Changan Chen[1]
N
[1]University c

Al

We introduce the novel-
task: given the sight and s
point, can we synthesize t
unseen target viewpoint?
approach: Visually-Guided
work that learns to synth
point in space by analyzi
To benchmark this task, v
large-scale multi-view aud
and one real. We show tha
about the spatial cues and s
datasets. To our knowledg
first formulation, dataset, a
view acoustic synthesis ta
applications ranging from
locked by this work, we bel
synthesis is in multi-modal

1. Introduction

Replaying a video recording from a new viewpoint[1] has many applications in cinematography, video enhancement, and virtual reality. For example, it can be used to edit a video, simulate a virtual camera, or, given a video of a personal memory, even enable users to experience a treasured moment again—not just on a 2D screen, but in 3D in a virtual or augmented reality, thus 'reliving' the moment.

While the applications are exciting, there are still many unsolved technical challenges. Recent advances in 3D reconstruction and novel-view synthesis (NVS) address the problem of synthesizing new *images* of a given scene [31, 33, 43]. However, thus far, the view synthesis problem is concerned with creating visuals alone; the output is silent or at best naively adopts the sounds of the original video (from the "wrong" viewpoint). Without sound, the emotional and cognitive significance of the replay is severely diminished.

In this work, we address this gap and introduce the new task of *novel-view acoustic synthesis* (NVAS). The goal of this task is to synthesize the sound in a scene from a new acoustic viewpoint, given only the visual and acoustic input from another source viewpoint in the same scene (Fig. 1).

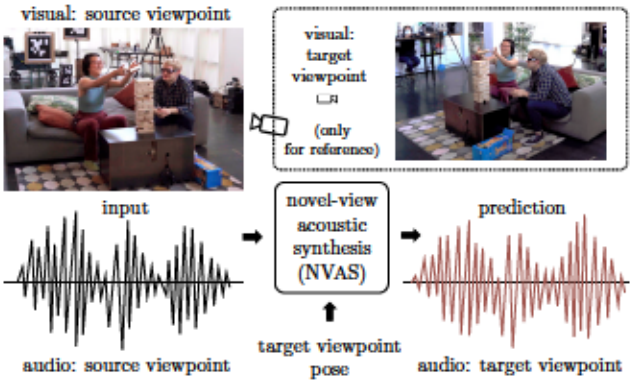

Figure 1. **Novel-view acoustic synthesis task.** Given audio-visual observations from one viewpoint and the relative target viewpoint pose, render the sound received at the target viewpoint. Note that the target is expressed as the desired pose of the microphones; the image at that pose (right) is neither observed nor synthesized.

NVAS is very different from the existing NVS task, where the goal is to reconstruct images instead of sounds, and these differences present new challenges. First, the 3D geometry of most real-life scenes changes in a limited manner during the recording. On the contrary, sound changes substantially over time, so the reconstruction target is highly dynamic. Secondly, visual and audio sensors are very different. A camera matrix captures the light in a highly-directional manner, and a single image comprises a large 2D array of pixels. In contrast, sounds are recorded with one or two microphones which are at best weakly-directional, providing only a coarse sampling of the sound field. Thirdly, the frequency of light waves is much higher than that of sound waves; the length of audio waves is thus larger to the point of being comparable to the size of geometric features of the scene, meaning that effects such as diffraction are often dominant, and spatial resolution is low. As a result, techniques that require spatial precision, such as triangulation and segmentation, are not applicable to audio. Lastly, sounds mix together, making it difficult to segment them,

[1]We use "viewpoint" to mean a camera or microphone pose.

and they are affected by environmental effects such as reverberation that are distributed and largely unobservable.

While the NVS and NVAS tasks are indeed very different, we hypothesize that NVAS is an inherently multimodal task. In fact, vision can play an important role in achieving accurate sound synthesis. First, establishing correspondences between sounds and their sources as they appear in images can provide essential cues for resynthesizing the sounds realistically. For instance, human speech is highly directional and sounds very differently if one faces the speaker or their back, which can only be inferred from visual cues. In addition, the environment acoustics also affect the sound one hears as a function of the scene geometry, materials, and emitter/receiver locations. The same source sounds very differently if it is located in the center of a room, at the corner, or in a corridor, for example. In short, vision provides cues about space and geometry that affect sound, and are difficult to estimate from the sound alone.

In order to validate our hypothesis, we propose a novel *visually-guided acoustic synthesis network* that analyzes audio and visual features and synthesizes the audio at a target location. More specifically, the network first takes as input the image observed at the source viewpoint in order to infer global acoustic and geometric properties of the environment along with the bounding box of the active speaker. The network then reasons how the speaker and scene geometry change in 3D based on the relative target pose with a fusion network. We inject the fused features into audio with a gated multi-modal fusion network and model the acoustic changes between viewpoints with a time-domain model.

In order to conduct our experiments on the new NVAS task, we require suitable training and benchmarking data, of which currently there is none available. To address that, we contribute two new datasets: one real (Replay-NVAS) and one synthetic (SoundSpaces-NVAS). The key feature of these datasets is to record the sight and sound of different scenes from multiple cameras/viewpoints. Replay-NVAS contains video recordings of groups of people performing social activities (e.g., chatting, watching TV, doing yoga, playing instruments) from 8 surrounding viewpoints simultaneously. It contains 72 hours of highly realistic everyday conversation and social interactions in one home-like environment. To our knowledge, Replay-NVAS represents the first large-scale real-world dataset enabling NVAS. This dataset would also greatly benefit many other existing tasks including NVS, active speaker localization, etc. For SoundSpaces-NVAS, we render 1.3K hours of audio-visual data based on the SoundSpaces [7] platform. Using this simulator, one can easily change the scene geometry and the positions of speakers, cameras, and microphones. This data serves as a powerful test bed with clean ground truth for a large collection of home environments, offering a good complement to Replay-NVAS. For both datasets, we capture binaural audio, which is what humans perceive with two ears. Together the datasets contain 1,372 hours of audio-visual capture, with 1,032 speakers across 121 3D scenes. We will release both datasets.

We show that our model outperforms traditional signal processing approaches as well as learning-based baselines, often by a substantial margin, in a quantitative evaluation and a human study. We show qualitative examples where the model predicts acoustic changes according to the viewpoint changes, e.g., left channel becomes louder when the viewpoint changes from left to right. In a nutshell, we present the first work that deals with novel-view acoustic synthesis, and contribute two large-scale datasets along with a novel neural rendering approach for solving the task.

2. Related Work

Novel-view synthesis (NVS). Kickstarted by advances in neural rendering [33, 51], many recent works consider variants of the NVS problem. Most approaches assume dozens of calibrated images for reconstructing a single static scene. Closer to monocular video NVS, authors have considered reducing the number of input views [19, 24, 37, 45, 60] and modelling dynamic scenes [26, 27, 41, 42, 53, 55]. However, none of these works tackle audio.

Acoustic matching and spatialization. NVAS requires accounting for (1) the environmental acoustics and (2) the geometric configuration of the target microphone(s) (e.g., monaural vs binaural). Modelling environmental acoustics has been addressed extensively by the audio community [4, 25]. Room impulse response (RIR) functions characterize the environment acoustics as a transfer function between the emitter and receiver, accounting for the scene geometry, materials, and emitter/receiver locations. Estimating the direct-to-reverberant ratio and the reverberation time, is sufficient to synthesize simple RIRs that match audio in a plausible manner [11, 15, 21, 29, 36, 59]. These methods do not synthesize for a target viewpoint, rather they resynthesize to match an audio sample. In [46, 47] sound from a moving emitter is spatialized towards a receiver conditioned on the tracked 3D location of the emitter.

Recently, the vision community explores using visual information to estimate environmental acoustics [6, 50]. However, these works only synthesize acoustics for a given viewpoint rather than a novel viewpoint. In addition, they have only addressed monaural audio, which is more forgiving than binaural because humans do not perceive subtle absolute acoustic properties, but can detect easily inconsistencies in the sounds perceived by the two ears. Recent work spatializes monaural sounds by upmixing them to multiple channels conditioned on the video, where the sound emitters are static [16, 35]. Because the environment, emitter and receiver are static, so are the acoustics. Other work predicts impulse responses in simulation either for a single

environment [28], or by using few-shot egocentric observations [30], or by using the 3D scene mesh [44]. While simulated results are satisfying, those models' impact on real-world data is unknown, especially for scenarios where human speakers move and interact with each other. Unlike any of the above, we introduce and tackle the NVAS problem, accounting for both acoustics and spatialization, and we propose a model that addresses the problem effectively on both synthetic and real-world data.

Audio-visual learning. Recent advances in multi-modal video understanding enable new forms of self-supervised cross-modal feature learning from video [2, 23, 34], sound source localization [18, 20, 54], and audio-visual speech enhancement and source separation [1, 13, 32, 39, 61]. All of these existing tasks and datasets only deal with a single viewpoint. We introduce the first audio-visual learning task and dataset that deals with multi-view audio-visual data.

3. The Novel-view Acoustic Synthesis Task

We introduce a new task, *novel-view acoustic synthesis* (NVAS). Assuming there are N sound emitters in the scene (emitter i emits C^i from location L^i), given the audio A_S and video V_S observed at the source viewpoint S, the goal is to synthesize the audio A_T at the target viewpoint T, as it would sound from the target location, specified by the relative pose P_T of the target microphone (translation and orientation) with respect to the source view (Fig. 1). Furthermore, we assume that the active sound emitters in the environment are visible in the source camera, but we make no assumptions about the camera at the target location.

The sound received at any point R can be expressed as:

$$A_R = \mathcal{F}(L^{1,\dots,N}, C^{1,\dots,N}, R \mid E), \tag{1}$$

where R is the receiver location (S or T) and E is the environment. The emitted sounds C^i are not restricted to speech but can be ambient noise, sounding objects, etc. Our goal here is to learn a transfer function $\mathcal{T}(\cdot)$ defined as $A_T = \mathcal{T}(A_S, V_S, P_T)$, where $S, T, L^{1,\dots,N}, C^{1,\dots,N}, E$ are not directly given and need to be inferred from V_S and P_T, which makes the task inherently multi-modal.

This task is challenging because the goal is to model the sound field of a dynamic scene and capture acoustic changes between viewpoints given one pair of audio-visual measurements. While traditional signal processing methods can be applied, we show in Sec. 6 that they perform poorly. In this work, we present a neural rendering approach.

4. Datasets

We introduce two datasets for the NVAS task: live recordings (Sec. 4.1), and simulated audio in scanned real-world environments (Sec. 4.2) (see Fig. 2). The former is real and covers various social scenarios, but offers limited diversity of sound sources, viewpoints and environments, and is noisy. The latter has a realism gap, but allows perfect control over these aforementioned elements.

Figure 2. Example source and target views for the two introduced datasets: Replay-NVAS (left) and SoundSpaces-NVAS (right).

Both datasets focus on human speech given its relevance in applications. However, our model design is not specific to speech. For both datasets, we capture binaural audio, which best aligns with human perception. Note that for both datasets, we collect multiple multi-modal views for training and evaluation; during inference the target viewpoint(s) (and in some cases target environment) are withheld. We will release both datasets to assist future research.

4.1. The Replay-NVAS Dataset

Replay-NVAS contains multi-view captures of acted scenes in apartments. We capture 109 different scenarios (e.g., having a conversation, having dinner, or doing yoga) from 8 different viewpoints. In total, we collect 72 hours of video data, involving 32 participants across all scenarios.

In each scenario, we invite 2–4 participants to act on a given topic. Each participant wears a near-range microphone, providing a clean recording of their own speech. The scene is captured by 8 DLSR cameras, each augmented with a 3Dio binaural microphone. In this way, the data captures video and audio simultaneously from multiple cameras, resulting in 56 possible source/target viewpoint combinations for each scene. The videos are recorded at 30 FPS and the audio is recorded with a 48k sampling rate. We use a clapper at the beginning of the recording for temporal synchronization. Each scenario lasts 3–8 min. We use off-the-shelf software for multi-view camera calibration (see Supp.).

To construct the dataset, we extract one-second long clips from each video with overlapping windows. We automatically remove silent and noisy clips based on the energy of near-range microphones, which results in 77K/12K/2K clips in total for train/val/test (details in Supp.) During training, for one sample, we randomly select two out of eight viewpoints, one as the source and one as the target.

This dataset is very challenging. It covers a wide range of social activities. It is harrowed by ambient sound, room

reverberation, overlapping speech and non-verbal sounds such as clapping and instruments. Participants can move freely in the environment. We believe that this data will be useful to the community beyond the NVAS task as it can be used for benchmarking many other problems, including active speaker localization, source separation, and NVS.

4.2. The SoundSpaces-NVAS Dataset

In this dataset, we synthesize multi-view audio-visual data of two people having conversations in 3D scenes. In total, we construct 1.3K hours of audio-visual data for a total of 1,000 speakers, 120 3D scenes and 200K viewpoints.

Our goal is to construct audio-visual data with strong spatial and acoustic correspondences across multiple viewpoints, meaning that the visual information should indicate what the audio should sound like, e.g., observing speaker on the left should indicate the left ear is louder and observing speaker at a distance should indicate there is higher reverberation. We use the SoundSpaces 2.0 platform [7], which allows highly realistic audio and visual rendering for arbitrary camera and microphone locations in 3D scans of real-world environments [5, 52, 58]. It accounts for all major real-world acoustics phenomena: direct sounds, early specular/diffuse reflections, reverberation, binaural spatialization, and effects from materials and air absorption.

We use the Gibson dataset [58] for scene meshes and LibriSpeech [40] for speech samples. As we are simulating two people having conversations, for a given environment, we randomly sample two speaker locations within 3 m and insert two copyright-free mannequins (one male and one female) at these two locations.[2] We then randomly sample four nearby viewpoints facing the center of the two speakers at a height of 1.5 m (Fig. 2, right). For each speaker, we select a speech sample from LibriSpeech with matching gender. We render images at all locations as well as binaural impulse response for all pairs of points between speakers and viewpoints. The received sound is obtained by convolving the binaural impulse response with the speech sample.

During training, for one sample, we randomly sample two out of four rendered viewpoints, one as the source and one as the target. We also randomly choose one speaker to be active, simulating what we observe on the real data (i.e., usually only one person speaks at a time).

5. Visually-Guided Acoustic Synthesis

We introduce a new method, **Vi**sually-**G**uided **A**coustic **S**ynthesis (ViGAS), to address the NVAS problem, taking as input sound and an image and outputting the sound from a different target microphone pose.

ViGAS consists of five components: ambient sound separation, active speaker localization, visual acoustic network, acoustic synthesis, and temporal alignment. The high-level idea is to separate the observed sound into primary and ambient, extract useful visual information (active speaker and acoustic features), and use this information to guide acoustic synthesis for the primary sound. Temporal alignment is performed during training for better optimization. ViGAS is discussed in detail next and summarised in Fig. 3.

5.1. Ambient Sound Separation

ViGAS starts by decomposing the input sound into primary and ambient (traffic, electric noise from a fridge or the A/C, etc.). Ambient sound is important for realism, but it also interferes with learning the model because it can carry significant energy, making the model focus on it rather than on the primary sounds, and its spatial distribution is very different from the primary sounds.

By explicitly separating primary and ambient sounds, ViGAS: (1) accounts for the fact that the transfer functions of primary and ambient sounds are very different and thus difficult to model together; (2) avoids wasting representational power on modelling ambient sounds that might be difficult to reconstruct accurately and depend less on the viewpoint; and (3) prevents ambient sounds, which are noise-like and high-energy, from dominating learning and reconstruction. In practice, as we show in Sec. 6, without the ambient sound separation, the model performs poorly.

The goal of ambient sound separation is thus to construct a function $(A_C, A_N) = \mathcal{P}(A_S)$ that separates the input sound A_S into primary sound A_C and ambient sound A_S. Existing approaches to this problem are based on signal processing [3, 12] or learning [10, 14]. We find that pretrained speech enhancement models such as Denoiser [10] tend to aggressively remove the noise including the primary sound, which hinders re-synthesis. We thus opt for bandpass filtering, passing frequencies within a certain range and rejecting/attenuating frequencies outside of it, which we found to work well. We cut frequencies below 80 Hz for SoundSpaces-NVAS and 150 Hz for Replay-NVAS.

5.2. Active Speaker Localization

Knowing where the emitters of different primary sounds are located in the environment can help to solve the NVAS task. In this paper, we focus on localizing the active speaker, although there can be other important primary sound events like instruments playing, speakers interacting with objects, etc. The goal of active speaker localization is to predict the bounding box of the active speaker in each frame of the video (examples in Fig. 4). The bounding box is in the format of $(y_{\min}, y_{\max}, x_{\min}, x_{\max})$ and x, y are normalized to $[0, 1]$ by the image width and height, respectively.

On SoundSpaces-NVAS, this task is relatively easy because of the strong correspondence between the appearance of the speaker and the gender of the speech sample, which enables to easily train a classifier for active speakers. How-

[2] https://renderpeople.com/free-3d-people

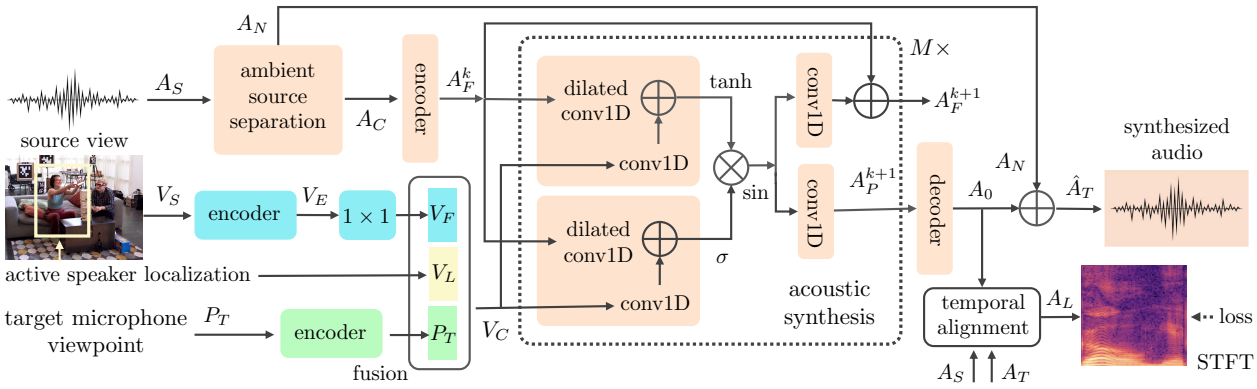

Figure 3. **Visually Guided Acoustic Synthesis (ViGAS)**. Given the input audio A_S, we first separate out the ambient sound to focus on the sound of interest. We take the source audio and source visual to localize the active speaker on the 2D image. We also extract the visual acoustic features of the environment by running an encoder on the source visual. We concatenate the active speaker feature, source visual features, and the target pose, and fuse these features with a MLP. We feed both the audio stream A_C and fused visual feature V_C into the acoustic synthesis network, which has M stacked audio-visual fusion blocks. In each block, the audio sequence is processed by dilated conv1d layers and the visual features are processed by conv1d layers. Lastly, the previously separated ambient sound is added back to the waveform. During training, our temporal alignment module shifts the prediction by the amount of delay estimated between the source and the target audio to align the prediction well with the target.

ever, this is much harder on Replay-NVAS because cameras record speakers from a distance and from diverse angles, meaning that lip motion, the main cue used by speaker localization methods [20, 48, 54], is often not visible. Hence, the model has to rely on other cues to identify the speaker (such as body motion, gender or identity). Furthermore, sometimes people speak or laugh over each other.

Since our focus is not speaker localization, for the Replay-NVAS we assume that this problem is solved by an external module that does audio-visual active speaker localization. To approximate the output of such a module automatically, we rely on the near-range audio recordings. Specifically, we first run an off-the-shelf detection and tracker [9] on the video at 5 FPS and obtain, with some manual refinement, bounding boxes B_t^i for $i = 1, \ldots, N$ at each frame t. We manually assign the near-range microphone audio A_N^i to each tracked person. We select the active speaker D based on the maximum energy of each near-range microphone, i.e., $D = \text{argmax}_i \left\{ \sum A_N^i[t : t+\Delta t]^2 \right\}$, where Δt is the time interval we use to calculate the audio energy. We output bounding box B^D as the localization feature V_L.

5.3. Visual Acoustic Network and Fusion

The active speaker bounding box B^D only disambiguates the active speaker from all visible humans on 2D, which is not enough to indicate where the speaker is in 3D. To infer that, the visual information is also needed. Since there is usually not much movement in one second (the length of the input video clip), the video clip does not provide much extra information compared to a single frame. Thus, we choose the middle frame to represent the clip and extract the visual acoustic features V_E from the input RGB image with a pretrained ResNet18 [17] before the average pooling layer to preserve spatial information. To reduce the feature size, we feed V_E into a 1D convolution with kernel size 1 and output channel size 8. We then flatten the visual features to obtain feature V_F.

The target pose is specified as the translation along x, y, z axes plus difference between orientations of the source "view" and the target "view" expressed via rotation angles: $+y$ (roll), $+x$ (pitch) and $+z$ (yaw). We encode each angle α as its sinusoidal value: $(\sin(\alpha), \cos(\alpha))$.

Similarly, the target pose is not enough by itself to indicate where the target viewpoint T is in the 3D space; to infer that, the source view V_S is again needed. For example, in top row of Fig 4, for target viewpoint 3, "two meters to the right and one meter forward" is not enough to indicate the target location is in the corridor, while the model can reason that based on the source view.

We use a fusion network to predict a latent representation of the scene variables S, T, L^D, E (*cf*. Sec. 3) by first concatenating $[V_L, P_T, V_F]$ and then feeding it through a multilayer perceptron (MLP). See Fig. 3 for the network.

5.4. Acoustic Synthesis

With the separated primary sound A_C and the visual acoustic feature V_C as input, the goal of the acoustic synthesis module is to transform A_C guided by V_C. We design the acoustic synthesis network to learn a non-linear transfer function (implicitly) that captures these major acoustic phenomena, including the attenuation of sound in space, the directivity of sound sources (human speech is directional), the reverberation level, the head-related transfer function, as well as the frequency-dependent acoustic phenomena. Training end-to-end makes it possible to capture these subtle and complicated changes in the audio.

Inspired by recent advances in time-domain signal modeling [38, 47], we design the network as M stacked synthesis blocks, where each block consists of multiple conv1D layers. We first encode the input audio A_C into a latent space, which is then fed into the synthesis block. The key of the synthesis block is a gated multimodal fusion network that injects the visual information into the audio as follows:

$$z = \tanh(p_A^k(A_F^k) + p_V^k(V_C)) \odot \sigma(q_A^k(A_F^k) + q_V^k(V_C)), \quad (2)$$

where $\odot$ indicates element-wise multiplication, σ is a logistic sigmoid function, $k = 1, \ldots, M$ is the layer index and p, q are both learnable 1D convolutions.

After passing z through a sinusoidal activation function, the network uses two separate conv1D layers to process the feature, one producing the residual connection A_F^{k+1} and one producing the skip connection A_P^{k+1}. All skip connections A_P^{k+1} are mean pooled and fed into a decoder to produce the output A_O. We add back the separated ambient sound A_N as the target audio estimate: $\hat{A}_T = A_O + A_N$. See Supp. for more details on the architecture.

5.5. Temporal Alignment

In order for the model to learn well, it is important that input and output sounds are temporally aligned. While the Replay-NVAS data is already synchronised based on the clapper sound, due to the finite speed of sound, the sounds emitted from different locations may still arrive at microphones with a delay slightly different from the one of the clapper, causing misalignments that affect training.

To align source and target audio for training, we find the delay τ that maximizes the generalized cross-correlation:

$$\mathcal{R}_{A_S, A_T}(\tau) = \mathbb{E}_t[h_S(t) \cdot h_T(t - \tau)], \quad (3)$$

where h_S and h_T are the feature embedding for A_S and A_T respectively at time t. We use the feature extractor h from the generalized cross-correlation phase transform (GCC-PHAT) algorithm [22], which whitens the audio by dividing by the magnitude of the cross-power spectral density. After computing τ, we shift the prediction A_O by τ samples to align with the A_T and obtain A_L. Note that alignment is already exact for SoundSpaces-NVAS.

5.6. Loss

To compute the loss, we first encode the audio with the short-time Fourier transform (STFT), a complex-valued matrix representation of the audio where the y axis represents frequency and the x axis is time. We then compute the magnitude of the STFT, and optimize the L1 loss between the the predicted and ground truth magnitudes as follows:

$$L = \big| ||STFT(A_L)||_2 - ||STFT(A'_T)||_2 \big|, \quad (4)$$

where A'_T is the primary sound separated from A_T with $\mathcal{P}(\cdot)$. By taking the magnitude, we do not model the exact phase values, which we find hinders learning if being included in the loss. See implementation details in Supp.

6. Experiments

We compare with several traditional and learning-based baselines and show that ViGAS outperforms them in both a quantitative evaluation and a human subject study.

Evaluation. We measure performance from three aspects: 1. closeness to GT as measured by the **magnitude spectrogram distance (Mag)**. 2. correctness of the spatial sound as measured by the **left-right energy ratio error (LRE)**, i.e., the difference of ratio of energy between left and right channels and 3. correctness of the acoustic properties measured by **RT60 error (RTE)** [6, 50], i.e., the error in reverberation time decaying by 60dB (RT60). We use a pretrained model [6] to estimate RT60 directly from speech.

Baselines. We consider the following baselines: 1. **Input audio**. Copying the input to the output. 2. **TF Estimator [56] + Nearest Neighbor**, i.e. storing the transfer function estimated during training and retrieving the nearest neighbor during test time. We estimate transfer functions with a Wiener filter [56] and index them with the ground-truth locations of the speaker, source viewpoint, and target viewpoint for the single environment setup and their relative pose for the novel environment setup. At test time, this method searches the database to find the nearest transfer function and applies it on the input audio. 3. **Digital Signal Processing (DSP) [8]** approach that takes the distance, azimuth, and elevation of the sound source, applies an inverse a head-related transfer function (HRTF) to estimate the speech spoken by the speaker and then applies another HRTF to estimate the audio at the target microphone location. This baseline adjusts the loudness of the left and right channels based on where the speaker is in the target view. We supply GT coordinates for SoundSpaces-NVAS and speakers' head positions estimated with triangulation on Replay-NVAS. 4. **Visual Acoustic Matching (VAM) [6]**, recently proposed for a related task of matching acoustics of input audio with a target image. This task only deals with single viewpoint and single-channel audio. We adapt their model with minimal modification by feeding in the image from the *source* viewpoint and concatenating the position offset of the target microphone at the multimodal fusion step. See Supp. for details.

6.1. Results on SoundSpaces-NVAS

Table 1 shows the results. For synthetic data, we consider two evaluation setups: 1. single environment: train and test on the same environment and 2. novel environment: train and test on multiple non-overlapping Gibson environments (90/10/20 for train/val/test).

In the single environment setup, our model largely outperforms all baselines as well as our audio-only ablation on all metrics. TF Estimator performs poorly despite being indexed by the ground truth location values because es-

	SoundSpaces-NVAS						Replay-NVAS		
	Single Environment			*Novel Environment*			*Single Environment*		
	Mag	LRE	RTE	Mag	LRE	RTE	Mag	LRE	RTE
Input audio	0.225	1.473	0.032	0.216	1.408	0.039	0.159	1.477	0.046
TF Estimator [56]	0.359	2.596	0.059	0.440	3.261	0.092	0.327	2.861	0.147
DSP [8]	0.302	3.644	0.044	0.300	3.689	0.047	0.463	1.300	0.067
VAM [6]	0.220	1.198	0.041	0.235	1.131	0.051	0.161	0.924	0.070
ViGAS w/o visual	0.173	0.973	0.031	0.181	1.007	0.036	0.146	0.877	**0.046**
ViGAS	**0.159**	**0.782**	**0.029**	**0.175**	**0.971**	**0.034**	**0.142**	**0.716**	0.048

Table 1. **Results on SoundSpaces-NVAS and Replay-NVAS.** We report the magnitude spectrogram distance (Mag), left-right energy ratio error (LRE), and RT60 error (RTE). Replay-NVAS does not have novel environment setup due to data being collected in a single environment. For all metrics, lower is better. In addition to baselines, we also evaluate ViGAS w/o visual by removing the active speaker localization and visual features. Note that reverberation time is mostly invariant of the receiver location in the same room and thus input audio has low RTE. A good model should preserve this property while synthesizing the desired acoustics for the target viewpoint.

timating a transfer function directly from two audio clips is non-trivial and noisy for low-energy parts of the signal. DSP also performs badly despite having the ground truth 3D coordinates of the sound source. This is because head related transfer functions are typically recorded in anechoic chambers, which does not account for acoustics of different environments, e.g., reverberation. Both traditional approaches perform worse than simply copying the input audio, indicating that learning-based models are needed for this challenging task. The recent model VAM [6] performs much better compared to the traditional approaches but still underperforms our model. There is a significant difference between ViGAS w/o visual and the full model; this shows that the visual knowledge about the speaker location and the environment is important for this task.

Fig. 4 shows an example where given the same input source viewpoint, our model synthesizes audio for three different target viewpoints. The model reasons about how the geometry and speaker locations changes based on the source view and the target pose, and predicts the acoustic difference accordingly. See Supp. video to listen to sounds.

For the novel environment setup, our model again outperforms all baselines. Compared to ViGAS in the single environment setup, both the magnitude spectrogram distance and the left-right energy ratio error increase. This is expected because for novel (unseen) environments, single images capture limited geometry and acoustic information. The model fails sometime when there is a drastic viewpoint change, e.g., target viewpoint 3 in Fig. 4. This setup requires the model to reason or "imagine" the environment based on single audio-visual observation, which poses great challenge for NVAS as well as NVS, where typically synthesis is performed in a fully observed environment.

Ablations. Table 2 shows ablations on the model design. To understand if the model uses visual information, we ablate the visual features V_F and the active speaker feature V_L. Removing the active speaker feature leads to less damage on the model performance, because without the explicitly localized active speaker, the model can still implicitly reason about the active speaker location based on the image and audio. If both are removed ("ViGAS w/o visual" in Table 1), the performance suffers most.

ViGAS	SS-NVAS		Replay-NVAS	
	Mag	LRE	Mag	LRE
full model	**0.159**	0.782	**0.142**	0.716
w/o visual features	0.171	0.897	0.146	0.920
w/o ASL	0.161	0.814	0.143	0.757
w/o alignment	0.176	**0.771**	0.144	**0.706**
w/o separation	0.165	0.840	0.182	0.859

Table 2. **Ablations of the model on both datasets**.

To study the effectiveness of the temporal alignment and ambient sound separation modules, we ablate them separately. Removing the temporal alignment leads to higher Mag error and slightly lower LRE. As for ambient sound separation, the results show that optimizing for the high-energy noise-like ambient sound degrades the performance.

6.2. Results on Replay-NVAS

Table 1 (right) shows the Replay-NVAS results. Compared to SoundSpaces-NVAS, the magnitudes of all errors are smaller because there are less drastic acoustic changes between viewpoints (8 DLSR cameras form a circle around the participants). Traditional approaches like TF Estimator and DSP still perform poorly despite using the 3D coordinates of the camera and the speaker (triangulated from multiple cameras). VAM performs better due to end-to-end learning; however, our model outperforms it. Compared to ViGAS w/o visual, the full model has much lower left-right energy ratio error and slightly higher reverberation time error, showing that the model takes into account the speaker position and viewpoint change for synthesizing the audio.

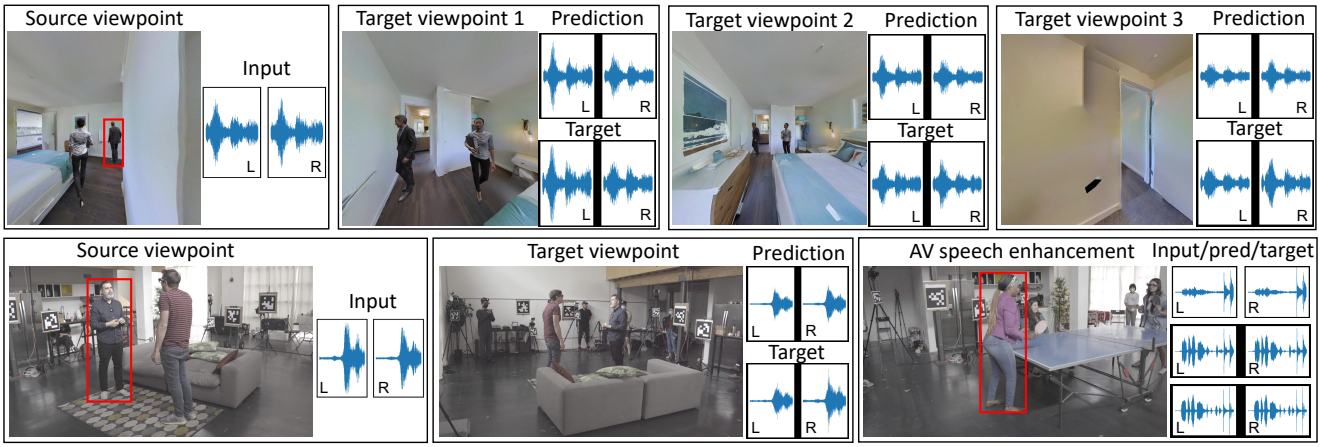

Figure 4. **Qualitative examples**. For all binaural audio, we show the left-channel and the right-channel waveforms side-by-side. Row 1: SoundSpaces-NVAS example where given the source viewpoint and input audio, the model synthesizes audio for three different target viewpoints (target views are for reference only). In this case, the active speaker is the male speaker as indicated by the bounding box. For target viewpoint 1, the view rotates about 90 degrees and the male speaker is on the left side and the predicted left channel is louder than the right channel. Viewpoint 2 moves away from the speaker and thus yields lower amplitude compared to the first prediction. For target viewpoint 3, it is completely located outside of the living room, in which case, the sound could only come from the door open on the right (louder right channel) and the reverberation also greatly increases due to the vanishing direct sound. Row 2: Replay-NVAS example where the speaker is located on the left in the source viewpoint which becomes the right and further from the camera in target viewpoint 2, the model also predicts lower amplitude and louder right channel. On the right side, we show an example of the audio-visual speech enhancement for the active speaker. The model enhances the speech to largely match with the near-range audio (target).

	Mag	RTE
Input	0.279	0.376
ViGAS (ours)	**0.234**	**0.122**

Table 3. **Speech enhancement on Replay-NVAS**.

Fig. 4 (row 2, left) shows a qualitative example. In the source viewpoint, the active speaker is on the left, while in the target viewpoint, he is further from the camera and on the right. The model synthesizes an audio waveform that captures the corresponding acoustic change, showing that our model successfully learns from real videos.

Audio-visual speech enhancement. In some real-world applications, e.g., hearing aid devices, the goal is to obtain the enhanced clean speech of the active speaker. This can be seen as a special case of NVAS, where the target viewpoint is the active speaker. Our model is capable of performing audio-visual speech enhancement without any modification. We simply set the target audio to the near-range audio recording for the active speaker. We show the results in Table 3. Our model obtains cleaner audio compared to the input audio (example in Fig. 4, row 2, right).

Human subject study. To supplement the quantitative metrics and evaluate how well our synthesized audio captures the acoustic change between viewpoints, we conduct a human subject study. We show participants the image of the target viewpoint V_T as well as the audio A_T as reference. We provide three audio samples: the input, the prediction of ViGAS, and the prediction of DSP (the most naturally sounding baseline) and ask them to select a clip that sounds closest to the target audio. We select 20 examples from SoundSpaces-NVAS and 20 examples from Replay-NVAS and invite 10 participants to perform the study.

Dataset	Input	DSP	ViGAS
SoundSpaces-NVAS	24%	2%	**74%**
Replay-NVAS	43%	6%	**51%**

Table 4. **Human Study**. Participants favor our approach over the two most realistic sounding baselines, (1) copying the input signal, and (2) a digital signal processing baseline.

See Table 4 for the results. On the synthetic dataset SoundSpaces-NVAS, our approach is preferred over the baselines by a large margin. This margin is lower on the real-world Replay-NVAS dataset but is still significant.

7. Conclusion

We introduce the challenging novel-view acoustic synthesis task and a related benchmark in form of both real and synthetic datasets. We propose a neural rendering model that learns to transform the sound from the source viewpoint to the target viewpoint by reasoning about the observed audio and visual stream. Our model surpasses all baselines on both datasets. We believe this research unlocks many potential applications and research in multimodal novel-view synthesis. In the future, we plan to incorporate active-speaker localization model into the approach and let the model jointly learn to localize and synthesize.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, 2018. 3

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 3

[3] M. Berouti, Richard Schwartz, and John Makhoul. Enhancement of speech corrupted by acoustic noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979. 4

[4] J.S. Bradley. Review of objective room acoustics measures and future needs. *Applied Acoustics*, 2011. 2

[5] 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. *3DV*, 2017. MatterPort3D dataset license available at: http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf. 4

[6] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022. 2, 6, 7, 13

[7] 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. 2, 4

[8] Corey I. Cheng and Gregory H. Wakefield. Introduction to head-related transfer functions (hrtfs): representations of hrtfs in time, frequency, and space. *journal of the audio engineering society*, 49(4):231–249, april 2001. 6, 7

[9] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 5

[10] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020. 4

[11] James Eaton, Nikolay Gaubitch, Allistair Moore, and Patrick Naylor. Estimation of room acoustic parameters: The ACE challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10), 2016. 2

[12] Yariv Ephraim and Harry L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 1995. 4

[13] 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. In *SIGGRAPH*, 2018. 3

[14] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning (ICML)*, 2019. 4

[15] Hannes Gamper and Ivan J Tashev. Blind reverberation time estimation using a convolutional neural network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 136–140, 2018. 2

[16] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[18] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *NeurIPS*, 2020. 3

[19] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proc. ICCV*, 2021. 2

[20] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *CVPR*, 2022. 3, 5

[21] Florian Klein, Annika Neidhardt, and Marius Seipel. Real-time estimation of reverberation time for selection of suitable binaural room impulse responses. In *Audio for Virtual, Augmented and Mixed Realities: Proceedings of 5th International Conference on Spatial Audio (ICSA)*, pages 145–150, 2019. 2

[22] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976. 6

[23] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 3

[24] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. ViewFormer: NeRF-free neural rendering from few images using transformers. In *Proc. ECCV*, 2022. 2

[25] Heinrich Kuttruff. *Room Acoustics*. Boca Raton, 6th edition, 2016. 2

[26] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. *arXiv.cs*, abs/2011.13084, 2020. 2

[27] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural humans: Pose-controlled free-view synthesis of human actors with template-guided neural radiance fields. In *arXiv*, 2021. 2

[28] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *NeurIPS*, 2022. 3

[29] Wolfgang Mack, Shuwen Deng, and Emanuël AP Habets. Single-channel blind direct-to-reverberation ratio estimation using masking. In *INTERSPEECH*, pages 5066–5070, 2020. 2

[30] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 3

[31] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, June 2021. 1

[32] 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. In *arXiv*, 2020. 3

[33] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[34] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *NeurIPS*, 2020. 3

[35] Pedro Morgado, Nono Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *NeurIPS*, 2018. 2

[36] Prateek Murgai, Mark Rau, and Jean-Marc Jot. Blind estimation of the reverberation fingerprint of unknown acoustic environments. In *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017. 2

[37] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. CVPR*, 2022. 2

[38] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. cite arxiv:1609.03499. 6

[39] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 3

[40] 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. 4

[41] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. *CoRR*, abs/2011.12948, 2020. 2

[42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *arXiv.cs*, abs/2011.13961, 2020. 2

[43] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, June 2021. 1

[44] Anton Ratnarajah, Zhenyu Tang, Rohith Chandrashekar Aralikatti, and Dinesh Manocha. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *ACM Multimedia*, 2022. 3

[45] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *Proc. CVPR*, 2021. 2

[46] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022. 2

[47] Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Butler, Fernando de la Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021. 2, 6

[48] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. Ava-activespeaker: An audio-visual dataset for active speaker detection. In *ICASSP*, 2020. 5

[49] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 12

[50] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, 2021. 2, 6

[51] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. *Proc. NeurIPS*, 2019. 2

[52] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 4

[53] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Surface-free human 3d pose refinement via neural rendering. *arXiv.cs*, abs/2102.06199, 2021. 2

[54] 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*, page 3927–3935, 2021. 3, 5

[55] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a deforming scene from monocular video. *arXiv.cs*, abs/2012.12247, 2020. 2

[56] Norbert Wiener. Extrapolation, interpolation, and smoothing of stationary time series. *Report of the Services 19, Research Project DIC-6037 MIT*, 1942. 6, 7, 13

[57] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 12

[58] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 4, 12

[59] Feifei Xiong, Stefan Goetze, Birger Kollmeier, and Bernd T Meyer. Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(2):255–267, 2018. 2

[60] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. PixelNeRF: Neural radiance fields from one or few images. In *Proc. CVPR*, 2021. 2

[61] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *ICCV*, 2019. 3

8. Supplementary M

In this supplementary details about:

1. Supplementary vide model's performanc
2. Replay-NVAS datas of the main paper).
3. SoundSpaces-NVAS
4. Implementation and 5.6).
5. Baseline details (ref

8.1. Supplementary V

This video includes examples for the Replay-NVAS dataset and the SoundSpaces-NVAS dataset as well our model's prediction on both datasets. Listen with a headphone for the spatial sound.

8.2. Replay-NVAS Dataset Details

Multi-view camera calibration. We estimate camera poses with COLMAP [49] Structure-from-Motion (SfM) framework on each scene separately. Each scene is filmed with 8 static DSLR cameras and 3 wearable GoPro cameras (the latter are not used in our acoustic synthesis experiments). We first run SfM on the segments of the GoPro recordings where the wearers move significantly; followed by registration of the static camera frames to the model and a final round of bundle adjustment where we enforce constant relative poses between static camera frames taken at the same timestamp. This two-stage procedure greatly reduces the scale of the problem by making SfM focus first on the most diverse part of the data. Upon feature extraction stage, we cull the local features belonging to potentially dynamic object categories (such as persons or animals) as detected by Detectron2 instance segmentation [57]. We then exploit the stationarity of DSLRs by picking a medoid camera pose among the frames filmed by each camera. Finally, we rotate and scale the coordinate system so that Z axis is pointing roughly upwards and the distances between cameras match the approximate distances in centimeters. Fig. 5 plots all camera coordinates and orientations projected to XY plane.

Training data construction. We align different DSLR videos with the clapper sound, which gives us synchronized multi-view audio-visual data. However, this data is not directly usable for training because some data are noisy (e.g., people frequently talking over each other) or silent, which leads to additional learning challenges for the model. Thus, we design an automatic process for filtering out noisy clips. More specifically, we first extract all one-second audio clips

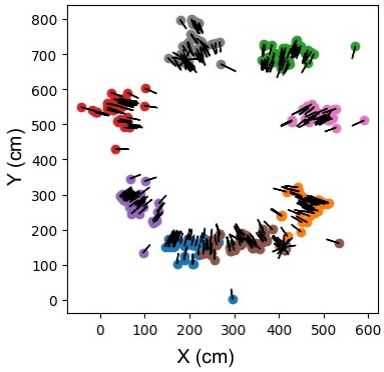

Figure 5. Camera coordinates estimated from COLMAP.

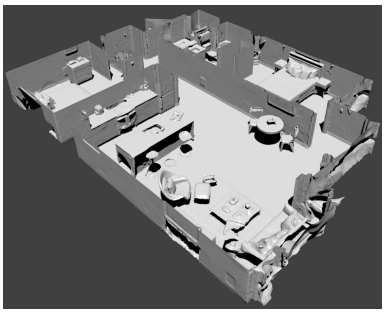

Figure 6. Environment mesh.

of all videos and obtain the corresponding near-range audio clips and bounding boxes for each speaker. As described in Sec. 5.2, we select the active speaker based on the maximum energy of near-range audio with $\Delta t = 0.2$. For a one-second video clip, we obtain 5 candidate bounding boxes. We choose a threshold $\delta\%$ and only keep clips where more than $\delta\%$ of the bounding boxes belong to the same person. We set δ to 80. In this way, we keep clips where there is one main speaker talking, and this speaker's bounding box is used as the localization feature V_L.

8.3. SoundSpaces-NVAS Dataset Details

For the single environment experiment, we use an apartment environment from the Gibson dataset [58][3]. Fig. 6 shows the mesh of the environment (the ceiling is removed). For the novel environment experiment, we use the public train/val/test splits.

For all images, we render with a resolution of 256×256 and a field of view of 120 degrees. We render binaural audio at a sample rate of 16000.

8.4. Implementation and Training Details

All audio clips during training are one second long with a sample rate of 16000. The shape of A_S and A_T is thus 2×16000. The audio encoder is a conv1d layer that encodes audio from 2 channel (binaural) to latent features of

[3] http://gibsonenv.stanford.edu/models/?id=Oyens

64 channels, i.e., A_F^k is of shape 64×16000. For acoustic synthesis, we have $M = 30$ gated multi-modal fusion layers, which are equally divided into 3 blocks. In each block, the dilation of the dilated conv1d increases exponentially with base 3. The kernel size for each dilated conv1d is also 3. Both the skip and residual layers are conv1d layers with kernel size 1. The decoder network is a conv1d layer that encodes the latent audio features from 64 channels back to 2 channels.

The image resolutions are downsampled to 216×384 and 256×256 for Replay-NVAS (downsampled) and SoundSpaces-NVAS respectively. After being processed by a cond1d layer and flattened, the output visual feature V_F is of size 672 for Replay-NVAS and 512 for SoundSpaces-NVAS. The fusion layer consists of two fully connected layers with the first output dimension being 512 and the second being 256.

We train all models for 1000 epochs on the SoundSpaces-NVAS dataset and for 600 epochs on the Replay-NVAS dataset with a learning rate of 0.001. We evaluate the checkpoint with the lowest validation loss on the test set.

8.5. Baseline Details

For the Digital Signal Processing (DSP) baseline, we use the head-related transfer function (HRTF) measured by a KEMAR Dummy-Head Binaural Microphone. We apply a Wiener filter [56] to estimate the inverse HRTF. We adjust the gain of the HRTF by performing a binary search on the validation dataset and selecting the best gain value for testing. For the VAM [6] baseline, we take the original model from the paper, and we make minimal modifications by concatenating the visual feature with the target viewpoint pose P_T. We train the model with the same hyper-parameters described in the paper until convergence on both datasets.