($\mathcal{I}_G$) represents the geometry and sound absorption properties of materials in the environment. Physics-based RIR simulators take the 3D geometry of the environment and the material absorption coefficients ($\mathcal{AC}$) of each material present in the environment as input to accurately generate RIR for that environment [7, 8, 12]. On the other hand, Changhan *et al.* [14] show that leveraging depth maps improves the performance of audio-visual dereverberation. Inspired by these techniques, we propose Geo-Mat $\mathcal{I}_G$ to improve AV-RIR's understanding of geometry and material information of the environment. Our proposed $\mathcal{I}_G$ is a 3-channel feature map constructed before training our AV-RIR using $\mathcal{AC}$ as the first two channels and the depth map as the last channel. We represent $\mathcal{I}_G$ with 3 channels to encode $\mathcal{I}_G$ using commonly used image encoder [29].

Obtaining material absorption coefficients ($\mathcal{AC}$). To obtain absorption coefficients $\mathcal{AC}$ of all materials in the environment from image $\mathcal{I}_P$, we employ a language-guided pipeline with SOTA pre-trained models. We first use Tag-2-Text [31], a SOTA object tagging model that identifies all objects in the image. This is followed by Grounding DINO [38, 51], which provides bonding box locations of each object identified by Tag-2-Text. Next, we use a large-scale room acoustic database with measured frequency-dependent $\mathcal{AC}$ to match the $\mathcal{AC}$ for each detected object [39]. For the matching operation, we adhere to a simple semantic matching technique to match the material names in the database to the ones detected using Tag-2-Text using embeddings from sentence transformer [72]. Precisely, we calculate an embedding $e_{\mathcal{I}_P} \in \mathbb{R}^{768}$ for every object detected by Tag-2-Text and an embedding $e_{\mathcal{M}} \in \mathbb{R}^{768}$ for every object in the database. Then we take the coefficient of the material in the database with the highest cosine similarity to $e_{\mathcal{I}_P}$. The acoustic $\mathcal{AC}$ are frequency-dependent [85] and, therefore, we use the sub-band $\mathcal{AC}$ at 125 Hz, 500 Hz, 2000 Hz, and 8000 Hz to create the Geo-Mat feature $\mathcal{I}_G$.

Feature Map Construction. After obtaining the absorption coefficients ($\mathcal{AC}$) for each material in the panoramic image ($\mathcal{I}_P$), we finally construct our 3-channel Geo-Mat feature $\mathcal{I}_G$ (Eq. 2). **Channel 1:** The material $\mathcal{AC}$ at low frequencies (i.e., 125 Hz and 500 Hz) **Channel 2:** The $\mathcal{AC}$ of the materials at high frequencies (i.e., 2000 Hz and 8000 Hz). **Channel 3:** The last channel of the $\mathcal{I}_G$ represents the monocular depth map ($\mathcal{I}_D$) of the $\mathcal{I}_P$. Most datasets used in our experiments provide depth maps; however, for datasets that do not, we use the system provided by Godard *et al.* [28] to compute the depth map from a single image. We notice that the order of the channels does not matter.

$$\begin{aligned}
\mathcal{I}_{\rm G}[:,:,0] &= \mathcal{AC}_{125} + \mathcal{AC}_{500} * 16 \\
\mathcal{I}_{\rm G}[:,:,1] &= \mathcal{AC}_{2000} + \mathcal{AC}_{8000} * 16 \\
\mathcal{I}_{\rm G}[:,:,2] &= \mathcal{I}_D
\end{aligned} \tag{2}$$

3.4. Training AV-RIR

$\mathcal{D}_{QIR}$ and $\mathcal{D}_{QS}$ are RIR and clean speech quantizers followed by a decoder, respectively, and $\mathcal{CRIP}$ represents CRIP. We estimate clean speech ($\hat{\mathcal{S}_C}$), RIR ($\mathcal{R\hat{I}R}$), and reverberant speech ($\hat{\mathcal{S}_R}$) as shown below.

$$\begin{aligned}
&\hat{\mathcal{S}_{\rm C}} = \mathcal{D}_{\rm QS}(\mathcal{E}_{\rm P}(\mathcal{I}_{\rm P}),\ \mathcal{E}_{\rm S}(\mathcal{S}_{\rm R})). \\
&\mathcal{R\hat{I}R} = \mathcal{D}_{\rm QIR}(\mathcal{E}_{\rm GM}({\rm I}_{\rm GM}),\ \mathcal{E}_{\rm IR}(\mathcal{S}_{\rm R}))\ + \mathcal{CRIP}({\rm I}_{\rm P}). \\
&\hat{\mathcal{S}_{\rm R}} = \hat{\mathcal{S}_{\rm C}} \circledast \mathcal{R\hat{I}R}.
\end{aligned} \tag{3}$$

RIR Estimation Loss. We calculate the time-domain mean squared error (MSE) for the estimated RIR as follows.

$$\mathcal{L}_{\rm MSE} = \mathbb{E}[||\mathcal{RIR}\text{ - }\mathcal{R\hat{I}R}||_2]. \tag{4}$$

To train our RIR estimation RVQ codebook, we use the exponential moving average loss proposed in [87] as our vector quantizer (VQ) loss $\mathcal{L}_{VQ1}$.

Speech Dereverberation Loss. For solving the speech dereverberation task, we optimize three losses:

(1) Mel-Spectrogram (Mel) loss ($\mathcal{L}_{\rm Mel}$). The mel-spectrogram loss helps improve the perceptual quality of the predicted enhanced speech [34]. The 1D waveform ($\hat{\mathcal{S}_R}$) output from the HiFi-GAN vocoder is first converted to the Mel-spectrogram, transforming it from the time domain to the frequency domain representation. A drawback of the Mel loss is its fixed resolution. The window length determines whether it has a good frequency or time resolution (i.e., a wide window has good frequency resolution and poor time resolution). Therefore, we calculate $\mathcal{L}_{Mel}$, over a range of window lengths W_L = {64, 128, 256, 512, 1024, 2048, 4096}. Formally, $\mathcal{L}_{Mel}$ can be defined as:

$$\begin{aligned}
\mathcal{L}_{\rm MEL} = \mathbb{E}[||&{\rm MEL}(\mathcal{S}_{\rm R})\text{-}{\rm MEL}(\hat{\mathcal{S}_{\rm R}}))||_1 + \\
&||{\rm MEL}(\mathcal{S}_{\rm C})\text{-}{\rm MEL}(\hat{\mathcal{S}_{\rm C}}))||_1],
\end{aligned} \tag{5}$$

where MEL() is the operation that converts time-domain speech into its mel-spectrogram representation.

(2) Short-Time Fourier Transform (STFT) loss ($\mathcal{L}_{\rm STFT}$). The STFT loss helps in the high-fidelity reconstruction of the predicted enhanced speech [14]. The 1D waveform is first converted to the frequency domain by applying the STFT() operation. The STFT of a waveform can be represented as a complex spectrogram where $\mathcal{M}_S$ represents the magnitude of the STFT and $\mathcal{P}_S$ is the phase of the STFT. We map the phase angle of the STFT to the rectangular coordinate on the unit circle to avoid phase wraparound issues [14]. Our $\mathcal{L}_{STFT}$ is the sum of magnitude loss $\mathcal{L}_{MAG}$ and phase loss $\mathcal{L}_{PH}$. Similar to $\mathcal{L}_{Mel}$, we calculate $\mathcal{L}_{STFT}$ over a range of window lengths in a similar setting.

$$\begin{aligned}
&\mathcal{L}_{\rm MAG} = \mathbb{E}[||{\rm M}_{\rm S}(\mathcal{S}_{\rm R})\text{-}{\rm M}_{\rm S}(\hat{\mathcal{S}_{\rm R}}))||_2 + ||{\rm M}_{\rm S}(\mathcal{S}_{\rm C})\text{-}{\rm M}_{\rm S}(\hat{\mathcal{S}_{\rm C}}))||_2]. \\
&\mathcal{L}_{\rm P}({\rm x},\hat{\rm x}) = \mathbb{E}[||\sin({\rm P}_{\rm S}({\rm x}))\text{-}\sin({\rm P}_{\rm S}(\hat{\rm x})))||_2 \\
&\qquad\qquad + ||\cos({\rm P}_{\rm S}({\rm x}))\text{-}\cos({\rm P}_{\rm S}(\hat{\rm x}))||_2]. \\
&\mathcal{L}_{\rm PH} = \mathcal{L}_{\rm P}((\mathcal{S}_{\rm R}, (\hat{\mathcal{S}_{\rm R}}) + \mathcal{L}_{\rm P}(\mathcal{S}_{\rm C}, \hat{\mathcal{S}_{\rm C}}). \\
&\mathcal{L}_{\rm STFT} = \mathcal{L}_{\rm MAG} + \mathcal{L}_{\rm PH}
\end{aligned} \tag{6}$$

Our total metric loss $\mathcal{L}_{METRIC}$ (including RIR estimation and speech derverberation) is described in Eq. 7.

$$\mathcal{L}_{\rm METRIC} = \mathcal{L}_{\rm MEL} + \lambda_1 \mathcal{L}_{\rm STFT} + \lambda_2 \mathcal{L}_{\rm MSE}, \tag{7}$$

where $\lambda_1$and λ_2 are the weights.

(3) Adversarial loss ($\mathcal{L}_{\rm ADV}$) In addition to metric loss, we train our network using adversarial loss. We train separate discriminator networks $\mathcal{D}_R$ and $\mathcal{D}_S$ for reverberant and clean speech respectively. Our $\mathcal{L}_{ADV}$ is described in Eq. 8.

$$\mathcal{L}_{\rm ADV} = \mathbb{E}[\max(0, 1\text{-}\mathcal{D}_{\rm R}(\hat{\mathcal{S}_{\rm R}}) + \max(0, 1\text{-}\mathcal{D}_{\rm S}(\hat{\mathcal{S}_{\rm C}})], \tag{8}$$

We train speech dereverbaration RVQ codebook using VQ loss $\mathcal{L}_{VQ2}$ [87]. Eq. 9 presents our total generator loss $\mathcal{L}_{Gen}$. In Eq. 9, λ_1 and λ_2 are the weights.

$$\mathcal{L}_{\rm GEN}({\rm x}) = \mathcal{L}_{\rm METRIC} + \lambda_1 \mathcal{L}_{\rm ADV} + \lambda_2(\mathcal{L}_{\rm VQ1} + \mathcal{L}_{\rm VQ2}), \tag{9}$$

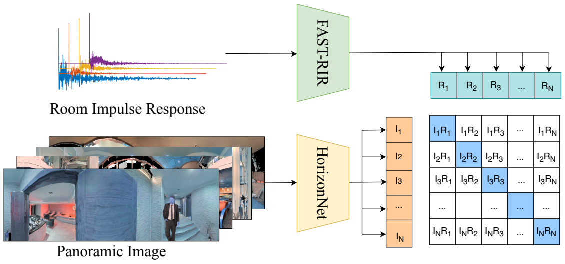

Figure 4. Illustration of **CRIP** training. Like CLIP [62], we propose two networks, one to encode a panoramic image and the other to encode the RIR to learn a joint embedding space between both. We use our CRIP-based image-to-RIR retrieval during inference to improve late reverberation in the estimated RIR from AV-RIR.

Discriminator ($\mathcal{D}_\mathcal{R}, \mathcal{D}_\mathcal{S}$). We use the multi-period discriminator network (MPD) proposed by HiFi-GAN [40] and the multi-scale discriminator network (MSD) proposed in MelGAN [43]. MPD effectively captures the periodic details by having several sub-discriminators, each handling different parts of the input audio. MSD captures consecutive patterns and long-term dependencies.

3.5. Contrastive RIR-Image Pretraining (CRIP)

We propose CRIP, a model built on the fundamentals of CLIP [62], that learns a joint embedding space between the panoramic image ($\mathcal{I}_P$) and their corresponding RIRs. Similar to CLIP, CRIP employs two encoders, a pre-trained HorizonNet encoder [83] $\mathcal{E}_H$ which serves as our $\mathcal{I}_P$ encoder, and the discriminator network proposed in FAST-RIR [68], which serves as our RIR encoder $\mathcal{E}_F$. Formally, $\mathcal{E}_H$ takes as input $\mathcal{I}_P$ and outputs an embedding $\mathcal{I}_{EM} \in \mathbb{R}^{N\times 1024}$, and $\mathcal{E}_F$ takes as input $\mathcal{RIR} \in \mathbb{R}^{N\times 1\times 4096}$ and output an embedding $\mathcal{R}_{EM} \in \mathbb{R}^{N\times 1024}$. Finally, we measure similarity by calculating the dot product between the two as follows:

$$C_{r\text{-}2\text{-}i} = \tau * \left(\mathcal{RIR} \cdot \mathcal{I}_{EM}^\top\right)$$
$$C_{i\text{-}2\text{-}r} = \tau * \left(\mathcal{I}_{EM} \cdot \mathcal{RIR}^\top\right) \tag{10}$$

where τ is the temperature. This is followed by calculating the RIR-to-Image loss $\ell_{r\text{-}2\text{-}i}$ and the Image-to-RIR loss $\ell_{r\text{-}2\text{-}i}$ as follows:

$$\ell_{r\text{-}2\text{-}i} = \frac{1}{N}\sum_{i=0}^{N} \log \operatorname{diag}(\operatorname{softmax}(C_{r\text{-}2\text{-}i}))$$
$$\ell_{r\text{-}2\text{-}i} = \frac{1}{N}\sum_{i=0}^{N} \log \operatorname{diag}(\operatorname{softmax}(C_{i\text{-}2\text{-}r})) \tag{11}$$

Finally, we optimize the average of both losses:

$$\mathcal{L} = 0.5 * (\ell_{r\text{-}2\text{-}i} + \ell_{i\text{-}2\text{-}r}) \tag{12}$$

Why CRIP? Neural-network-based RIR estimators are known to inaccurately approximate the late components of the RIR as a sum of decaying filtered noise [81]. Similarly, while our codec-based AV-RIR can accurately estimate the early components with structured impulsive patterns, it cannot precisely estimate late reverberation, which generally contains noise-like components. Thus, we propose CRIP to fill this gap. CRIP uses HorizonNet that captures room geometry/layout information [83]. The late components of RIR depend on the geometry of the room [61].

CRIP for AV-RIR Inference. During inference, for $\mathcal{I}_P$ of the target scene, we retrieve an RIR from a datastore $\mathcal{DS}$. The retrieval is performed by calculating cosine similarity between the CRIP embeddings of $\mathcal{I}_P$, denoted as $\mathcal{I}^t_{EM}$, and the RIR embeddings for all RIRs in datastore $\mathcal{DS}$. $\mathcal{DS}$ is generally a large collection of RIRs in the wild, which in our case is composed of synthetic RIRs, more details on which can be found in Section 4. The final estimated RIR is obtained by replacing the late components of the original estimated RIR $R\hat{I}R$ with the late components of the retrieved RIR $\mathcal{RIR}_{CRIP}$. We perform hyper-parameter tuning to find the optimal number of samples S from $R\hat{I}R$ to replace with $\mathcal{R}_{CRIP}$, and we found S = 2000 to give us the best improvements. This whole process can be formalized as:

$$\mathcal{R}\hat{\mathcal{I}}\mathcal{R}[2000:4000] = \mathcal{RIR}_{CRIP}[2000:4000]. \tag{13}$$

4. Experiments and Results

Datasets. For training and evaluation, we use the widely adopted SoundSpaces dataset [10, 12]. The SoundSpaces dataset provides paired reverberant speech and its RIR. The data is sourced by convolving simulated RIRs with clean speech from the LibriSpeech [59] dataset. The RIRs are simulated using a geometrical acoustic simulation techniques [74, 75] with environments taken from the Matterport3D dataset [9]. To additionally evaluate how AV-RIR fairs in speech dereverberation in real-world scenarios, we use web videos in the filtered AVSpeech dataset [22] proposed in VAM [11]. Since AVSpeech does not have ground truth (GT) RIR, we only used it to evaluate our speech dereverberation pipeline. Our datastore $\mathcal{DS}$ comprises synthetic RIRs generated from SoundSpaces, excluding test set RIRs.

Hyperparameters. We train AV-RIR with a batch size of 16 for 400 epochs with only metric loss (Eq. 7) and VQ loss ($\mathcal{L}_{VQ1}$, $\mathcal{L}_{VQ2}$). Later, we train with total loss (Eq. 9) for 1K epochs. We use Adam Optimizer [36] with β_1 = 0.5, β_2 = 0.9 and learning rate 5×10^{-5}. For every 200K steps, we decay the learning rate by 0.5.

4.1. RIR Estimation

Evaluation Metrics. We quantitatively measure the accuracy of estimated RIR using standard room acoustic metrics. Reverberation time (T_{60}), direct-to-reverberant ratio (DRR), and early decay time (EDT) are the commonly used room acoustic statistics. We calculate the mean absolute difference between the acoustic statistics of estimated and ground truth (GT) RIRs as the error. T_{60} measures the time taken for the sound pressure to decay by 60 decibels (dB), and EDT is 6 times the time taken for the sound pressure to decay by 10 dB. T_{60} depends on the room size and room materials, and EDT depends on the type and location of the sound source [66]. DRR is the ratio between the sound pressure level of the direct sound source and the reflected sound [58]. We also report the mean square difference (MSE) between the GT and estimated early component (EMSE) and late component (LMSE) of the RIR in time domain. We show the benefit of RIR estimated from our approach in SLP tasks in our supplementary.

Baselines. We compare the performance of AV-RIR with six other baselines. (1) **Image2Reverb [79]:** Predicts RIR from the camera-view image. (2) **Visual Acoustic Matching (VAM) [11]:** Takes as input the source audio and the target environment image and outputs resynthesized audio matching the target environment. VAM does not explicitly estimate the RIR of the target environment. Therefore, we compare VAM only on our perceptual evaluation. (3) **FAST-RIR++ [68]**: Takes as input the room geometry, source and listener positions, and T_{60} to generate the RIR. We modified the architecture to the input panoramic image ($\mathcal{I}_P$) of the environment and estimate the RIR (FAST-RIR++). (4) **CRIP-only** *(ours)*: CRIP retrieves the closest RIR from the large synthetic dataset $\mathcal{DS}$ using a panoramic image $\mathcal{I}_P$ as input. We use it as a baseline to evaluate how much improvement our audio-visual network contributes. (5) **Filtered Noise Shaping Network (FiNS) [81]:** An audio-only time domain RIR estimator from reverberant speech. (6) **S2IR-GAN [69]:** An audio-only GAN-based reverberant speech-to-IR estimator.

Results. Table 1 compares our approach AV-RIR with audio-only and visual-only baselines. We can see that the audio-only baselines outperform the visual-only baselines. Audio and visual cues provide complementary information for RIR estimation, and we can see a significant boost in performance in our AV-RIR, which inputs audio and visual cues. AV-RIR outperforms the SOTA audio-only approach S2IR-GAN by 36%, 42%, 63%, 89% and 98% on T_{60}, DRR, EDT, EMSE, and LMSE, respectively.

4.2. Speech Dereverberation

Evaluation Metrics. We evaluate speech dereverberation using our AV-RIR on two downstream tasks: Automatic Speech Recognition (ASR) and Speaker Verification (SV). We use standard metrics such as Word Error Rate (WER) to measure ASR performance and Equal Error Rate (EER) to measure the SV. Following prior works [14, 17], we used pre-trained models from SpeechBrain [70] to evaluate the ASR and SV performance. AVSpeech dataset does not have parallel clean speech to perform ASR and SV tasks and we use the Reverberation Time Error (RTE) metric [17] to evaluate the estimated clean speech from our AV-RIR.

Table 1. We compare the RIR estimated using our AV-RIR with prior visual-only method (Image2Reverb [79]) and audio-only methods (FiNS [81] and S2IR-GAN [69]). We perform an ablation study to show the benefit of each component of our network.

	Method	T_{60} Error (ms)	DRR Error (dB)	EDT Error (ms)	EMSE ($x10^{-5}$)	LMSE ($x10^{-5}$)
	Image2Reverb [79]	131.7	4.94	382.1	4907	1126
	FAST-RIR++ [68]	126.4	3.62	334.2	2630	990
	FiNS [81]	87.7	3.30	235.7	924	561
	S2IR-GAN [69]	63.1	3.04	168.3	730	310
ABLATION	AV-RIR (Audio-Only)	88.8	2.96	122.4	176	51
	AV-RIR w Random	77.6	2.67	109.2	124	6
	AV-RIR w/o CRIP	61.7	2.07	79.8	79	42
	AV-RIR w/o Geo-Mat	55.7	1.98	74.1	104	6
	AV-RIR w/o Multi-Task	77.8	2.56	105.4	144	6
	AV-RIR w/o STFT Loss	59.4	1.94	77.2	123	6
	CRIP-only *(ours)*	118.9	3.14	298.4	212	6
	AV-RIR ***(ours)***	**40.2**	**1.76**	**62.1**	**82**	**6**

Table 2. Performance comparison of AV-RIR with audio-only baselines (marked with ‡) and audio-visual baselines (marked with ∗) on the SLP tasks on the SoundSpaces dataset. "Reverberant" refers to clean speech convolved with ground-truth RIR. We also report RTE for real-world audio from the AVSpeech dataset.

	Method	Speech Recognition[†] WER (%) ↓	Speaker Verification[†] EER (%) ↓	RTE* ↓ (in sec)
	Clean (Upper bound)	2.89	1.53	-
	Reverberant	8.20	4.51	0.382
	MetricGAN+ [26]‡	7.48 (+9%)	4.67 (-4%)	0.187 (+51%)
	DEMUCS [21]‡	7.97 (+3%)	3.82 (+15%)	0.129 (+66%)
	HiFi-GAN [82]‡	9.31 (-14%)	4.32 (+4%)	0.196 (+49%)
	WPE [57]‡	8.43 (-3%)	5.90 (-31%)	0.173 (+55%)
	VoiceFixer [49]‡	5.66 (+31%)	3.76 (+16%)	0.121 (+68%)
	SkipConvGAN [42]‡	7.22 (+12%)	4.86 (-8%)	0.119 (+69%)
	Kotha *et al.* [41]‡	5.32 (+35%)	3.71 (+17%)	0.124 (+68%)
	VIDA [14]∗	4.44 (+46%)	3.97 (+12%)	0.155 (+59%)
	AdVerb [17]∗	**3.54 (+57%)**	3.11 (+31%)	0.101 (+74%)
ABLATION	AV-RIR (Audio-Only)	5.24 (+36%)	2.67 (+41%)	0.055 (+86%)
	AV-RIR w Random Image	4.85 (+41%)	2.56 (+43%)	0.049 (+87%)
	AV-RIR w/o Multi-Task	4.57 (+44%)	2.55 (+43%)	0.048 (+87%)
	AV-RIR w/o CRIP	4.67 (+43%)	2.66 (+41%)	0.049 (+87%)
	AV-RIR w/o Geo-Mat	4.54 (+45%)	2.21(+51%)	0.044 (+88%)
	AV-RIR w/o MEL Loss	4.44 (+46%)	2.44 (+46%)	0.048 (+87%)
	AV-RIR ***(ours)***	4.17 (+49%)	**2.02 (+55%)**	**0.042 (+89%)**

Baselines. We compared the speech dereverberation using our approach with the following prior audio-only and audio-visual speech enhancement networks. (1) **Audio-Only:** WPE [57] is a statistical-model-based speech dereverberation network that cancels late reverberation without the knowledge of RIR. MetricGan+ [26], DEMUCS [21], HiFi-GAN [82], VoiceFixer [49], SkipConvGAN [42] and Kotha *et al.* [41] are learning-based speech dereverberation networks. (2) **Audio-Visual :** VIDA [14] is the first audio-visual speech dereverberation network that takes $\mathcal{I}_P$ as visual cues. Recently, geometry-aware AdVerb [17] has been shown to achieve SOTA results in downstream speech tasks.

Results. Table 2 shows the benefits of speech dereverberation from our AV-RIR in the ASR and SV task. We can see that AV-RIR outperforms SOTA network AdVerb in SV tasks by 35% and outperforms all the baselines except for AdVerb in the ASR. To evaluate the robustness of AV-RIR,

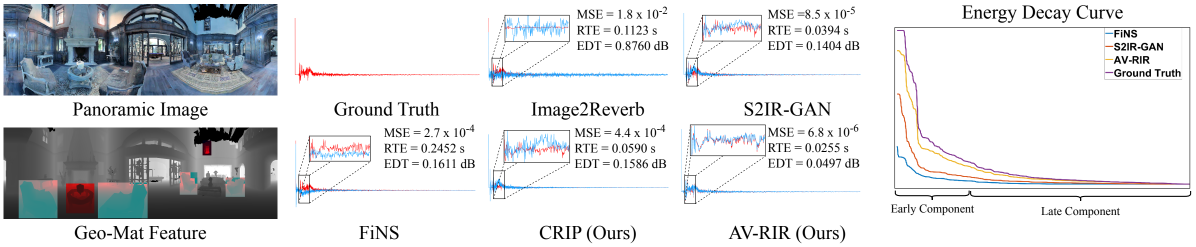

Figure 5. **Qualitative Results**. (Left) We show the Geo-Mat feature generated using our approach. The cushion chairs with a similar material absorption property are represented in green. The table and window with similar material are represented in red. (Right) We plot the time-domain representation of the RIRs estimated using prior methods and our approach with the ground truth (GT) RIR (GT: Red, Estimated: Blue). We also report the MSE (Eq. 4), T_{60} error (RTE), and EDT error (EDT). It can be seen that the RIR estimated using our AV-RIR matches closely with GT RIR when compared with the baseline. Also, we can see that the RIR retrieved from our CRIP has similar late components as the GT RIR. However, the early component of the retrieved RIR (shown in zoom) significantly differs from the GT. Our full AV-RIR pipeline estimates the early components of the RIR using audio-visual features and adds the late component of the RIR from our CRIP to accurately predict the full RIR. The energy decay curve (EDC) depicts the energy remaining in the RIR over time [78]. We can see that the EDC of the late component of RIR estimated from AV-RIR (yellow) matches closely with the GT RIR (purple).

we test our network on recorded speech not used for training in the AVSpeech using the RTE metric. We can see that our work outperforms all the baselines by 60%.

Ablation Study. We perform a comprehensive ablation study to show the benefit of different components of our AV-RIR. (1) **Multi-task learning.** To prove the effectiveness of the multi-task learning approach, we train the branches separately (AV-RIR w/o Multi-Task). Table 1 and Table 2 show that our multi-task learning approach benefits both tasks mutually. While the RIR estimation performance improves by 31% - 48%, speech dereverberation performance improves by 13% - 21%. (2) **CRIP.** Table 1 shows that adding CRIP during inference for late reverberation improves the late component MSE (LMSE) by 86%. (3) **Geo-Mat feature.** Table 1 shows that the Geo-Mat feature improves RIR estimation accuracy by 11% - 28%. (4) **Visual cues.** To prove the effectiveness of visual cues, we discard them while training and directly pass RIR and speech encoder outputs to RVQ. Additionally, we also discard CRIP and directly estimate the full RIR. Table 1 and Table 2 show that AV-RIR outperforms our audio-only AV-RIR variation in RIR estimation tasks by 41% - 55% and speech dereverbation task by around 24%.

4.3. Perceptual Evaluation

In Table 3 we report scores for perceptual evaluation of our estimated RIRs. For each environment, we provide the ground truth (GT) speech, generated speech using Image2Reverb [79], VAM [11], our AV-RIR and the environment image. The participants were asked to select the generated speech that sounds closer to the GT speech.

Deatils on initial pre-screening of participants is described in the Appendix. We select 6 scenes with varying complexity with T_{60} ranging from 0.2 seconds to 0.7 seconds. We can see that, irrespective of the T_{60} and the environment complexity, 56% to 79% of the participants said that the generated speech from AV-RIR closely matched the GT speech.

Table 3. **Perceptual Evaluation**. Participants find that the reverberant speech generated using our AV-RIR is closer to GT reverberant speech when compared to visual-only baselines.

Scene	T_{60}	Image2Reverb [79]	VAM[11]	AV-RIR (Ours)
Scene 1	0.22	2%	19%	**79%**
Scene 2	0.31	16%	26%	**58%**
Scene 3	0.35	14%	16%	**70%**
Scene 4	0.38	5%	40%	**56%**
Scene 5	0.47	16%	16%	**67%**
Scene 6	0.65	14%	12%	**74%**

5. Conclusion, Limitations and Future Work

We propose AV-RIR, a novel multi-modal multi-task learning approach for RIR estimation. AV-RIR leverages both audio and visual cues using a novel neural codec-based multi-modal architecture and solves speech dereverberation as the auxiliary task. We also propose Contrastive RIR-Image Pre-training (CRIP), which improves late reverberation components in estimated RIR using retrieval. Both quantitative metrics and perceptual studies show that our AV-RIR significantly outperforms all the baselines. We evaluate the speech dereverberation performance on the recorded AVSpeech dataset not used for training and observe that our approach outperforms the baselines by 60%.

AV-RIR assumes stationary single-talker input speech or single-source audio without noise. Future work aims to tackle multi-channel RIR estimation and RIR estimation from noisy, multi-source environment with moving sources.

References

[1] *About this reverberation business*. 1977. 2

[2] Steam audio, 2018. 1

[3] Oculus spatializer, 2019.

[4] Microsoft project acoustics, 2019. 1

[5] Nobuharu Aoshima. Computer-generated pulse signal applied for sound measurement. *Journal of the Acoustical Society of America*, 69:1484–1488, 1981. 2

[6] Michael Barron and Timothy J. Foulkes. Auditorium Acoustics and Architectural Design. *The Journal of the Acoustical Society of America*, 96(1):612–612, 1994. 1

[7] Carlos Alberto Brebbia and Robert D. Ciskowski. Boundary element methods in acoustics. 1991. 4

[8] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Trans. Graph.*, 35(6):180:1–180:11, 2016. 4

[9] 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. 6

[10] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3d environments. In *ECCV*, 2020. 2, 6

[11] 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 (CVPR)*, pages 18858–18868, 2022. 3, 6, 7, 8

[12] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022. 4, 6

[13] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6409–6419, 2023. 3, 4

[14] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3, 4, 5, 7

[15] Mingfei Chen, Kun Su, and Eli Shlizerman. Be everywhere - hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7853–7862, 2023. 2

[16] Ziyang Chen, Israel D. Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. 2024. 2

[17] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7884–7896, 2023. 3, 7

[18] K. Crammer and D.D. Lee. Room impulse response estimation using sparse online prediction and absolute loss. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, pages III–III, 2006. 1, 2

[19] M. Dahl, I. Claesson, and S. Nordebo. Simultaneous echo cancellation and car noise suppression employing a microphone array. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 239–242 vol.1, 1997. 3

[20] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. 4

[21] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real Time Speech Enhancement in the Waveform Domain. In *Proc. Interspeech 2020*, pages 3291–3295, 2020. 7

[22] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4), 2018. 6

[23] Ori Ernst, Shlomo E. Chazan, Sharon Gannot, and Jacob Goldberger. Speech dereverberation using fully convolutional networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 390–394, 2018. 3

[24] Angelo Farina. Advancements in impulse response measurements by sine sweeps. *Journal of The Audio Engineering Society*, 2007. 2

[25] Maozhong Fu, Jesper Rindom Jensen, Yuhan Li, and Mads Græsbøll Christensen. Sparse modeling of the early part of noisy room impulse responses with sparse bayesian learning. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590, 2022. 1, 2

[26] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538*, 2021. 7

[27] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual Speech Enhancement. In *Proc. Interspeech 2018*, pages 1170–1174, 2018. 3

[28] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. 2019. 5

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[30] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018. 3

[31] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023. 5

[32] Aleksandar Z Ivanov, Andrew J King, Ben DB Willmore, Kerry MM Walker, and Nicol S Harper. Cortical adaptation to sound reverberation. *eLife*, 11:e75090, 2022. 3

[33] F. Jabloun and B. Champagne. A multi-microphone signal subspace approach for speech enhancement. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, pages 205–208 vol.1, 2001. 3

[34] Takuhiro Kaneko, Shinji Takaki, Hirokazu Kameoka, and Junichi Yamagishi. Generative Adversarial Network-Based Postfilter for STFT Spectrograms. In *Proc. Interspeech 2017*, pages 3389–3393, 2017. 5

[35] Matti Karjalainen, Poju Ansalo, Aki Mäkivirta, Timo Peltonen, and Vesa Välimäki. Estimation of modal decay parameters from noisy response measurements. *Journal of the Audio Engineering Society*, 50(11):867–878, 2002. 1, 2

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[37] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka. A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research. *Journal on Advances in Signal Processing*, 2016, 2016. 3

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 5

[39] Christoph Kling. Absorption coefficient database, 2018. 5

[40] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, pages 17022–17033. Curran Associates, Inc., 2020. 4, 6

[41] Vinay Kothapally and John H.L. Hansen. Complex-Valued Time-Frequency Self-Attention for Speech Dereverberation. In *Proc. Interspeech 2022*, pages 2543–2547, 2022. 7

[42] Vinay Kothapally and John H. L. Hansen. Skipconvgan: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1600–1613, 2022. 3, 7

[43] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 6

[44] Sungho Lee, Hyeong-Seok Choi, and Kyogu Lee. Yet another generative model for room impulse response estimation. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, 2023. 2

[45] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP J. Adv. Signal Process*, 2007(1):187, 2007. 2

[46] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *arXiv preprint arXiv:2302.02088*, 2023. 3

[47] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. In *NeurIPS*, 2023. 2

[48] Yuanqing Lin and D.D. Lee. Bayesian regularization and nonnegative deconvolution for room impulse response estimation. *IEEE Transactions on Signal Processing*, 54(3): 839–847, 2006. 1, 2

[49] Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration. In *Proc. Interspeech 2022*, pages 4232–4236, 2022. 7

[50] Shiguang Liu and Dinesh Manocha. *Sound Synthesis, Propagation, and Rendering*. Morgan & Claypool Publishers, 2022. 1

[51] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5

[52] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *Advances in Neural Information Processing Systems*, pages 3165–3177. Curran Associates, Inc., 2022. 2

[53] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In *Advances in Neural Information Processing Systems*, 2022. 2, 4

[54] C. Marro, Y. Mahieux, and K.U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3):240–259, 1998. 3

[55] Ravish Mehra, Nikunj Raghuvanshi, Lakulish Antani, Anish Chandak, Sean Curtis, and Dinesh Manocha. Wave-based sound propagation in large open scenes using an equivalent source formulation. *ACM Trans. Graph.*, 32(2), 2013. 2

[56] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2):145–152, 1988. 3

[57] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1717–1731, 2010. 7

[58] Patrick A. Naylor and Nikolay D. Gaubitch. *Speech Dereverberation*. Springer Publishing Company, Incorporated, 1st edition, 2010. 3, 7

[59] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. 6

[60] Ashutosh Pandey and DeLiang Wang. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6875–6879, 2019. 3

[61] Ricardo Falcon Perez, Georg Götz, and Ville Pulkki. Machine-learning-based estimation of reverberation time using room geometry for room effect rendering. In *Proceedings of the 23rd International Congress on Acoustics: integrating 4th EAA Euroregio*, page 13, 2019. 6

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[63] Nikunj Raghuvanshi and Hannes Gamper. *Interactive and Immersive Auralization*. Springer, 2022. 1

[64] Nikunj Raghuvanshi, John Snyder, Ravish Mehra, Ming Lin, and Naga Govindaraju. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. *ACM Trans. Graph.*, 29(4), 2010. 2

[65] Anton Ratnarajah and Dinesh Manocha. Listen2scene: Interactive material-aware binaural sound propagation for reconstructed 3d scenes. *arXiv preprint arXiv:2302.02809*, 2023. 2

[66] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. IR-GAN: room impulse response generator for far-field speech recognition. In *Interspeech*, pages 286–290. ISCA, 2021. 2, 7

[67] Anton Ratnarajah, Zhenyu Tang, Rohith Aralikatti, and Dinesh Manocha. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 924–933, New York, NY, USA, 2022. Association for Computing Machinery. 2

[68] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–575, 2022. 2, 6, 7

[69] Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia. Towards improved room impulse response estimation for speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2, 3, 4, 7

[70] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624. 7

[71] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 4

[72] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. 5

[73] Lorenzo Rizzi, Gabriele Ghelfi, and Maurizio Santini. Small-rooms dedicated to music: From room response analysis to acoustic design. *Journal of The Audio Engineering Society*, 2016. 1

[74] Carl Schissler and Dinesh Manocha. Gsound: Interactive sound propagation for games. In *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society, 2011. 6

[75] Carl Schissler and Dinesh Manocha. Interactive sound propagation and rendering for large multi-source scenes. *ACM Trans. Graph.*, 36(4), 2016. 6

[76] Carl Schissler, Christian Loftin, and Dinesh Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1246–1259, 2018. 2

[77] M. R. Schroeder. Integrated-impulse method measuring sound decay without using impulses. *The Journal of the Acoustical Society of America*, 66(2):497–500, 1979. 2

[78] M. R. Schroeder. New Method of Measuring Reverberation Time. *The Journal of the Acoustical Society of America*, 37 (3):409–412, 2005. 8

[79] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, pages 286–295. IEEE, 2021. 2, 3, 7, 8

[80] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching. In *NeurIPS*, 2023. 3

[81] Christian J. Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. Filtered noise shaping for time domain room impulse response estimation from reverberant speech. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 221–225, 2021. 2, 3, 6, 7

[82] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. In *Proc. Interspeech 2020*, pages 4506–4510, 2020. 7

[83] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1047–1056, 2019. 6

[84] 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*, 26(5):1991–2001, 2020. 1

[85] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. Gwa: A large high-quality acoustic dataset for audio processing. In *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022. Association for Computing Machinery. 2, 5

[86] Micah Taylor, Anish Chandak, Qi Mo, Christian Lauterbach, Carl Schissler, and Dinesh Manocha. Guided multiview ray tracing for fast auralization. *IEEE Transactions on Visualization and Computer Graphics*, 18(11):1797–1810, 2012. 2

[87] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 4, 5

[88] A. Vasuki and P.T. Vanathi. A review of vector quantization techniques. *IEEE Potentials*, 25(4):39–47, 2006. 4

[89] Heming Wang and DeLiang Wang. Cross-domain diffusion based speech enhancement for very noisy speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 3

[90] Yi-Chiao Wu, Israel D. Gebru, Dejan Marković, and Alexander Richard. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 4

[91] Karren Yang, Dejan Marković, Steven Krenn, Vasu Agrawal, and Alexander Richard. Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8217–8227, 2022. 2, 3, 4

[92] Hengchin Yeh, Ravish Mehra, Zhimin Ren, Lakulish Antani, Dinesh Manocha, and Ming Lin. Wave-ray coupling for interactive sound propagation in large complex scenes. *ACM Trans. Graph.*, 32(6), 2013. 2

[93] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. 4