AdVerb: Visually Guided Audio Dereverberation

Sanjoy Chowdhury[1]* Sreyan Ghosh[1]* Subhrajyoti Dasgupta[2]
Anton Ratnarajah[1] Utkarsh Tyagi[1] Dinesh Manocha[1]

[1]University of Maryland, College Park [2]Mila and Université de Montréal

{sanjoyc,sreyang,jeran,utkarsht,dmanocha}@umd.edu subhrajyoti.dasgupta@umontreal.ca
Project page - https://gamma.umd.edu/researchdirections/speech/adverb

Abstract

We present AdVerb, a novel audio-visual dereverberation framework that uses visual cues in addition to the reverberant sound to estimate clean audio. Although audio-only dereverberation is a well-studied problem, our approach incorporates the complementary visual modality to perform audio dereverberation. Given an image of the environment where the reverberated sound signal has been recorded, AdVerb employs a novel geometry-aware cross-modal transformer architecture that captures scene geometry and audio-visual cross-modal relationship to generate a complex ideal ratio mask, which, when applied to the reverberant audio predicts the clean sound. The effectiveness of our method is demonstrated through extensive quantitative and qualitative evaluations. Our approach significantly outperforms traditional audio-only and audio-visual baselines on three downstream tasks: speech enhancement, speech recognition, and speaker verification, with relative improvements in the range of 18% - 82% on the LibriSpeech test-clean set. We also achieve highly satisfactory RT60 error scores on the AVSpeech dataset.

1. Introduction

Reverberation occurs when an audio signal reflects from multiple surfaces and objects in the environment to alter the dry sound thereby degrading its quality. Far-field speech recorded at a considerable distance from the speaker is significantly degraded by the strong reverberation effects caused by the environment. The amount of reverberation is highly correlated to the geometry of the surroundings and the materials present in the vicinity [9, 11]. For instance, the auditory experience changes drastically when listening to a pleasant symphony in a large empty hallway vs. a relatively small furnished living room (Fig. 2). Recent studies have shown that the reverberation effects can be estimated from a single image of the environment with reasonable accuracy [69, 46, 36]. Removal of reverberation in recorded speech signals is highly desirable and would help improve the performance of several other auxiliary downstream tasks like automatic speech recognition (ASR), speaker verification (SV), source separation (SP), speech enhancement (SE), etc., which are widely used in several day-to-day tools.

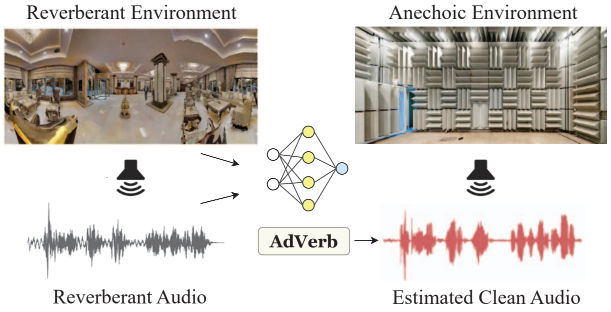

Figure 1: We present AdVerb, a novel audio-visual dereverberation framework that leverages visual cues of the environment to estimate clean audio from reverberant audio. E.g, given a reverberant sound produced in a large hall, our model attempts to remove the reverb effect to predict the anechoic or clean audio.

Audio-only dereverberation is a well-studied problem with various systems achieving encouraging results [53, 34, 99, 91, 90]. In contrast, using the visual stream as an additional cue to solve this task is a particularly understudied problem. We attribute the lack of research in this space to the scarcity of datasets. Most open-source datasets, both real and synthetic, contain only room impulse responses (RIRs) with no information about their source of origin [77, 83]. Note that obtaining such RIRs can be challenging as doing so requires access to the physical environment, thereby limiting their applicability. However, in real-world settings, reverberant audio is naturally accompanied by a visual stream; video conferencing, augmented reality (AR),

*Equal contribution.

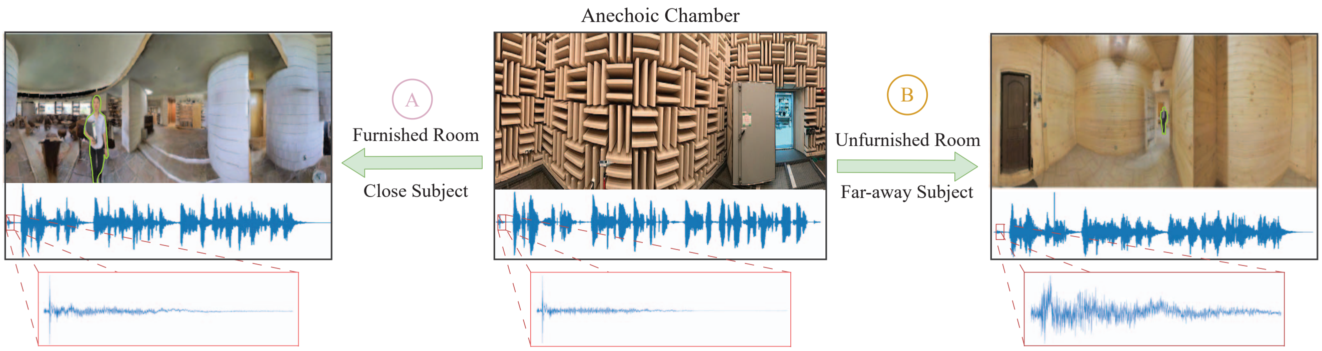

Figure 2: Reverberation is a function of the speaker's relative position and the surrounding environment. The visual signals present critical details that determine the nature of the distortion. E.g, Ⓐ in a relatively small furnished room when the speaker is nearby, reverb is less evident, whereas for Ⓑ in a large hallway (especially when the speaker is far away) the reverb effect is very strong. The audio waveform illustrates the nature of reverberation, with the magnified section clearly depicting a stronger reverberation effect in case Ⓑ over Ⓐ.

and web video indexing are some examples.

Recently, audio-visual speech enhancement methods [95, 79, 12, 48] have shown significant improvements over the audio-only speech enhancement approaches. These tasks benefit from the presence of sound-producing objects in the visual scene, which allows the model to effectively utilize these strong stimuli for accomplishing the task. Many of these approaches track the lip movements of the speaker to separate the noise from the voice components in degraded speech which builds on the assumption that a speaker is always close to and facing the camera. These assumptions might not always hold in our case as the scope of the problem under consideration (mid/far-field) makes it difficult to obtain such cues. Thus, in a real-world setting, the available cue for audio-visual dereverberation is a panoramic view of the environment with or without a speaker in the field of view. Effectively utilizing visual cues in order to perform audio-visual dereverberation would require the model to understand the room's implicit geometric and material properties, which poses its own challenges.

Our Contributions: We propose AdVerb, comprising a modified conformer block [24] with specially designed positional encoding to learn audio-visual dereverberation. The network takes corrupted audio and the corresponding visual image[1] of the surrounding environment (from where the RIR is obtained) as input to perform this task (Fig. 1). Our approach employs a *novel geometry-aware module* with cross-modal attention between the audio and visual modalities to generate a *complex ideal ratio mask*, which is applied to the reverberant spectrogram to obtain the estimated clean spectrogram. This conformer block consists of a *modified (Shifted) Window Block* [44] and *Panoptic Blocks* to combine local and global geometry relations. We discuss key motivations behind our approach in Section 4. To learn audio-visual dereverberation, AdVerb solves two objectives, *Spectrogram Prediction Loss* and *Acoustic Token Matching Loss*, which makes the output audio retain phonetic and prosodic properties. To summarize, our main contributions are as follows:

(1) We propose AdVerb, *a novel cross-modal framework* for dereverberating audio by exploiting complementary low-level visual cues and specially designed relative position embedding.

(2) To this end, AdVerb employs *a novel geometry-aware conformer network* to capture 3D spatial semantic information to equip the network with salient vision cues through (Shifted) Window Blocks and Panoptic Blocks.

(3) Our architecture involves the prediction of *complex ideal ratio mask* and simultaneous optimization of two objective functions to estimate the dereverbed speech.

(4) On objective evaluation our approach significantly outperforms traditional audio-only and audio-visual [12] baselines with a relative improvement in the range 18% - 82% on three downstream tasks: speech enhancement, speech recognition, and speaker verification, when evaluated on the LibriSpeech test-clean set on all difficulty levels. It also achieves highly satisfactory RT60 error scores on the AVSpeech dataset.

(5) User study analysis reveals our method outperforms prior approaches on perceptual audio quality assessment.

2. Related Works

Audio Dereverberation: In communication and speech processing applications, reverberation can reduce intelligibility and weaken a dry audio signal [53, 34, 99, 91, 90]. Lately, there has been a paradigm shift from using the traditional signal processing-based methods to neural networks and, subsequently, deep learning-based methods for dereverberation. Kinoshita *et al.* [35] presents a deep neural network to estimate the power spectrum of the target sig-

[1] We use panoramic images to train; inference can be done on both panoramic and non-panoramic images.

nal for weighted prediction error. Extending this, Wang *et al.* [87] deploy a CNN-based model to separate the real and imaginary parts of clean speech. Typically, there are two prominent ways of training such models: through supervised learning [92, 45] or through adversarial networks (GANs) [73, 75]. Audio reverberation in nature is heavily influenced by room acoustics [43]. We find studies in the literature that try to capture room-specific information for finer modeling of acoustic environments [72, 23]. Another line of work [80, 41] attempts to extract visual features of target lip movements. Work from Chen *et al.* [12] is most similar in spirit to our proposed approach. These studies motivate us to pursue audio-visual dereverberation by leveraging room-aware geometric cues. Our framework exploits panoramic image features and is applicable even for out-of-view speaker cases.

Room Impulse Response and Geometry Awareness: For a given environment, the amount of reverberation in the speech signal is mathematically described using a function known as room impulse response (RIR). RIR generators are used to simulate large-scale speech training data [63, 64]. While [28, 71, 78] engage dedicated in-room amenities to estimate this function, another line of research [5, 50, 9, 81, 62] choose to produce RIRs synthetically. These works [37, 69] estimate RIRs from an RGB and depth image. One downside of these approaches is that they require access to paired image and impulse response data. In contrast, some prior methods [32, 33, 51] for generating RIR operate by using images taken at arbitrary distances from the point of audio capture.

Video streams, by nature, capture the natural association between visual and audio modalities. Wang *et al.* [85] propose a geometry-aware approach for room layout estimation by horizon depth, which is only effective in the horizontal direction. Hu *et al.* [31] and Eder *et al.* [16] introduce gradient of depth and plane aware loss, respectively for improved depth estimation of panoramic images. These works inspire us to leverage room geometry to model this problem.

Audio-Visual Learning: Cross-modal learning powered by large-scale video datasets has been pushing boundaries in applications like audio-visual sound separation [98, 97, 22, 19, 93], audio-visual speech enhancement [1, 2, 27, 94], active speaker detection [3, 4, 82, 67], talking head generation [86, 13, 60], embodied AI for audio-visual navigation [7, 10, 47, 96], etc. In addition, many recent works have utilized paired audio-visual data for representation learning. Owens *et al.* [56] learned visual representations for materials from impact sounds. Another line of work learns features, scene structure, and geometric properties [14, 57, 20] respectively from audio. However, our approach to estimating the geometric cues for audio-visual dereverberation is complementary to these methods.

3. Problem Formulation

We propose a novel framework that takes reverberant speech $\mathcal{A}_r$ and the corresponding environment panoramic image $\mathcal{V}_r$ as input and outputs estimated clean audio $\mathcal{A}_e$. Both $\mathcal{V}_r$ and $\mathcal{A}_r$ are captured from the listener position focusing on the environment surrounding the speaker (considers far, mid, and near field examples). The reverberation effects can be described using a transfer function known as room impulse response $\mathcal{R}(t)$. $\mathcal{A}_r$ can be obtained by convolving clean speech $\mathcal{A}_s$ with $\mathcal{R}(t)$ (Equation 1) [54]. Here, $\mathcal{R}$ depends on the listener and speaker positions, room geometry, and acoustic material characteristics.

$$\mathcal{A}_r(t) = \mathcal{A}_s(t) * \mathcal{R}(t) \tag{1}$$

4. Our Approach: AdVerb

Fig. 3 depicts a pictorial representation of AdVerb, our proposed audio-visual dereverberation model. Our primary objective is to learn the inverse function given a reverberant audio signal by exploiting the audio and visual cues. Elaborations on the individual components are as follows:

4.1. Feature Encoder

Vision Encoder: To encode geometric layout-specific visual features $\mathcal{E}_{\mathcal{V}}(\cdot)$, we use HorizonNet [76], which is based on ResNet-50 [26] backbone. HorizonNet takes a panoramic image of the surroundings as input $\mathcal{V}$, with dimensions 512 × 1024 × 3. The output is a 2D feature map of 4 different scales. For each feature map, the height is down-sampled, and the width $\mathcal{N}$ is up-sampled to obtain 1D spatial property-infused feature sequences with dimension $\mathbb{R}^{\mathcal{D}/4}$ and connect all the feature maps to obtain $\mathbb{R}^{\mathcal{D}}$, where $\mathcal{D}$ is 1024 in our case.

Audio Encoder: For audio features, we employ Short-Time Fourier Transform (STFT) $\mathcal{E}_{\mathcal{A}}(\cdot)$ on the reverberant 1D audio $\mathcal{A}$ to obtain a 2D spectrogram $\mathcal{A}(t, f)$, where t and f index time and frequency, respectively. In contrast to prior work, which learns a convolution network for this transformation [8], we employ STFT with the motivation of using complex masks for learning dereverberation. We calculate STFT with a window of size 400 samples or 25 milliseconds, a hop length of 160 samples or 10 milliseconds, and a 512-point FFT. All our audios are sampled at 16kHz.

4.2. Complex Ideal Ratio Masks

Intuition Behind Masks: We hypothesize that learning to generate clean anechoic speech in an end-to-end fashion might not be effective owing to the nature of the task. Traditionally, the input audio learns to align to the visual cues, which proves to be effective for the visual acoustic matching [8] task. Similarly, synthesizing speech directly has also seen huge success in audio-visual speech enhancement, and

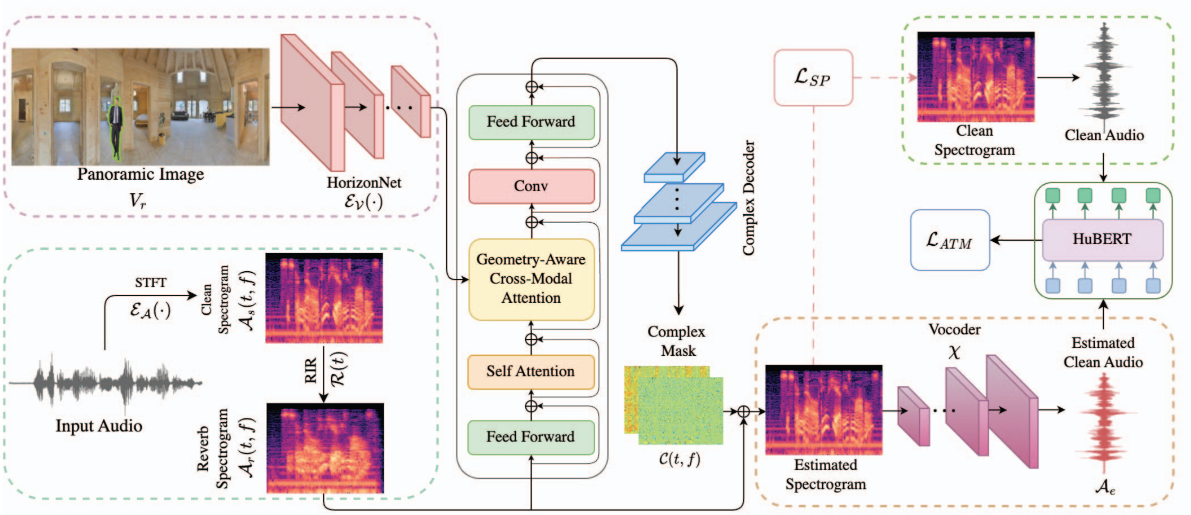

Figure 3: Overview of **AdVerb**. AdVerb estimates clean source audio from a reverberant speech signal leveraging two primary components: ① The visual stream processing path comprises a HorizonNet-based backbone $\mathcal{E}_{\mathcal{V}}(\cdot)$ to obtain 1D feature sequences, which are subsequently passed to the cross-modal geometry-aware attention subnetwork. ② The audio processing module applies STFT $\mathcal{E}_{\mathcal{A}}(\cdot)$ to get 2D spectrograms which are fed to the cross-modal encoder. The cross-attention subnetwork powered by geometry-aware (Shifted) Window Blocks, Panoptic Blocks, and Relative Position Embedding generates a complex ideal ratio mask.

separation [25, 94], where the visual cues have a high correlation with the contents of the speech, e.g., lip movements. However, the dereverberation algorithm tries to learn an inverse function making the task intrinsically challenging. From Fig. 2 it is evident that the same speech content incurs heavy reverberation artifacts when the speaker is far away in a reverberant environment (Ⓑ) while the corruption of the speech signal is not significant when the speaker is closer in a relatively less reverberant environment (Ⓐ). Thus, we hypothesize such visual cues can be instead used to learn a mask that, when applied to the reverberated speech, suppresses reverberation effects. STFT mask prediction has seen success in the past in a variety of tasks, including source separation [98, 22], speech enhancement [89], etc.

Complex Ideal Ratio Mask Construction: A complex ideal ratio mask (cIRM) [88] is an extension of the conventional ideal ratio mask to process the real and imaginary components of an audio signal separately. The product of cIRM and reverberant speech results in estimated clean speech. It is calculated in the time-frequency (T-F) domain, and thus learning to generate cIRM enhances both the magnitude and phase of reverberant speech, improving overall perceptual speech quality. Given the STFT of reverberant speech, $\mathcal{A}_r(t, f)$, and the cIRM, $\mathcal{C}(t, f)$, clean speech, $\mathcal{A}_s(t, f)$, is computed as follows:

$$\mathcal{A}_s(t, f) = \mathcal{C}(t, f) * \mathcal{A}_r(t, f) \tag{2}$$

where t and f are index time and frequency respectively. Since the STFT is complex, $*$ indicates complex multiplication. $\mathcal{C}(t, f)$ is computed by dividing the STFT of direct speech, by the STFT of reverberant speech:

$$\mathcal{A}_s^r(t, f) + j\mathcal{A}_s^i(t, f) = \mathcal{C}(t, f) * \mathcal{A}_r^r(t, f) + j\mathcal{A}_r^i(t, f) \tag{3}$$

$$\begin{aligned}\mathcal{C}(t, f) &= \frac{\mathcal{A}_s^r(t, f) + j\mathcal{A}_s^i(t, f)}{\mathcal{A}_r^r(t, f) + j\mathcal{A}_r^i(t, f)} * \frac{\mathcal{A}_r^r(t, f) - j\mathcal{A}_r^i(t, f)}{\mathcal{A}_r^r(t, f) - j\mathcal{A}_r^i(t, f)} \\ &= \frac{\mathcal{A}_s^r(t, f)\mathcal{A}_r^r(t, f) + \mathcal{A}_s^i(t, f)\mathcal{A}_r^i(t, f)}{\mathcal{A}_r^{r2}(t, f) - \mathcal{A}_r^{i2}(t, f)} \\ &+ j\frac{\mathcal{A}_s^i(t, f)\mathcal{A}_r^r(t, f) - \mathcal{A}_s^r(t, f)\mathcal{A}_r^i(t, f)}{\mathcal{A}_r^{r2}(t, f) - \mathcal{A}_r^{i2}(t, f)}\end{aligned} \tag{4}$$

4.3. Cross-Modal Geometry-Aware Conformer

Overview: In this module, we aim to learn cross-modal attention between audio and visual features, which enables incorporating fine-grained interactions between them in a geometry-aware fashion. The visual and audio feature maps obtained from corresponding encoders are used as inputs here. The sequence of features from each time step represents a part of the input stream. These sequences are then passed to the conformer-based [24] cross-modal encoder. For the audio stream, we obtain a complex ideal ratio mask by employing a complex-valued self-attention block. This is then fed into the geometry-aware cross-modal self-attention (GCA) block for audio-visual modeling. We specially design a relative position embedding to encode position-specific information. Finally, the learned representations are passed through a complex-valued decoder to generate the predicted cIRM. We next describe these components in detail.

Complex Self-Attention: The self-attention mechanism [84] transforms a sequence into a set of vectors, where each vector is computed as the weighted sum of all other vectors. Here the weights are determined by a learnable function based on the similarities between the input and output. The primary difference between conventional and complex self-attention (CSA) is that the latter operates on complex-valued representations and calculates self-attention separately on the real and imaginary parts. We use CSA instead of vanilla SA layers because of the nature of our input spectrogram. For our implementation, we use Complex-Valued Time-Frequency SA (CTSA) proposed in [39], which improves over CSA by accurately modeling inter-dependencies between real and imaginary components of the encoded audio features.

Geometry-Aware Cross-Modal Encoder: A wealth of studies [8, 14] establish that direct concatenation of cross-modal features [21, 55] might lead to suboptimal performance. A key observation here is such techniques don't seem suitable in our case, as our application demands more robust reasoning on how different regions of the 3D space contribute to the acoustics differently. For instance, if the sound originates from inside a highly absorptive chamber, less reverberation will be noticeable. In contrast, in the case of a reflective surface, an extended reverberation effect will persist. Hence, it is imperative to attend to image patches to study how they contribute to the overall acoustics.

Inspired by Swin-Transformer [44], our novel GCA module exploits window partitioning for robust spatial modeling ability. However, we observe that using window partition alone limits the conception of the holistic representation of the visual scene. As a result, we equip our Transformer module with (Shifted) Window Blocks and Panoptic Blocks to combine the local and global geometry relations efficiently. Each loop contains four consecutive blocks: Window Block, Panoptic Block, and Shifted Window Block, followed by another Panoptic Block. As shown in Fig.4, the individual blocks follow the Transformer [84] architecture, with modifications done before and after the multi-head attention layer. Note that the dimension of the sequence and corresponding positions of tokens don't get altered in any block.

In Window Block, we use a patchwise partition on the input feature sequence to obtain $\frac{\mathcal{N}}{\mathcal{N}_w}$ window feature sequences $\mathbb{R}^{\mathcal{N}_w \times \mathcal{D}}$ where $\mathcal{N}_w$ is the window length and is set to 16 in our case. The window partition captures local geometry relations and facilitates the calculation of self-attention by reducing computation while calculating attention. Subsequently, window features are combined after the multi-head attention, as depicted in Fig. 4 Ⓐ.

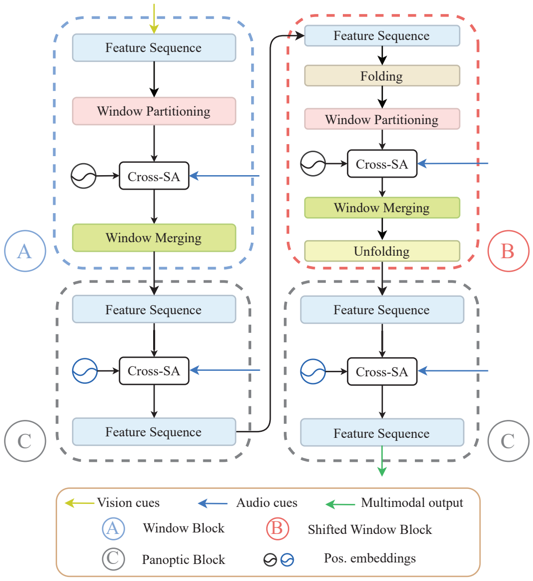

Figure 4: Overview of the Geometry-Aware Cross-Modal Attention block. Window and Panoptic Relative Position Embedding (RPE) are fused into Cross-modal Self-Attention (CSA) blocks. In Window Block Ⓐ, partitioning and merging of windows before and after CSA. In Ⓑ, Folding and Unfolding of sequence features before and after CSA, respectively. Ⓒ integrates another RPE to CSA.

Inspired by [44], we deploy Shifted Window Block, which connects adjacent windows to facilitate the exchange of information flow between nearby patches. Here a fold and unfold operation is performed by a fraction of $\frac{\mathcal{N}_w}{2}$ to retain the original positions of the feature sequence even after merging: refer to Fig.4 Ⓑ. Finally, the Panoptic Block follows the native Transformer [84] encoder to enhance holistic geometry-aware relations of the visual scene (Fig. 4 Ⓒ).

To model the natural association between visual and audio streams by ensuring cross-modal information flow, we employ the conformer variant [24] of encoder blocks, which adjoins a convolution layer inside the block for modeling local interactions of audio features. Building on this, we insert one cross-modal attention layer ξ_{cm} after the first feed-forward layer, described as follows:

$$\xi_{cm}\left(\mathcal{A}_i, \mathcal{V}_i\right) = \text{softmax}\left(\frac{\mathcal{A}_i^Q \mathcal{V}_i^{K^T}}{\sqrt{\mathcal{S}}}\right)\mathcal{V}_i^V. \qquad (5)$$

where superscripts K, Q, and V indicate Key, Query, and Value, respectively. Here, we compute the attention scores between the visual ($\mathcal{V}_i$) and the audio ($\mathcal{A}_i$) sequences by dot-product. This is followed by softmax normalization and scaling by $\frac{1}{\sqrt{\mathcal{S}}}$, which is then used to factor $\mathcal{V}_i$. The key observation here is that cross-modal attention thus designed enables the model to attend to spatial regions in the visual stream and comprehend its acoustic nature.

Position Embedding: The conventional attention module is found to be insensitive to the positions of the tokens producing suboptimal results. To this end, we introduce specially designed relative position embedding (RPE) [61] to strengthen its spatial identification ability. We denote the input sequence of multi-head cross-modal self-attention as $\mathcal{X} = \{x_i\}_{i=1}^{\mathcal{Z}}$, where $\mathcal{Z}$ is the sequence length and $x_i \in \mathbb{R}^{\mathcal{D}}$. A bias matrix $\mathcal{B} \in \mathbb{R}^{\mathcal{Z}\times\mathcal{Z}}$ is added to Scaled Query-Key product [84]:

$$\begin{aligned} \alpha_{ij} &= \frac{1}{\sqrt{\mathcal{D}}}\left(x_i \mathcal{W}^Q\right)\left(x_j \mathcal{W}^K\right)^T + \mathcal{B}_{ij}, \\ \text{Attention}\left(\mathcal{X}\right) &= \text{Softmax}(\alpha)\left(\mathcal{X}\mathcal{W}^V\right), \end{aligned} \tag{6}$$

where $\mathcal{W}^Q, \mathcal{W}^K, \mathcal{W}^V \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ are learnable project matrices and each bias $\mathcal{B}_{ij}$ comes from a learnable scalar table. In (Shifted) Window Block, $\mathcal{Z} = \mathcal{N}_w$. We denote the learnable scalar table as $\{b_k\}_{k=-\mathcal{N}_w+1}^{\mathcal{N}_w-1}$, and $\mathcal{B}_{ij}$ corresponds to b_{j-i}. This Patch RPE is fed into multi-head attention.

For Panoptic Block, we consider $\mathcal{Z} = \mathcal{N}$. Here we propose a symmetric representation of only distance and denote the learnable scalar table as $\{b_k\}_{k=0}^{n}$, where $n = \frac{\mathcal{N}}{2}$. When $|j-i| \leq \frac{\mathcal{N}}{2}$, B_{ij} corresponds to $b_{|j-i|}$, otherwise $\mathcal{B}_{ij}$ corresponds to $b_{\mathcal{N}-|j-i|}$.

4.4. Complex Mask Decoder

The complex mask decoder takes input from the conformer and generates a complex ideal ratio mask $\mathcal{C}$. The decoder comprises a complex-valued ReLU activation function followed by a complex-valued convolutional layer, a self-attention module, a dense block, and finally a normalization layer.

4.5. Vocoder

After generating the complex ideal ratio mask $\mathcal{C}$, we decode the output spectrogram $\mathcal{G}$ by performing the complex multiplication operation between $\mathcal{C}$ and $\mathcal{G}$. Next, we use a pre-trained vocoder χ [73] to convert the spectral representation of the audio signal to the waveform. We perform this step specifically to calculate the SSL-based HuBERT Loss, which we describe later.

4.6. Model Optimization

Spectrogram Prediction Loss: The first of the two objective functions we use for model optimization is the Spectrogram Prediction Loss (SP). Learning to reconstruct the clean spectrogram is a common optimization methodology used in speech enhancement and dereverberation [10, 40]. It computes the L_2 norm between the spectrogram predicted by our network Θ and the ground truth clean spectrogram $\mathcal{A}_s$. It is defined as:

$$\mathcal{L}_{SP} = \mathbb{E}_{(\mathcal{A}_r,\mathcal{V})\sim\mathcal{U}} \left\|\phi(\Theta(\mathcal{A}_r,\mathcal{V})) - \phi(\mathcal{A}_s)\right\|_2, \tag{7}$$

where $\mathcal{A}_r$ is the reverberant audio and $\mathcal{V}$ is the corresponding panoramic image in some distribution $\mathcal{U}$. ϕ is the function that transforms the speech waveform to the corresponding spectrogram representation.

Acoustic Token Matching Loss: Inspired by the recent success of self-supervised speech representation learning [49], we introduce Acoustic Token Matching Loss (ATM). The traditional MSE loss ignores the inherent speech characteristics, like phonetic and prosodic properties, that are essential for learning and reconstructing speech information [29]. Speech representations learned with SSL effectively encode such characteristics in their latent representations [59]. Thus, we propose a simple yet effective method to enforce the output speech from AdVerb to encode such information by solving the Acoustic Token Matching Loss . To calculate ATM loss, we first generate latent representations $\tilde{\mathcal{H}} \in \mathbb{R}^{\mathcal{J}\times d}$ from the clean waveform $\mathcal{A}_s$ with a pre-trained HuBERT [30] model $\mathrm{e}(\cdot)$, where d is the HuBERT embedding dimension and $\mathcal{J}$ is the sequence length. Next, we cluster these latent representations using *K-means* to generate a sequence of pseudo-labels $\mathcal{P} = \{p_t\}_{t=1}^{\mathcal{J}}$. These pseudo-labels are representative of the latent space in our speech input. Finally, we predict these pseudo-labels from latent representations of $\mathcal{A}_e$ (estimated output audio from AdVerb) obtained after passing it through HuBERT. The ATM Loss function can be expressed as follows:

$$\mathcal{L}_{ATM}(e;\mathcal{A}_s,\mathcal{A}_e) = \sum_{t\in\mathcal{J}} \log p_f\left(p_t \mid \tilde{\mathcal{H}}, t\right) \tag{8}$$

where p_f is the distribution over the target indices at each timestep t. Finally, we optimize our model with a total loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \lambda\mathcal{L}_{SP} + \mu\mathcal{L}_{ATM} \tag{9}$$

where $\lambda, \mu \in \mathbb{R}$ are hyper-parameters to balance the contribution of each loss component.

5. Experiments and Results

For a fair assessment of our model, we evaluate our model through speech dereverberation on three downstream tasks: speech enhancement (SE), automatic speech recognition (ASR), and speaker verification (SV), respectively. The environments are taken from Matterport3D [6], with speech samples from the LibriSpeech dataset [58].

5.1. Dataset

SoundSpaces-Speech Dataset: We use the SoundSpaces-Speech dataset proposed in [12] for our experiments. It comes with paired anechoic and reverberant audio with

	Method	Speech Enhancement (SE)†	Speech Recognition (SR)†		Speaker Verification (SV)†		RTE* ↓
		PESQ ↑	WER (%) ↓	WER-FT (%) ↓	EER (%) ↓	EER-FT (%) ↓	(in sec)
	Anechoic (Upper bound)	4.72	2.89	2.33	1.53	1.57	-
	Reverberant	1.49	8.20	4.44	4.51	4.88	0.382
	MetricGAN+ [18]‡	2.45 (+64%)	7.48 (+9%)	4.86 (-9%)	4.67 (-4%)	2.85 (+42%)	0.187
	HiFi-GAN [38]‡	1.83 (+23%)	9.31 (-14%)	5.59 (-26%)	4.32 (+4%)	2.49 (+49%)	0.196
	WPE [52]‡	1.63 (+9%)	8.43 (-3%)	4.30 (+3%)	5.90 (-31%)	4.11 (+16%)	0.173
	SkipConvGAN [40]‡	2.10 (+41%)	7.22 (+12%)	4.17 (+6%)	4.86 (-8%)	3.98 (+18%)	0.119
	VIDA [12]	2.37 (+59%)	4.44 (+46%)	3.66 (+18%)	3.97 (+12%)	2.40 (+51%)	0.155
ABLATION	AdVerb w/o Image	2.31 (+55%)	3.92 (+52%)	3.41 (+23%)	3.67 (+19%)	2.19 (+55%)	0.119
	AdVerb w/ Random Image	2.54 (+70%)	4.12 (+50%)	3.62 (+18%)	3.76 (+17%)	2.26 (+54%)	0.110
	AdVerb w/o ATM Loss	2.89 (+94%)	4.67 (+43%)	3.66 (+18%)	3.17 (+30%)	2.07 (+58%)	0.117
	AdVerb w/o Complex SA	2.91 (+95%)	3.63 (+56%)	2.98 (+33%)	3.21 (+29%)	2.10 (+57%)	0.117
	AdVerb w/o Geometry Aware Block	2.30 (+54%)	4.01 (+51%)	3.12 (+30%)	3.68 (+18%)	2.12 (+57%)	0.113
	AdVerb w/o RPE	2.79 (+87%)	3.54 (+57%)	3.01 (+32%)	3.17 (+30%)	2.11 (+57%)	0.107
	AdVerb w/o Window Block	2.81 (+89%)	3.61 (+56%)	2.99 (+33%)	3.14 (+30%)	2.12 (+57%)	0.108
	AdVerb w/o Panoptic Block	2.92 (+96%)	3.59 (+56%)	2.92 (+34%)	3.29 (+27%)	2.01 (+59%)	0.107
	AdVerb (ours)	**2.96 (+98%)**	**3.54 (+57%)**	**2.91 (+34%)**	**3.11 (+31%)**	**1.98 (+59%)**	**0.101**

Table 1: Comparison of AdVerb with various baselines on multiple spoken language processing tasks based on the LibriSpeech test-clean set (marked with †) and on sim-to-real transfer based on the AVSpeech dataset (marked with *). "Anechoic (Upper bound)" refers to clean speech, while "Reverberant" refers to clean speech convolved with RIR. WER-FT and EER-FT denote evaluations when the SR and SV models are finetuned with the audio-enhanced data. Numbers in parentheses denote the relative improvement compared to Reverberant. Methods marked with ‡ are audio-only.

camera views from 82 Matterport3D [6] environment convolved with speech clips from LibriSpeech [58] samples. SoundSpaces [9] provide precomputed RIRs $\mathcal{R}(t)$, which are convolved with speech waveforms to obtain reverberant signal $\mathcal{A}_r(t)$ for a total of 49,430/2,700/2,600 train/validation/test samples, respectively.

Acoustic AVSpeech Web Videos: Web videos offer natural supervision between visuals and acoustics in abundance. To be consistent with prior work, we use the collection from [8], which is a subset of the AVSpeech[17] dataset. The clip durations range between 3-10 seconds with a visible human subject in each video frame. To evaluate our model on real-world data in addition to synthetic data, we use these 3K samples only for testing purposes.

Evaluation Tasks And Metrics: We follow the standard practice of reporting Perceptual Evaluation of Speech Quality (PESQ) [66], Word Error Rate (WER), and Equal Error Rate (EER) to compare our method with the baselines for the three tasks. Following [12], we employ the pre-trained models from the SpeechBrain [65] for ASR and SV tasks. These models were evaluated on the LibriSpeech test-clean set. SV evaluation was done on a set of 80K randomly sampled utterance pairs from the test-clean set.

5.2. Baselines

WPE [52]: A statistical method that estimates an inverse system for late reverberation. It deploys variance normalization to improve dereverberation results with relatively short observations.

MetricGan+ [18]: We use the implementation by [65] for benchmarking. As presented by the authors, it can be used to optimize different metrics. We optimize PESQ to report values from the best model for individual downstream tasks.

SkipConvGAN [40]: A recent model where the generator network estimates a complex time-frequency mask and the discriminator aids in driving the generator to restore the lost formant structure. The model achieves SOTA results on the dereverberation task.

HiFi-GAN [38]: A GAN-based high-fidelity speech synthesis system that shows satisfactory results on speech dereverberation. It models periodic patterns of audio to enhance sample quality.

VIDA [12]: An end-to-end vision backed speech dereverberation framework. It combines RGB-D image information to estimate clean speech.

5.3. Results

Evaluation Setup on LibriSpeech: We compare model performance on three speech tasks: Speech Enhancement (SE), Automatic Speech Recognition (ASR), and Speaker Verification (SV). To evaluate our trained models, we use the dereverbed version of the test-clean set split of the LibriSpeech dataset. Similar to [12], for SR and SV, we either use pre-trained models from SpeechBrain [65] or fine-tune a model from scratch using dereverbed LibriSpeech train-clean-360 split.

Quantitative Analysis on LibriSpeech: Table 1 compares the performance of AdVerb with the baselines. Experimental results show AdVerb outperforms all audio-only baselines by a significant margin on all three tasks. We achieved relative improvements of 41%, 51%, and 36% over the best audio-only baseline, SkipConvGAN, on SE, SR, and SV, respectively, in terms of relative gain from reverberant speech. AdVerb also outperforms VIDA by 25%, 20%, and 22% on SE, SR, and SV, respectively, which shows the superiority of AdVerb in audio-visual dereverberation tasks.

Quantitative Analysis on AVSpeech: To examine the robustness of our proposed approach in real-world settings, we evaluate our model on an in-the-wild AVSpeech audio-visual dataset collected from YouTube [17]. The AVSpeech dataset has non-panoramic images; therefore the field-of-view is limited in the test dataset, and the performance of our network trained on panoramic images is not optimal. In the absence of the ground truth clean speech, we use the average reverberation time (RT) of the dereverberated speech signal for evaluation. RT is the time taken to decay the sound pressure in RIR by 60 decibels. We can estimate RT from the reverberant speech signal [8]. According to Equation 1, in clean speech, RIR will be an impulse response ($\delta(t)$) and ≈ 0. The dereverberated speech with the least amount of reverberation will have the lowest RT. Therefore, reverberation time error (RTE) is the average RT of the dereverberated test speech samples. From Table 1, we can see that AdVerb reports the lowest RTE.

Ablation Study: To show the importance of the individual components in AdVerb, we perform an extensive ablation study shown in Table 1. Note that AdVerb sees the steepest fall in performance across tasks when trained and evaluated w/o images, i.e., in an audio-only setup. In this setup, our GCA block is replaced with a simple uni-modal self-attention block. There is also a considerable drop in performance across all tasks w/o the geometry-aware module, thus underlining the importance of this block. In this setup, our GCA block is replaced with a simple cross-modal self-attention block with queries as audio cues and keys and values as visual cues. We carry out further ablations to study the contributions of the individual components of the cross-modal geometry-aware attention block. Interestingly, the drop in performance when removing the individual elements is much less than the entire GCA block. This underlines that these components combine to have a telling impact on the overall setup. Finally, we show that ATM loss improves AdVerb's SR performance by a significant margin. Refer to the supplementary for further ablations.

More Comparison Against Audio-only Methods: Table 2 demonstrates the performance of our model against SOTA audio-only methods. AdVerb outperforms existing audio-only methods and sets new benchmarks.

Results on Noisy Dataset: To evaluate the robustness of the proposed AdVerb model to outdoor unwanted noise, we add ambient sounds from urban environments to the LibriSpeech test-clean dataset using the MUSAN dataset [70]. Following [10], we maintain an SNR of 20 for our mixture. Table 3 compares the performance of AdVerb on three downstream speech-based tasks. All experiments were done for the non-fine-tuned version of our experimental setup, where a pre-trained model was used from SpeechBrain. Though we see a drop in performance compared to the noise-free dataset, AdVerb outperforms all our baselines and maintains similar margins compared to the original noise-free dataset.

	SE	SR		SV		RTE ↓
Method	**PESQ ↑**	**WER (%) ↓**	**WER-FT (%) ↓**	**EER (%) ↓**	**EER-FT (%) ↓**	**(in sec)**
Reverberant	1.49	8.20	4.44	4.51	4.88	0.382
DEMUCS [15]	2.17	7.97 (+2.8%)	5.20 (-17%)	3.82 (+15%)	2.96 (+39%)	0.129
VoiceFixer [42]	2.41	5.66 (+31%)	4.19 (+5%)	3.76 (+16%)	2.79 (+42%)	0.121
H-GAN [74]	1.94	8.14 (+1%)	5.01 (-12%)	4.22 (+6%)	3.13 (+35%)	0.196
Kotha. *et al.* [39]	2.54	5.32 (+35%)	4.13 (+6%)	3.71 (+17%)	2.68 (+44%)	0.124
AdVerb	**2.96**	**3.54 (+57%)**	**2.91 (+34%)**	**3.11 (+31%)**	**1.98 (+59%)**	**0.101**

Table 2: Comparison of AdVerb with more audio-only approaches. AdVerb results in a relative gain of **14%-56%**. Percentages in bracket represent an improvement on reverberant audio.

	SE	SR	SV
Method	**PESQ ↑**	**WER ↓**	**EER ↓**
Anechoic (Upper bound)	4.72	2.89	1.53
Reverberant	1.57	11.45	4.76
MetricGAN+	2.29	8.92	4.89
HiFi-GAN	1.95	10.55	4.73
WPE	1.88	9.10	5.11
SkipConvGAN	2.06	7.28	4.94
VIDA	2.14	4.97	4.01
AdVerb (Ours)	**2.52**	**4.20**	**3.46**

Table 3: Result comparison of AdVerb with baseline methods on noise added dataset splits for 3 speech tasks.

Ablation on Noisy Dataset: Table 4 illustrates the results of the ablation study on the noisy LibriSpeech dataset. The noise addition process is the same as before.

	SE	SR	SV
Method	**PESQ ↑**	**WER ↓**	**EER ↓**
w/o Image	2.03	4.68	3.81
w/o ATML	2.28	5.10	3.87
w/o Geom. aware	2.29	4.99	3.64
w/o Window block	2.34	4.43	3.43
w/o Panoptic block	2.39	4.37	3.51

Table 4: Ablation on LibriSpeech noisy data. AdVerb performs considerably well on noisy data with the individual modules contributing to the overall gain.

Analysis On Visual Features: To underline the importance of the visual cues, we show the activations of the network using Grad-CAM[68] in Fig. 5. Note that the network attends to the sides of the hallway or empty regions with al-

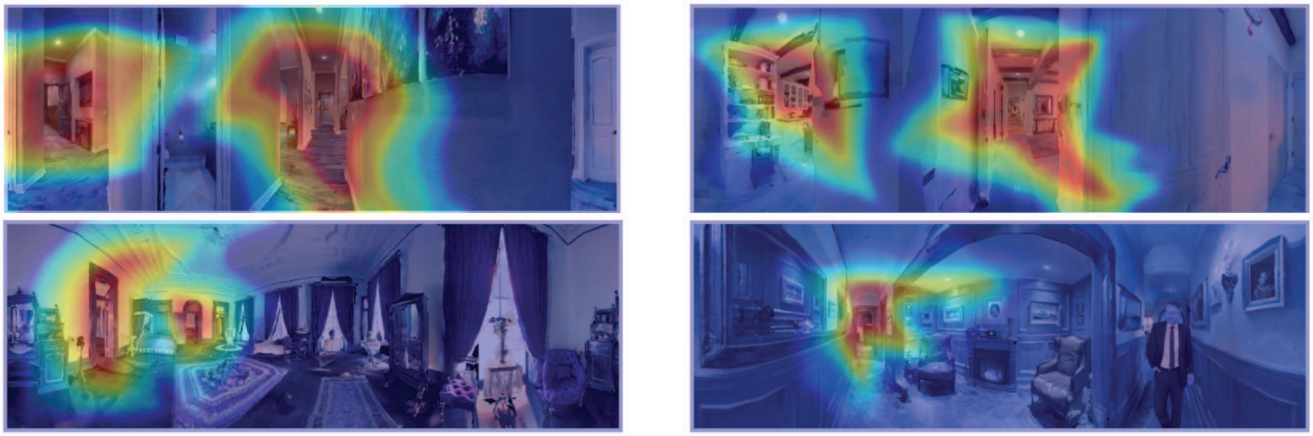

Figure 5: Grad-CAM visualization of activated regions. Our model attends to regions that cause heavy reverberation effects.

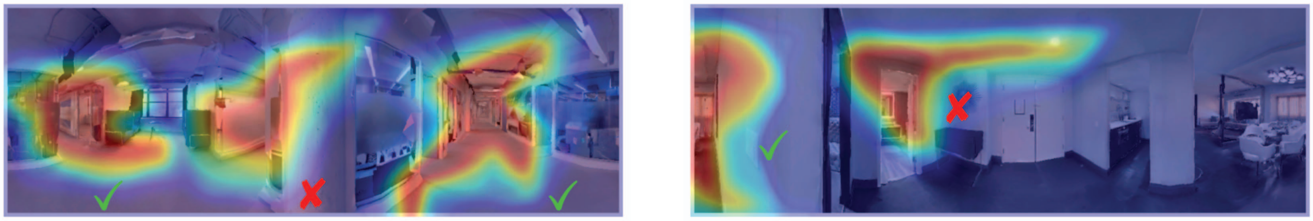

Figure 6: Some failure cases. The ✓ denotes the regions with correct activation while ✗ spurious detections.

Baseline	SoundSpaces(in %)	AVSpeech(in %)
Method	(A% / B% / C%)	(A% / B% / C%)
Clean Speech	**61.3** / 8.1 / 30.6	– / – / –
VIDA[12]	16.5 / 6.5 / **77.0**	13.5 / 0.0 / **86.5**
WPE[52]	8.8 / 3.5 / **87.7**	3.7 / 7.4 / **88.9**
SCGAN[40]	9.2 / 0.0 / **90.8**	0.0 / 8.0 / **92.0**

Table 5: User study results. A% of participants find the baseline audio samples better, B% have no preference, and C% prefer AdVerb.

most or no sound absorbers which lead to longer reverberation effects. Fig. 6 demonstrates some cases where our model attends to spurious regions.

5.4. User Study For Subjective Evaluation

In addition to objective metric evaluation, we perform a subjective human listening study on a synthetic (generated using SoundSpaces) and an in-the-wild (AVSpeech) dataset over Amazon MTurk. We believe this can be a good measure to understand how realistic and aesthetically pleasing the output produced by our model is. Moreover, through this, we try to understand other aural artifacts not captured in an objective measure like PESQ. In our study, a total of 89 participants were presented with 8 sets of samples containing the reverberant speech, clean speech (not present for AVSpeech), and estimated dereverberant speech. Table 5 demonstrates that users find samples generated by our method better than the three other baselines VIDA [12], WPE [52] and SkipConvGAN [40], in both cases.

6. Conclusions and Future Works

In this paper, we present a novel audio-visual dereverberation framework. To this end, we introduce the GCA module with a specially designed position embedding scheme to capture the local and global spatial relations of the 3D environment. The experimental analysis demonstrates how modeling the visual information efficiently can lead to improved performance of such a system. We believe our work will encourage further research in this space. One limitation of our approach is that the efficacy of the method drops for non-panoramic images. Future work can aim towards finding more sophisticated ways of modeling the acoustic property of the environment and combining cross-modal information. Although our framework achieves highly satisfactory results at all difficulty levels on both simulated and real-world samples, we notice the performance of our model can be improved for situations with extreme reverb effects, and far away subjects. A potential use case of our work can be to leverage the properties of target visual scenes to provide immersive experiences to users in AR/VR applications. This work can also find applications in the audio/speech simulation domain.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019.

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020.

[4] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2020.

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

[6] 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. *arXiv preprint arXiv:1709.06158*, 2017.

[7] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021.

[8] 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.

[9] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020.

[10] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*, 2020.

[11] 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 *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[12] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. *arXiv preprint arXiv:2106.07732*, 2021.

[13] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, pages 35–51. Springer, 2020.

[14] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. *arXiv preprint arXiv:2111.05846*, 2021.

[15] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*, 2020.

[16] Marc Eder, Pierre Moulon, and Li Guan. Pano popups: Indoor 3d reconstruction with a plane-aware network. In *2019 International Conference on 3D Vision (3DV)*, pages 76–84. IEEE, 2019.

[17] 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. *arXiv preprint arXiv:1804.03619*, 2018.

[18] 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.

[19] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[20] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 658–676. Springer, 2020.

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

[22] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019.

[23] Ritwik Giri, Michael L Seltzer, Jasha Droppo, and Dong Yu. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5014–5018. IEEE, 2015.

[24] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[25] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[27] Sindhu B Hegde, KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Visual speech enhancement without a real visual stream. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1926–1935, 2021.

[28] Martin Holters, Tobias Corbach, and Udo Zölzer. Impulse response measurement techniques and their applicability in the real world. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, pages 108–112, 2009.

[29] Tsun-An Hsieh, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao. Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement. *arXiv preprint arXiv:2010.15174*, 2020.

[30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[31] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019.

[32] Marco Jeub, Magnus Schäfer, Hauke Krüger, Christoph Nelke, Christophe Beaugeant, and Peter Vary. Do we need dereverberation for hand-held telephony? In *Proc. Int. Congress on Acoustics (ICA), Sydney, Australia*, 2010.

[33] Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *2009 16th International Conference on Digital Signal Processing*, pages 1–5. IEEE, 2009.

[34] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, et al. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016:1–19, 2016.

[35] Keisuke Kinoshita, Marc Delcroix, Haeyong Kwon, Takuma Mori, and Tomohiro Nakatani. Neural network-based spectrum estimation for online wpe dereverberation. In *Interspeech*, pages 384–388, 2017.

[36] Homare Kon and Hideki Koike. Estimation of late reverberation characteristics from a single two-dimensional environmental image using convolutional neural networks. *Journal of the Audio Engineering Society*, 2019.

[37] Homare Kon and Hideki Koike. Estimation of late reverberation characteristics from a single two-dimensional environmental image using convolutional neural networks. *Journal of the Audio Engineering Society*, 67(7/8):540–548, 2019.

[38] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[39] Vinay Kothapally and John HL Hansen. Complex-valued time-frequency self-attention for speech dereverberation. *arXiv preprint arXiv:2211.12632*, 2022.

[40] Vinay Kothapally and John HL 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.

[41] Guinan Li, Jianwei Yu, Jiajun Deng, Xunying Liu, and Helen Meng. Audio-visual multi-channel speech separation, dereverberation and recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6042–6046. IEEE, 2022.

[42] 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. *arXiv preprint arXiv:2204.05841*, 2022.

[43] Shiguang Liu and Dinesh Manocha. In *Sound Synthesis, Propagation, and Rendering*, pages 7–28. Springer, 2020.

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[45] Yi Luo and Nima Mesgarani. Real-time single-channel dereverberation and separation with time-domain audio separation network. In *Interspeech*, pages 342–346, 2018.

[46] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[47] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 551–569. Springer, 2022.

[48] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021.

[49] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[50] Damian Murphy, Antti Kelloniemi, Jack Mullen, and Simon Shelley. Acoustic modeling using the digital waveguide mesh. *IEEE Signal Processing Magazine*, 24(2):55–66, 2007.

[51] Damian T Murphy and Simon Shelley. Openair: An interactive auralization web resource and database. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[52] 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.

[53] Patrick A Naylor, Nikolay D Gaubitch, et al. *Speech dereverberation*, volume 2. Springer, 2010.

[54] Stephen T Neely and Jont B Allen. Invertibility of a room impulse response. *The Journal of the Acoustical Society of America*, 66(1):165–169, 1979.

[55] 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.

[56] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[57] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 801–816. Springer, 2016.

[58] 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. IEEE, 2015.

[59] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE, 2021.

[60] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020.

[61] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[62] 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*, MM '22, page 924–933, New York, NY, USA, 2022. Association for Computing Machinery.

[63] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition. In *Proc. Interspeech 2021*, pages 286–290, 2021.

[64] 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.

[65] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

[66] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001.

[67] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020.

[68] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[69] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 286–295, 2021.

[70] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[71] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio engineering society*, 50(4):249–262, 2002.

[72] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Acoustic matching by embedding impulse responses. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 426–430. IEEE, 2020.

[73] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. *arXiv preprint arXiv:2006.05694*, 2020.

[74] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. *arXiv preprint arXiv:2006.05694*, 2020.

[75] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Hifi-gan-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 166–170. IEEE, 2021.

[76] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d represen-

tation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019.

[77] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černockỳ. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876, 2019.

[78] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černocký. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876, 2019.

[79] Ke Tan, Yong Xu, Shi-Xiong Zhang, Meng Yu, and Dong Yu. Audio-visual speech separation and dereverberation with a two-stage multimodal network. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):542–553, 2020.

[80] Ke Tan, Yong Xu, Shi-Xiong Zhang, Meng Yu, and Dong Yu. Audio-visual speech separation and dereverberation with a two-stage multimodal network. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):542–553, 2020.

[81] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. Gwa: A large high-quality acoustic dataset for audio processing. SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery.

[82] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021.

[83] James Traer and Josh H McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[85] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360deg layout estimation via differentiable depth rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12956–12965, 2021.

[86] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.

[87] Zhong-Qiu Wang and DeLiang Wang. Deep learning based target cancellation for speech dereverberation. *IEEE/ACM transactions on audio, speech, and language processing*, 28:941–950, 2020.

[88] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 24(3):483–492, 2015.

[89] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for joint enhancement of magnitude and phase. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2016.

[90] Bo Wu, Kehuang Li, Fengpei Ge, Zhen Huang, Minglei Yang, Sabato Marco Siniscalchi, and Chin-Hui Lee. An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1289–1300, 2017.

[91] Bo Wu, Kehuang Li, Minglei Yang, and Chin-Hui Lee. A reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM transactions on audio, speech, and language processing*, 25(1):102–111, 2016.

[92] Xiong Xiao, Shengkui Zhao, Duc Hoang Ha Nguyen, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li. Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation. *EURASIP Journal on Advances in Signal Processing*, 2016(1):1–18, 2016.

[93] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 882–891, 2019.

[94] Karren Yang, Dejan Marković, Steven Krenn, Vasu Agrawal, and Alexander Richard. Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8227–8237, 2022.

[95] 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.

[96] Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. *arXiv preprint arXiv:2202.10910*, 2022.

[97] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019.

[98] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[99] Yan Zhao, Zhong-Qiu Wang, and DeLiang Wang. Two-stage deep learning for noisy-reverberant speech enhancement. *IEEE/ACM transactions on audio, speech, and language processing*, 27(1):53–62, 2018.