Visually-Guided Audio Spatialization in Video with Geometry-Aware Multi-Task Learning

Rishabh Garg[1*], Ruohan Gao[2] and Kristen Grauman[1,3]

[1*]The University of Texas at Austin, Austin, TX, USA.
[2]Stanford University, Palo Alto, CA, USA.
[3]Facebook AI Research, Austin, TX, USA.

*Corresponding author(s). E-mail(s): rishabhg@utexas.edu;
Contributing authors: rhgao@cs.stanford.edu; grauman@cs.utexas.edu;

Abstract

Binaural audio provides human listeners with an immersive spatial sound experience, but most existing videos lack binaural audio recordings. We propose an audio spatialization method that draws on visual information in videos to convert their monaural (single-channel) audio to binaural audio. Whereas existing approaches leverage visual features extracted directly from video frames, our approach explicitly disentangles the geometric cues present in the visual stream to guide the learning process. In particular, we develop a multi-task framework that learns geometry-aware features for binaural audio generation by accounting for the underlying room impulse response, the visual stream's coherence with the sound source(s) positions, and the consistency in geometry of the sounding objects over time. Furthermore, we introduce two new large video datasets: one with realistic binaural audio simulated for real-world scanned environments, and the other with pseudo-binaural audio obtained from ambisonic sounds in YouTube 360° videos. On three datasets, we demonstrate the efficacy of our method, which achieves state-of-the-art results.

Keywords: audio spatialization, binaural audio generation, video, audio-visual, multi-task learning

1 Introduction

Both sight and sound are key drivers of the human perceptual experience, and both convey essential spatial information. For example, a car driving past us is audible—and spatially trackable—even before it crosses our field of view; a bird singing high in the trees helps us spot it with binoculars; a chamber music quartet performance sounds spatially rich, with the instruments' layout on stage affecting our listening experience.

Spatial hearing is possible thanks to the *binaural* audio received by our two ears. The Interaural Level Difference (ILD) and the Interaural Time Difference (ITD) between the sounds reaching each ear, as well as the shape of the outer ears themselves, all provide spatial effects (Rayleigh, 1875). Meanwhile, the reflections and reverberations of sound in the environment are a function of the room acoustics—the geometry of the room, its major surfaces, and their materials. For example, we perceive the same audio differently in a long corridor versus a large room, or a room with heavy carpet versus a smooth marble floor.

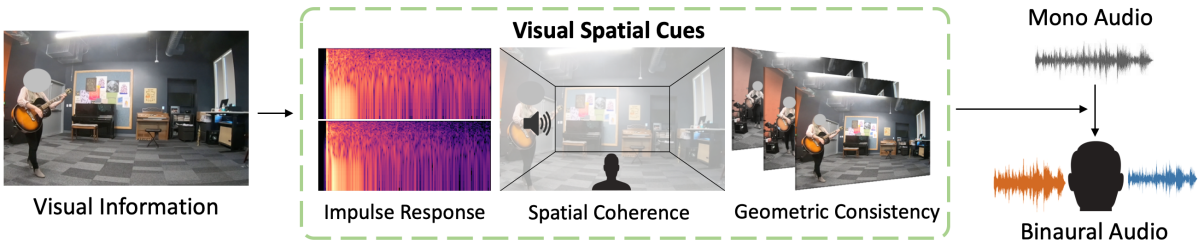


Fig. 1: To generate accurate binaural audio from monaural audio, the visuals provide significant cues that can be learnt jointly with audio prediction. Our approach learns to extract spatial information (e.g., the guitar player is on the left), geometric consistency of the position of the sound sources over time, and cues from the inferred binaural impulse response from the surrounding room.

Videos or other media with binaural audio imitate that rich audio experience for a user, making the media feel more real and immersive. This immersion is important for virtual reality and augmented reality applications, where the user should feel transported to another place and perceive it as such. However, collecting binaural audio data is a challenge. Presently, spatial audio is collected with an array of microphones or specialized dummy rig that imitates the human ears and head. The collection process is therefore less accessible and more costly compared to standard single-channel monaural audio captured with ease from today's ubiquitous mobile devices.

Recent work explores how monaural audio can be upgraded to binaural audio by leveraging the *visual* stream in videos (Gao & Grauman, 2019a; Morgado et al., 2018; H. Zhou et al., 2020). The premise is that the visual context provides hints for how to spatialize the sound due to the visible sounding objects and room geometry. While inspiring, existing models are nonetheless limited to extracting generic visual cues that only implicitly infer spatial characteristics.

Our idea is to explicitly model the spatial phenomena in video that influence the associated binaural sound. Going beyond generic visual features, our approach guides binauralization with those geometric cues from the object and environment that dictate how a listener receives the sound in the real world. In particular, we introduce a multi-task learning framework that accounts for three key factors (Fig. 1). First, we require the visual features to be predictive of the *room impulse response* (RIR), which is the transfer function between the sound sources, 3D environment, and camera/microphone position. Second, we require the visual features to be *spatially coherent* with the sound, i.e., they can understand the difference when audio is aligned with the visuals and when it is not. Third, we enforce the *geometric consistency* of objects over time in the video. Whereas existing methods treat audio and visual frame pairs as independent samples, our approach represents the spatio-temporal smoothness of objects in video, which generally do not have dramatic instantaneous changes in their layout.

The main contributions of this work are as follows. Firstly, we propose a novel multi-task approach to convert a video's monaural sound to binaural sound by learning audio-visual representations that leverage geometric characteristics of the environment and the spatial and temporal cues from videos. Second, to facilitate binauralization research, we create two new datasets: 1) SimBinaural, a large-scale dataset of simulated videos with binaural sound in photo-realistic 3D indoor scene environments, and 2) YouTube-Binaural, a video dataset of pseudo-binaural audio obtained by utilizing the information provided by ambisonic sounds in an existing collection of YouTube 360° videos (Morgado et al., 2018). Both datasets promote deeper experimentation for this problem, facilitating both learning and quantitative evaluation, while the new simulated dataset allows us to explore the impact of particular parameters in a controlled manner and support learning in real videos.

We show the efficacy of our method via extensive experiments in generating realistic binaural audio, achieving state-of-the-art results. We also show that simulated audio data can further improve learning and performance in

real videos. Supplementary video with examples of the datasets and results is available at: https://vision.cs.utexas.edu/projects/visually-guided-multitask-spatial.

2 Related Work

Visually-Guided Audio Spatialization

Recent work uses video frames to provide a form of self-supervision to implicitly infer the relative positions of sound-making objects. They formulate the problem as an upmixing task from mono to binaural using the visual information. Morgado et al. (2018) use 360 videos from YouTube to predict first order ambisonic sound useful for 360 viewing, while Lu et al. (2019) use a self-supervised audio spatialization network using visual frames and optical flow. Lu et al. (2019) use correspondence to learn audio synthesizer ratio masks, which does not necessitate understanding of sound making objects. In contrast, we enforce understanding of the sound location via spatial coherence in the visual features. For speech synthesis, using the ground truth position and orientation of the source and receiver instead of a video is also explored (Richard et al., 2021).

More closely related to our problem, the 2.5D visual sound approach of Gao and Grauman (2019a) generates binaural audio from video. Building on those ideas, H. Zhou et al. (2020) propose an associative pyramid network (APNet) architecture to fuse the modalities and jointly train on audio spatialization and source separation task. Xu et al. (2021) propose to generate binaural audio for training from mono audio by using spherical harmonics. In contrast to these methods, we explore a novel framework for learning geometric representations, and we introduce a large-scale photo-realistic simulated video dataset with acoustically accurate binaural information along with an in-the-wild video dataset augmented with pseudo-binaural sound (both of which will be shared publicly). We outperform the existing methods and show that the new datasets can be used to augment performance.

Audio and 3D Spaces

Recent work exploits the complementary nature of audio and the characteristics of the environment in which it is heard or recorded. Prior methods estimate the acoustic properties of materials (Schissler et al., 2017), estimate reverberation time and equalization of the room using an actual 3D model of a room (Tang et al., 2020), and learn audio-visual correspondence from video (C. Chen et al., 2022; Yang et al., 2020). C. Chen, Jain, et al. (2020) introduce the SoundSpaces audio platform to perform audio-visual navigation in scanned 3D environments, using binaural audio to guide policy learning. Ongoing work continues to explore audio-visual navigation models for embodied agents (C. Chen et al., 2021; C. Chen, Majumder, et al., 2020; Dean et al., 2020; Gan, Zhang, et al., 2020; Majumder et al., 2021; Majumder & Grauman, 2022). Other work predicts depth maps (Christensen et al., 2020) or floorplans (Purushwalkam et al., 2021) using spatial audio or learns representations via interaction using echoes recorded in indoor 3D simulated environments (Gao, Chen, et al., 2020). In contrast to all of the above, we are interested in a different problem of generating accurate spatial binaural sound from videos. We do not use it for navigation nor to explicitly estimate information about the environment. Rather, the output of our model is spatial sound to provide a human listener with an immersive audio-visual experience.

Audio-Visual Learning

Audio-visual learning has a long history, and has enjoyed a resurgence in the vision community in recent years.

Cross-modal learning is explored to understand the natural synchronisation between visuals and the audio (Arandjelovic & Zisserman, 2017; Aytar et al., 2016; Owens, Wu, et al., 2016). Audio-visual data is leveraged for audio-visual speech recognition (Chung et al., 2017; Hu et al., 2016; Yu et al., 2020; H. Zhou et al., 2019), audio-visual event classification and localization (C. Chen et al., 2022; Gao, Oh, et al., 2020; Tian et al., 2020, 2018; Wu et al., 2019) sound source localization (Arandjelović & Zisserman, 2018; Hu et al., 2020; Rouditchenko et al., 2019; Senocak et al., 2018; Tian et al., 2018), self-supervised representation learning (Gao, Chen, et al., 2020; Korbar et al., 2018; Morgado et al., 2020; Owens & Efros, 2018; Owens, Wu, et al., 2016), generating sounds from video (P. Chen et al., 2020; Gan, Huang, Chen, et al., 2020; Owens, Isola, et

al., 2016; Y. Zhou et al., 2018), and audio-visual source separation for speech (Afouras et al., 2019; Ephrat et al., 2018; Gabbay et al., 2018; Gao & Grauman, 2021; Owens & Efros, 2018), music (Gan, Huang, Zhao, et al., 2020; Gao & Grauman, 2019b; Xu et al., 2019; Zhao et al., 2019, 2018), and objects (Gao et al., 2018; Gao & Grauman, 2019b; Tzinis et al., 2021). In contrast to all these methods, we perform a different task: to produce binaural two-channel audio from a monaural audio clip using a video's visual stream.

Finally, this manuscript builds upon our previous work published in BMVC 2021 (Garg et al., 2021). Specifically, we make the following additional contributions in this work: (i) we propose a new in-the-wild binaural videos dataset and a method to obtain such videos from existing 360° videos (Sec 4.3), (ii) we evaluate our proposed method on this new dataset, and examine its utility for existing data (Sec. 5), (iii) we perform an ablation analysis to study the impact of the different multi-task components (Sec 5), (iv) we perform qualitative and quantitative analysis of the RIR prediction task (Sec 5), (v) we provide additional details about the SimBinaural dataset generation process and visualize various statistics of the data (Sec. 4.2).

3 Approach

Our goal is to generate binaural audio from videos with monaural audio. In this section, we first formally describe the problem (Section 3.1). Then we introduce our proposed multi-task setting (Section 3.2). Next we describe the training and inference method (Section 3.3), and finally we describe the proposed SimBinaural dataset (Section 4.2).

3.1 Problem Formulation

Our objective is to map the monaural sound from a given video to spatial binaural audio. The input video may have one or more sound sources, and neither their positions in the 3D scene nor their positions in the 2D video frame are given.

For a video $\mathcal{V}$ with frames $\{v^1 ... v^T\}$ and monaural audio a^t_M, we aim to predict a two channel binaural audio output $\{a^t_L, a^t_R\}$. Whereas a single-channel audio a^t_M lacks spatial characteristics, two-channel binaural audio $\{a^t_L, a^t_R\}$ conveys two distinct waveforms to the left and right ears separately and hence provides spatial effects to the listener. By coupling the monaural audio with the visual stream, we aim to leverage the spatial cues from the pixels to infer how to spatialize the sound. We first transfer the input audio waveforms into the time-frequency domain using the Short-Time Fourier Transformation (STFT). We aim to predict the binaural audio spectrograms $\{\mathcal{A}^t_L, \mathcal{A}^t_R\}$ from the input mono spectrogram $\mathcal{A}^t_M$, where $\mathcal{A}^t_X = \text{STFT}(a^t_X)$, conditioned on visual features v^t_f from the video frames at time t.

3.2 Geometry-Aware Multi-Task Binauralization Network

Our approach has four main components: the *backbone* for converting mono audio to binaural by injecting the visual information, the *spatial coherence* module that learns the relative alignment of the spatial sound and frame, an *RIR prediction* module that requires the room impulse response to be predictable from the video frames, and the *geometric consistency* module that enforces consistency of objects over time.

Backbone Loss

First, we define the backbone loss within our multi-task framework (Fig. 2, bottom). This backbone network is used to transform the input monaural spectrogram $\mathcal{A}^t_M$ to binaural ones. During training, the mono audio is obtained by averaging the two channels $a^t_M = (a^t_L + a^t_R)/2$ and hence the spatial information is lost. Rather than directly predict the two channels of binaural output, we predict the *difference* of the two channels, following Gao and Grauman (2019a). This better captures the subtle distinction of the channels and avoids collapse to the easy case of predicting the same output for both channels. We predict a complex mask M^t_D, which, multiplied with the original audio spectrogram $\mathcal{A}^t_M$, gives the predicted difference spectrogram $\mathcal{A}^t_{D(pred)} = M^t_D \cdot \mathcal{A}^t_M$. The true difference spectrogram of the training input $\mathcal{A}^t_D$ is the STFT of $a^t_L - a^t_R$. We minimize the distance between these two spectrograms: $\|\mathcal{A}^t_D - \mathcal{A}^t_{D(pred)}\|^2_2$. We also predict the two channels via two complex masks M^t_L and M^t_R, one for each

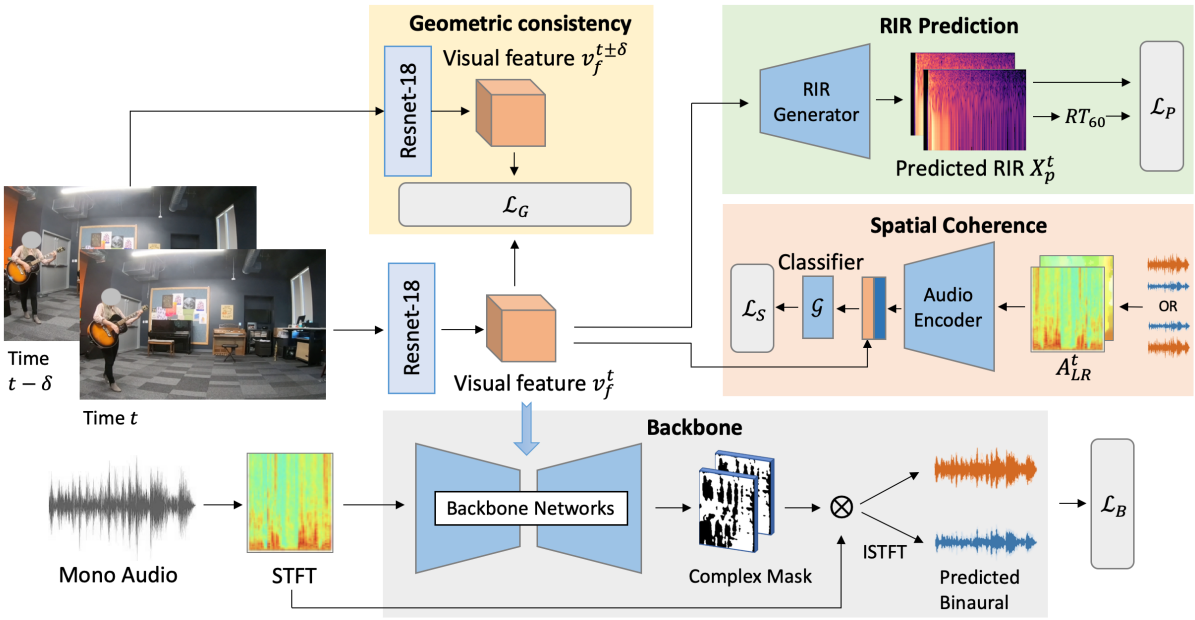


Fig. 2: Proposed network. The network takes the visual frames and monaural audio as input. The ResNet-18 visual features v_f^t are trained in a multi-task setting. The features v_f^t are used to directly predict the RIR via a decoder (top right). Audio features from binaural audio, which might have flipped channels, are combined with v_f^t and used to train a spatial coherence classifier $\mathcal{G}$ (middle right). Two temporally adjacent frames are also used to ensure geometric consistency (top center). The features v_f^t are jointly trained with the backbone network (bottom) to predict the final binaural audio output.

channel, to obtain the predicted channel spectrograms $\mathcal{A}^t_{L(pred)}$ and $\mathcal{A}^t_{R(pred)}$ like above. This gives us the overall backbone objective:

$$\mathcal{L}_B = \|\mathcal{A}^t_D - \mathcal{A}^t_{D(pred)}\|_2^2 + \Big\{\|\mathcal{A}^t_L - \mathcal{A}^t_{L(pred)}\|_2^2 + \|\mathcal{A}^t_R - \mathcal{A}^t_{R(pred)}\|_2^2\Big\}. \quad (1)$$

Spatial Coherence

We encourage the visual features to have geometric understanding of the relative positions of the sound source and receiver via an audio-visual feature alignment prediction term. This loss requires the predicted audio to correctly capture which channel is left and right with respect to the visual information. This is crucial to achieve the proper spatial effect while watching videos, as the audio needs to match the observed visuals' layout.

In particular, we incorporate a classifier to identify whether the visual input is aligned with the audio. The classifier $\mathcal{G}$ combines the binaural audio $\mathcal{A}_{LR} = \{\mathcal{A}^t_L, \mathcal{A}^t_R\}$ and the visual features v_f^t to classify if the audio and visuals agree. In this way, the visual features are forced to reason about the relative positions of the sound sources and learn to find the cues in the visual frames which dictate the direction of sound heard. During training, the original ground truth samples are aligned. We create misaligned samples by flipping the two channels in the ground truth audio to get $\mathcal{A}_{LR} = \{\mathcal{A}^t_R, \mathcal{A}^t_L\}$. We calculate the binary cross entropy (BCE) loss for the classifier's prediction of whether the audio is flipped or not, $c = \mathcal{G}(\mathcal{A}_{LR}, v_f^t)$, and the indicator $\hat{c}$ denoting if the audio is flipped, yielding the spatial coherence loss:

$$\mathcal{L}_G = \text{BCE}(\mathcal{G}(\mathcal{A}_{LR}, v_f^t),\ \hat{c}). \quad (2)$$

Room Impulse Response and Reverberation Time Prediction

The third component of our multi-task model trains the visual features to be predictive of the room impulse response (RIR). An impulse response gives a concise acoustic description of the environment, consisting of the initial direct sound,

the early reflections from the surfaces of the room, and a reverberant tail from the subsequent higher order reflections between the source and receiver. The visual frames convey information like the layout of the room and the sound source with respect to the receiver, which in part form the basis of the RIR. Since we want our audio-visual feature to be a latent representation of the geometry of the room and the source-receiver position pair, we introduce an auxiliary task to predict the room IR directly from the visual frames via a generator on the visual features.

Furthermore, we require the features to be predictive of the *reverberation time* RT_{60} metric, which is the time it takes the energy of the impulse to decay 60dB, and can be calculated from the energy decay curve of the IR (Schroeder, 1965). The RT_{60} is commonly used to characterize the sound properties of a room; we employ it as a low-dimensional target here to guide feature learning alongside the high-dimensional RIR spectrogram prediction.

We convert the ground truth binaural impulse response signal $\{r_L, r_R\}$ to the frequency domain using the STFT and obtain magnitude spectrograms $\mathcal{X}$ for each channel. The IR prediction network consists of a generator which performs upconvolutions on the visual features v_f^t to obtain a predicted magnitude spectrogram $\mathcal{X}^t_{(pred)}$. We minimize the euclidean distance between the predicted RIR $\mathcal{X}^t_{(pred)}$, and the ground truth $\mathcal{X}^t_{gt}$. Additionally, we obtain the RIR waveform from the predicted spectrogram $\mathcal{X}^t_{(pred)}$ via the Griffin-Lim algorithm (Griffin & Lim, 1984; Perraudin et al., 2013) and compute the $RT_{60(pred)}$. We minimize the L1 distance between the predicted $RT_{60(pred)}$ and the ground truth $RT_{60(gt)}$. Thus, the overall RIR prediction loss is:

$$\mathcal{L}_P = \|\mathcal{X}^t_{(pred)} - \mathcal{X}^t_{gt}\|_2^2 + |RT_{60(pred)} - RT_{60(gt)}|. \tag{3}$$

Geometric Consistency

Since the videos are continuous samples over time rather than individual frames, our fourth and final loss regularizes the visual features by requiring them to have spatio-temporal geometric consistency. The position of the source(s) of sound and the position of the camera—as well as the physical environment where the video is recorded—do not typically change instantaneously in videos. Therefore, there is a natural coherence between the sound in a video observed at two points that are temporally close. Since visual features are used to condition our binaural prediction, we encourage our visual features to learn a latent representation that is coherent across short intervals of time. Specifically, the visual features v_f^t and $v_f^{t\pm\delta}$ should be relatively similar to each other to produce audio with fairly similar spatial effects. Specifically, the geometric consistency loss is:

$$\mathcal{L}_S = \max(\|v_f^t - v_f^{t\pm\delta}\| - \alpha, 0), \tag{4}$$

where α is the margin allowed between two visual features. We select a random frame ± 1 second from t, so $-1 \leq \delta \leq 1$. This ensures that similar placements of the camera with respect to the audio source should be represented with similar features, while the margin allows room for dissimilarity for the changes due to time. Since the underlying visual features are regularized to be similar, the predicted audio conditioned on these visual features is also encouraged to be temporally consistent.

3.3 Training and Inference

During training, the mono audio is obtained by taking the mean of the two channels of the ground truth audio $a_m^t = (a_L^t + a_R^t)/2$. The visual features v_f^t are reduced in dimension, tiled, and concatenated with the output of the audio encoder to fuse the information from the audio and visual streams. The overall multi-task loss is a combination of the losses (Equations 1-4) described earlier:

$$\mathcal{L} = \lambda_B \mathcal{L}_B + \lambda_S \mathcal{L}_S + \lambda_G \mathcal{L}_G + \lambda_P \mathcal{L}_P, \tag{5}$$

where λ_B, λ_S, λ_G and λ_P are the scalar weights used to determine the effect of each loss during training, set using validation data.

To generate audio at test time, we only require the mono audio and visual frames. The predicted spectrograms are used to obtain the predicted difference signal $a^t_{D(pred)}$ and two-channel audio $\{a_L^t, a_R^t\}$ via an inverse Short-Time Fourier Transformation (ISTFT) operation.

4 Video Datasets with Binaural Audio

In this section, we describe the three video datasets with binaural audio that we use for evaluation of our method. We first describe FAIR-Play (Gao & Grauman, 2019a) (Sec. 4.1), an existing video dataset with binaural audio collected in a music room. Then, we introduce two *new* datasets to facilitate both learning and quantitative evaluation: SimBinaural (Sec. 4.2), a large-scale dataset of simulated videos with binaural sound in photo-realistic 3D indoor scene environments, and YouTube-Binaural (Sec. 4.3), a video dataset of pseudo-binaural audio obtained by utilizing the information provided by ambisonic sounds in YouTube 360° videos.

4.1 FAIR-Play

The FAIR-Play dataset collected by (Gao & Grauman, 2019a) is a large public video dataset with binaural audio, and is widely used for the task of visually-guided audio spatialization for video. It was recorded using a binaural microphone rig, composed of an ear shaped housing mounted on top of a video camera. The videos were captured in a large music room, with various combinations of instruments and people in different spatial contexts within the room. The videos consist of recordings of people playing instruments like cello, guitar, drum, piano etc. and are composed of solo, duet, and multi-player performances. The dataset totals 5.2 hours of video, which are broken into 1,871 10-second clips.

4.2 SimBinaural Dataset

We also experiment with video from scanned environments with high quality simulated audio. To facilitate large-scale experimentation—and to augment learning from real videos—we create a new dataset called SimBinaural of simulated videos in photo-realistic 3D indoor scene environments.[1] The generated videos, totaling over 100 hours, resemble real-world audio recordings and are sampled from 1,020 rooms in 80 distinct environments; each environment is a multi-room home. Using the publicly available SoundSpaces[2] audio simulations (C. Chen, Jain, et al., 2020) together with the Habitat simulator (Savva et al., 2019), we create realistic videos with binaural sounds for publicly available 3D environments in Matterport3D (Chang et al., 2017) (Fig. 3). Our resulting SimBinaural dataset is much larger and more diverse than the FAIR-Play dataset (Gao & Grauman, 2019a) which contains real videos but is limited to 5 hours of recordings in one room (Table 1).

To construct the dataset, we insert diverse 3D models from `poly.google.com` of various instruments like the guitar, violin, flute etc. and other sound-making objects like phones and clocks into the scene. Each object has multiple models of that type for diversity, so we do not associate a sound with a particular 3D model. We have a total of 35 objects from 11 classes. To generate realistic binaural sound in the environment as if it is coming from the source location and heard at the camera position, we convolve the appropriate SoundSpaces (C. Chen, Jain, et al., 2020) room impulse response (RIR) with an anechoic audio waveform (e.g., a guitar playing for an inserted guitar 3D object). We use sounds recorded in anechoic environments as input to the SoundSpaces RIRs, so that there are no existing reverberations to affect the data. The sounds are obtained from Freesound (Font et al., 2013) and OpenAIR data (Murphy & Shelley, 2010) to form a set of 127 different sound clips spanning the 11 distinct object categories.

Using this setup, we capture videos with simulated binaural sound just like in the real world using an agent equipped with a virtual camera and binaural microphone. The virtual camera and attached microphones are moved along trajectories such that the object remains in view, leading to diversity in views of the object and locations within each video clip (see Fig. 4). While generating a video, we use a fixed sound source position and the agent traverses a random path. Since the camera moves and rotates throughout the video,

[1] The SimBinaural dataset was constructed at, and will be released by, The University of Texas at Austin.

[2] SoundSpaces (C. Chen, Jain, et al., 2020) provides room impulse responses at a spatial resolution of 1 meter. These state-of-the-art RIRs capture how sound from each source propagates and interacts with the surrounding geometry and materials, modeling all the major real-world features of the RIR: direct sounds, early specular/diffuse reflections, reverberations, binaural spatialization, and frequency dependent effects from materials and air absorption.

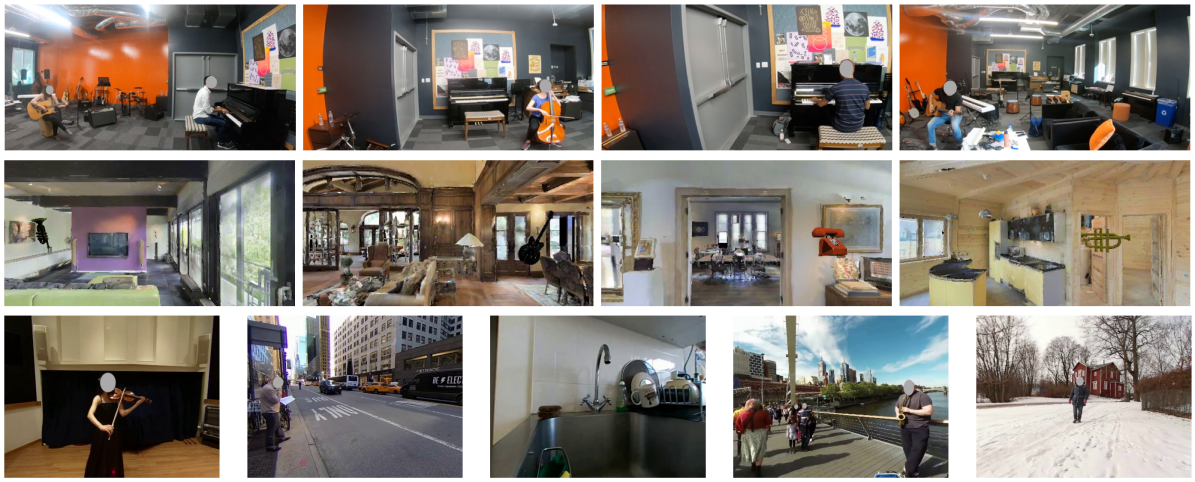


Fig. 3: Example frames from the FAIR-Play (top), SimBinaural (middle), and YouTube-Binaural (bottom) datasets. See supplementary video for audio-visual examples.

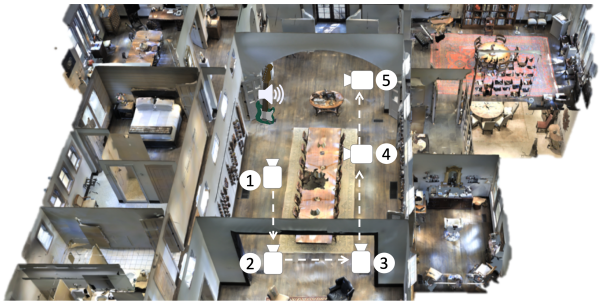


Fig. 4: The video generation process for SimBinaural. An object and corresponding audio is placed in the scene. A virtual camera traverses a trajectory in the room such that the object is seen from different distances and viewpoints.

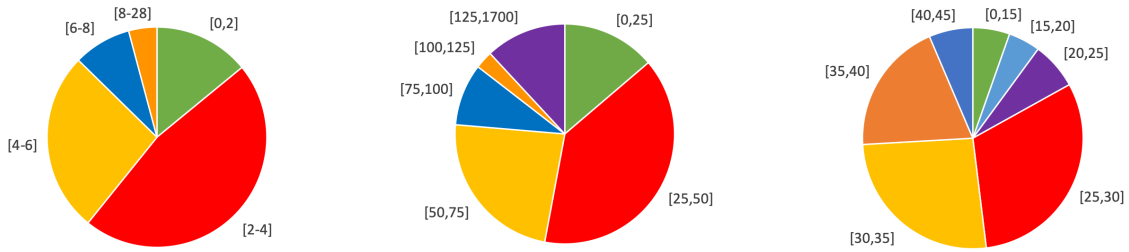
the view of the object changes resulting in various orientations and positions of the object within a video frame, for each video. We ensure that the object is in view of the camera using ray tracing, and the source positions are densely sampled from the 3D environments. The camera moves to a new position every 5 seconds and has a small translational motion during the five-second interval. The videos are generated at 5 frames per second, the average length of the videos in the dataset is 19.2s and the median length is 15s. Please see the supplementary video for examples.

The resulting data has a lot of inter-video and intra-video diversity. The objects are at various distances from the camera throughout the videos (Fig. 5a) and the distance changes within a video. The rooms in which the videos are captured are of very different sizes (Fig. 5b) and hence have different acoustic properties. The position of sources of sound with respect to the receiver also varies across the videos; Fig. 5c shows the distribution of the angle the object makes from the center of the frame.

4.3 YouTube-Binaural Dataset

To experiment with another source of real-world in-the-wild data from multiple scenes, we next augment an existing dataset of YouTube videos (Morgado et al., 2018). The dataset contains 360° videos and ambisonic omnidirectional audio. We augment it in two ways: 1) we transform the ambisonic audio into (pseudo)-binaural audio, and 2) we transform each 360° clip into a normal field-of-view clip in which the camera faces the prominent sound source.

Previous work using 360° video for the binauralization task uses the full equirectangular frame and converts the ambisonic audio to binaural directly (Gao & Grauman, 2019a; H. Zhou et al., 2020). However, the equirectangular frame offers a distorted view of the whole space and does not focus on sound sources. Additionally, the ground truth binaural obtained from ambisonics does not take into account the location of the sound sources, implicitly assuming that the visual



(a) Distance of the sound (m) (b) Size of the room (m^2) (c) Angle from the center (degrees)

Fig. 5: Statistics of the distance between the sound source and the receiver, the size of the room, and the angle of the object from the center of the frame for the SimBinaural dataset. This illustrates the diversity in the data in different aspects, both within a video, and across the dataset.

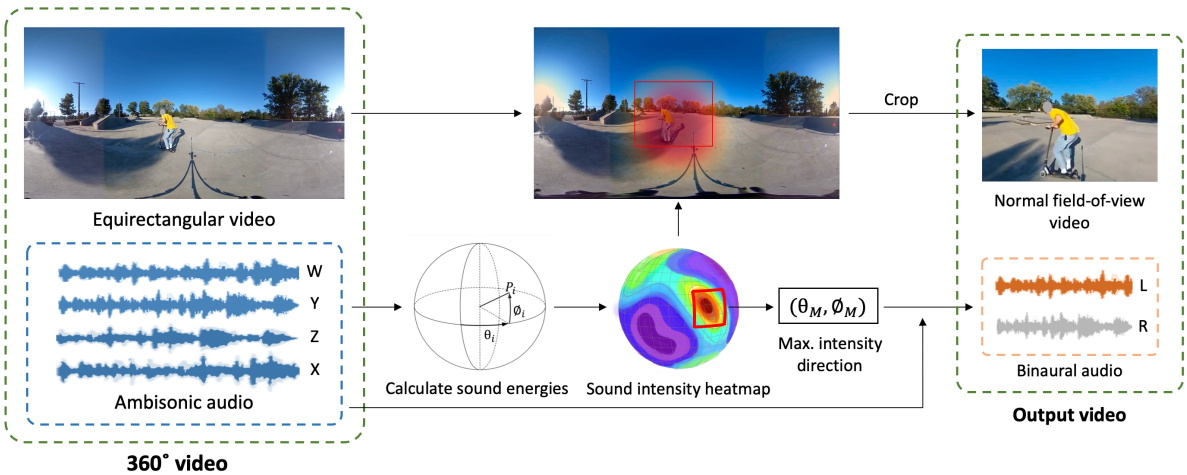


Fig. 6: The video processing pipeline to create the YouTube-Binaural dataset. We use a 360° video with ambisonic audio (from the existing YT-Clean dataset (Morgado et al., 2020)) to obtain a normal field-of-video video with binaural audio.

object(s) causing the sound are centered in the equirectangular projection of the frame.

Our insight is that the direction from which the primary sounds reach the 360° camera is also the direction where the visual stream is typically most relevant. For example, in a 360° video where a boat passes by, the direction of the noisy engine agrees with where one sees the visual evidence of the boat. We use this property to compose a normal field-of-view (perspective projection) video dataset relevant for our binauralization task, as follows.

We first use the actual ambisonic audio to calculate the audio intensities at densely sampled points $P_i = (\theta_i, \phi_i)$ in the full 360° view (Fig. 6). We use these points to obtain a spherical heatmap of the audio energies. From this audio energy map, we determine a rectangular region of size $25° \times 25°$, which has the highest average sound energy across the map. Since this is a 360° video, we should be able to see most sound sources somewhere in the frame if they are in view of the camera; the region we select represents the direction with maximum intensity and hence likelihood of a sound source of interest.[3]

[3]This is the typical case. However, there can be instances where a sound source is not visualized in the video at all; for example, if music is playing from a small radio and the radio is

Denote this direction by the angles (ϕ_M, θ_M) from the center of the frame in the spherical coordinates. We then crop the equirectangular video in this direction (ϕ_M, θ_M) to a normal field-of-view so that the resulting video will have the sound making region in view. To preserve the integrity of the scene geometry, we use a perspective projection of the cropped video onto a planar region tangent to the sphere, as opposed to simply cropping from the equirectangular projection (see Fig. 7 and 6).

To obtain the accompanying binaural audio, we convert the four-channel ambisonic audio to binaural audio such that the listener is facing the direction (ϕ_M, θ_M), which provides a realistic sound when paired with the normal field-of-view video. The ambisonic signal is convolved with a head-related transfer function (HRTF) along the time dimension and summed to produce the audio for each ear (Zaunschirm et al., 2018). Since the HRTF has to be truncated to the spherical harmonic order of the ambisonics (first order in our case), it results in some approximations in the generated binaural audio, as compared to that captured by a professional binaural microphone, as in FAIR-Play (Gao & Grauman, 2019a).

We stress that no matter what viewing angle is selected using the audio intensity heatmap above, the binaural audio we compute to pair with it is appropriate. That is, even if the viewpoint selection heuristic does not display the sounding object, it is still a relevant training/testing instance because the accompanying sound is spatialized for that same viewpoint.

For the 360° videos, we use the YT-Clean (Morgado et al., 2018) dataset, which contains in-the-wild YouTube videos collected by querying for terms related to spatial audio and consists of videos which have four-channel ambisonic audio. Most have a small number of super-imposed sources like people talking in a room, a person playing an instrument etc. which can be localized in the image. The resulting YouTube-Binaural dataset has 426 videos, totaling over 27 hours of normal field-of-view video and the corresponding binaural audio.

not visible to the camera. While the binaural data generated in such cases is still correct, it might be harder for any model (including ours) to learn from such samples. Empirically, the number of such clips forms a very small portion of the data.

Fig. 7: Examples of cropping the equirectangular frame. Direct cropping can lead to significant distortions. In contrast, a correct crop first projects the image to a plane tangent to the sphere centered at the target direction.

Table 1 offers a comparison of all three datasets. FAIR-Play (Gao & Grauman, 2019a) consists of real video, with audio recorded with a binaural mic rig. This dataset has over 5 hours of such video, but is limited to just one room and musical instruments. SimBinaural on the other hand is a large-scale dataset with over 100 hours of binaural audio and video in over 1,000 different rooms, and it includes RIR data as well. However, it is generated via a simulator and hence lacks the realism of FAIR-Play. YouTube-Binaural contains real video and is much larger (27 hours of video) and more diverse than FAIR-Play as it is composed of in-the-wild video in various scenes. However, unlike the other two datasets, the audio in YouTube-Binaural is derived from ambisonics, as opposed to directly sensed binaural sound. The datasets are available on the project page: https://vision.cs.utexas.edu/projects/visually-guided-multitask-spatial.

5 Experiments

We validate our approach on the FAIR-Play (Gao & Grauman, 2019a) (the existing real video benchmark), our new SimBinaural dataset, and the augmented YouTube-Binaural data. We compare to the following baselines:

- **Flipped-Visual:** We flip the visual frame horizontally to provide incorrect visual information while testing. Incorrect visual information ought to be a disadvantage if the visual frame is significant for our results. The other settings are the same as our method.

Dataset	#Videos	Length (hrs)	#Rooms	RIR
FAIR-Play (Gao & Grauman, 2019a)	1,871	5.2	1	No
SimBinaural	21,737	116.1	1,020	Yes
YouTube-Binaural	426	27.7	n/a	No

Table 1: A comparison of the data in FAIR-Play and the two additional datasets presented in this work.

- **Audio Only:** We provide only monaural audio as input, with no visual frames, to verify if the visual information is essential to learning.
- **Mono-Mono:** Both channels have the same input monaural audio repeated as the two-channel output to verify if we are actually distinguishing between the channels.
- **Mono2Binaural** (Gao & Grauman, 2019a): A state-of-the-art 2.5D visual sound model for this task. We use the authors' code to evaluate in the settings as ours.
- **APNet** (H. Zhou et al., 2020): A state-of-the-art model that handles both binauralization and audio source separation. We use the APNet network from their method and train only on binaural data (rather than stereo audio). We use the authors' code.
- **PseudoBinaural** (Xu et al., 2021): A state-of-the-art model that uses additional data to augment training. We use the authors' public pre-trained model.

We evaluate two standard metrics, following Gao and Grauman (2019a); Morgado et al. (2018); H. Zhou et al. (2020):

- **STFT Distance**: The euclidean distance between the predicted and ground truth STFT spectrograms, which directly measures how accurate is our produced spectrogram

$$D^{STFT} = \|A_L^t - A_{L(pred)}^t\|_2 + \|A_R^t - A_{R(pred)}^t\|_2.$$

- **Envelope Distance** (ENV): Perceptual similarity cannot be captured well by direct comparison of raw waveforms. The envelope distance metric measures the euclidean distance between the envelopes E_L^t of the predicted raw audio signal a_L^t and the ground truth and is defined as

$$D^{ENV} = \|E_L^t - E_{L(pred)}^t\|_2 + \|E_R^t - E_{R(pred)}^t\|_2.$$

Implementation details

All networks are written in PyTorch (Paszke et al., 2019). The backbone network is based upon the networks used for 2.5D visual sound (Gao & Grauman, 2019a) and APNet (H. Zhou et al., 2020). The visual network is a ResNet-18 (He et al., 2016) with the pooling and fully connected layers removed. The audio network consists of a U-Net (Ronneberger et al., 2015) type architecture. The U-Net consists of 5 convolution layers for downsampling and 5 upconvolution layers in the upsampling network and include skip connections. The encoder for spatial coherence follows the same architecture as the U-Net encoder for the audio feature extraction. The classifier combines the audio and visual features and uses a fully connected layer for prediction. The generator network is adapted from GANSynth (Engel et al., 2019), modified to fit the required dimensions of the audio spectrogram.

To preprocess all datasets, we follow the standard steps from (Gao & Grauman, 2019a). We resampled all the audio to 16kHz and computed the STFT using a FFT size of 512, window size of 400, and hop length of 160. For training the backbone, we use 0.63s clips of the audio and the corresponding frame. Frames are extracted at 10fps. The visual frames are randomly cropped to 448 $\times$ 224. For testing, we use a sliding window of 0.1s to compute the binaural audio for all methods.

For the YouTube-Binaural dataset, the sound energy intensity maps are calculated for the average intensity every one second. Thus, the maximum sound intensity region directions are computed and a new crop of the video is created in that direction for every second. The videos are cropped at a 90° field-of-view. Similarly, a corresponding binaural audio is computed from the ambisonics in the direction of max energy for that one-second period. The videos are created at 8 frames per second.

	FAIR-Play		SimBinaural				YouTube-Binaural	
			Scene-Split		Position-Split			
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Mono-Mono	1.215	0.157	1.356	0.163	1.348	0.168	4.715	0.261
Audio-Only	1.102	0.145	0.973	0.135	0.932	0.130	3.129	0.213
Flipped-Visual	1.134	0.152	1.082	0.142	1.075	0.141	3.298	0.214
Mono2Binaural (Gao & Grauman, 2019a)	0.927	0.142	0.874	0.129	0.805	0.124	2.892	0.208
APNet (H. Zhou et al., 2020)	0.904	0.138	0.857	0.127	0.773	0.122	2.733	0.204
Backbone+IR Pred	n/a	n/a	0.801	0.124	0.713	0.117	n/a	n/a
Backbone+Spatial	0.873	**0.134**	0.837	0.126	0.756	0.120	2.645	0.201
Backbone+Geom	0.874	0.135	0.828	0.125	0.731	0.118	2.580	0.196
Our Full Model	**0.869**	**0.134**	**0.795**	**0.123**	**0.691**	**0.116**	**2.544**	**0.196**

Table 2: Binaural audio prediction errors on all three datasets. For both metrics, lower is better.

We use the Adam optimizer (Kingma & Ba, 2015) and a batch size of 64. The initial learning rates are 0.001 for the audio and the fusion networks, and 0.0001 for all the other networks. We train the FAIR-Play dataset for 1000 epochs, the SimBinaural dataset for 100 epochs and the YouTube-Binaural dataset for 7000 epochs. We train the RIR prediction separately and use the weights for initialization while training jointly. The δ for choice of frame is set to 1s and the λ's used are set based on validation set performance to $\lambda_B = 10, \lambda_S = 1, \lambda_G = 0.01, \lambda_P = 1$.

SimBinaural results

We evaluate on two data splits: 1) ***Scene-Split***, where the train and test set have disjoint scenes from Matterport3D (Chang et al., 2017) and hence the room of the videos at test time has not been seen before; and 2) ***Position-Split***, where the splits may share the same Matterport3D scene/-room but the exact configuration of the source object and receiver position is not seen before.

Table 2 (center) shows the results. The table also ablates the parts of our model. Our model outperforms all the baselines, including the two state-of-the-art prior methods. In addition, Table 2 confirms that Scene-Split is a fundamentally harder task. This is because we must predict the sound, as well as other characteristics like the IR, from visuals distinct from those we have observed before. This forces the model to generalize its encoding to generic visual properties (wall orientations, major furniture, etc.) that have intra-class variations and geometry changes compared to the training scenes.

Method	STFT	ENV
APNet (H. Zhou et al., 2020)	1.291	0.162
PseudoBinaural (Xu et al., 2021)	1.268	0.161
Ours	1.234	0.160
Ours+SimBinaural	**1.175**	**0.154**

Table 3: Results on FAIR-Play when additional data is used for training.

The ablations shed light on the impact of each of the proposed losses in our multi-task framework. The full model uses all the losses as in Eqn 5. This outperforms other methods significantly on both splits. It also outperforms using each of the losses individually, which demonstrates the losses can combine to jointly learn better visual features for generating spatial audio.

FAIR-Play results

Table 2 (left) shows the results on the real video benchmark FAIR-Play using the standard split. Here, we omit the IR prediction network for our method, since FAIR-Play lacks ground truth impulse responses (which we need for training). The Backbone+Spatial and Backbone+Geom are the same as above. Both variants of our method outperform the state-of-the-art. Therefore, enforcing the geometric and spatial constraints is beneficial to the binaural audio generation task. We get the best results when we combine both the losses in our framework.

To further evaluate the utility of our SimBinaural dataset, we next jointly train with both SimBinaural and FAIR-Play, then test on a challenging split of FAIR-Play in which the test scenes are non overlapping, as proposed by Xu et al.

Method	STFT	ENV
APNet (H. Zhou et al., 2020)	2.733	0.204
Ours	2.544	0.196
Ours+SimBinaural	**2.491**	**0.194**

Table 4: Results on YouTube-Binaural trained along with SimBinaural. We cannot compare to PseudoBinaural (Xu et al., 2021) as their data augmentation method and the available pre-trained model focus on creating binaural audio for music and thus are tailored to FAIR-Play.

(2021). We compare our method with Augment-PseudoBinaural (Xu et al., 2021)[4] which also uses additional generated training data. Our method with SimBinaural outperforms other methods (Table 3). This is an important result, as it demonstrates that SimBinaural can be leveraged to improve performance on *real* video.

YouTube-Binaural results

Finally, we evaluate on our method on the YouTube-Binaural data. Table 2 (right) shows the results on this dataset. Like FAIR-Play, we omit the IR prediction network due to lack of ground truth impulse responses. The augmented binaural data is used for both training and evaluation. Using the additional proposed losses helps improve the results, and our overall method outperforms state-of-the-art methods, showcasing the efficacy of learning visual features on in-the-wild videos as well.

We also jointly train YouTube-Binaural with SimBinaural, and test on the same YouTube-Binaural clips. SimBinaural helps further improve binauralization performance, again demonstrating its utility for real video (Table 4).

User Studies

Next, we present two user studies to validate whether the predicted binaural sound does indeed provide an immersive and spatially accurate experience for human listeners. Twenty participants with normal hearing were presented with 20 videos from the test sets. They were asked to rate the quality in two ways: 1) users were given only the audio and asked to choose from which direction (left/right/center) they heard the audio; 2) given a pair of audios and a reference frame, the users were asked to choose which audio gives a binaural experience closer to the provided ground truth. As can be seen in Fig. 8, users preferred our method both for the accuracy of the direction of sound (left) and binaural audio quality (right).

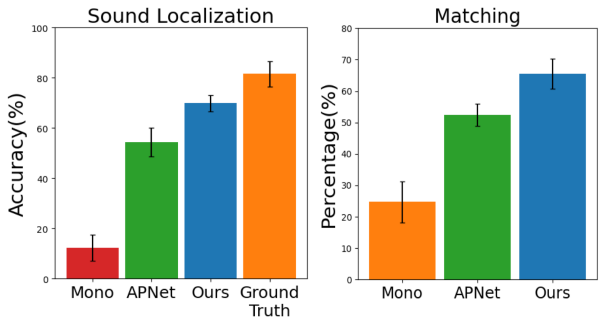


Fig. 8: User study results evaluating whether listeners can tell the correct direction of the sounds (left) and how often they find each model to better match the ground truth (right). See the supplementary video for examples.

Qualitative Visualization

Next we explore qualitatively what the features have learned. Figure 9 shows the t-SNE projections (Van der Maaten & Hinton, 2008) of the visual features from SimBinaural colored by the RT_{60} of the audio clip. While the features from our method (left) can infer the RT_{60} characteristics, the ones from APNet (H. Zhou et al., 2020) (center) are randomly distributed. Simultaneously, our features also accurately capture the angle of the object from the center (right). Fig. 10 shows the activation maps of the visual network. While APNet produces more diffuse maps, our method localizes the object better within the image. This indicates that the visual features in our method are better at identifying the regions which might be emitting sound to generate more accurate binaural audio.

Ablation Study

Table 2 illustrates that adding each component of our method individually to the visual features helps improve the binaural audio quality

[4]The pre-trained model provided by PseudoBinaural (Xu et al., 2021) is trained on a different split instead of the standard split from Gao and Grauman (2019a) and hence it is not directly comparable in Table 2. We evaluate on the new split in Table 3.

Fig. 9: t-SNE of visual features colored by RT_{60} for our method (left) and APNet (H. Zhou et al., 2020) (center); and colored by angle of the object from the center (right). Our method learns a representation that better reflects the RT_{60} characteristics and captures the angle of the sound source.

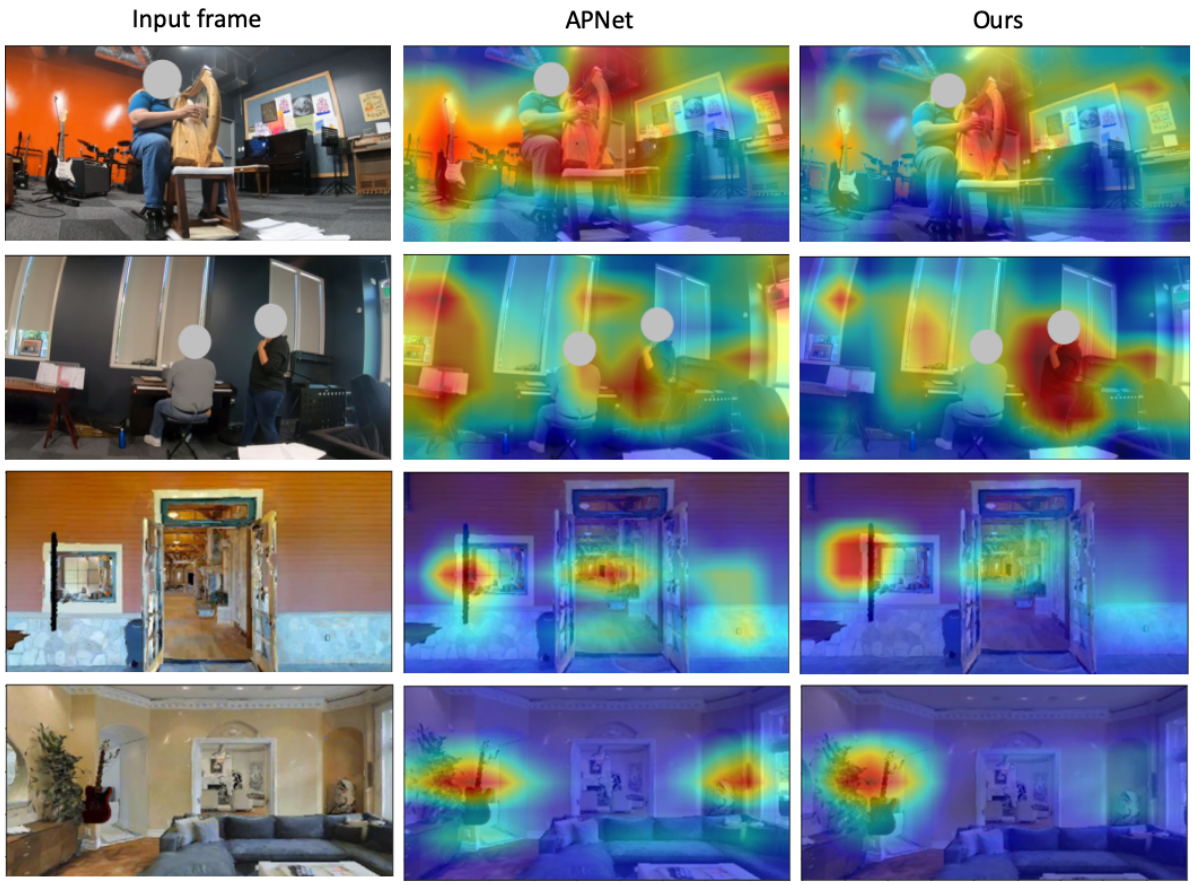


Fig. 10: Qualitative visualization of the activation maps for the visual network for APNet (H. Zhou et al., 2020) and ours. We can see that while the activation maps for APNet (H. Zhou et al., 2020) are diffused and focusing on non-essential parts like objects in the background, our method focuses more on the object/region producing the sound and its location.

Backbone	Spatial	IR Pred.	Geom.	Scene-Split STFT	Scene-Split ENV	Position-Split STFT	Position-Split ENV
✓				0.857	0.127	0.773	0.122
✓	✓			0.837	0.126	0.756	0.120
✓		✓		0.821	0.125	0.725	0.119
✓			✓	0.836	0.126	0.728	0.119
✓	✓	✓		0.804	**0.123**	0.702	0.117
✓	✓		✓	0.817	0.125	0.724	0.118
✓		✓	✓	0.809	0.124	0.707	0.117
✓	✓	✓	✓	**0.795**	**0.123**	**0.691**	**0.116**

Table 5: Ablations for the Scene-Split and Position-Split on SimBinaural with different combinations of constraints.

performance across all datasets. Table 5 provides additional analysis to evaluate the combination of different components for our multi-task framework for the SimBinaural for both Scene-Split and Position-Split. It can be seen that while adding each constraint helps improve performance, when the different components are combined, it improves the ability of the model to incorporate multiple facets of the task illustrating the efficacy of the multi-task formulation. Additionally, adding the IR prediction component to generate an approximate impulse response from the image frame without the audio has the most impact on the models' ability to accurately learn the characteristics of the room and generate more accurate binaural results. The tasks complement each other to learn better visual features, leading to better audio performance.

RIR Prediction Analysis

Finally, we analyze the RIR prediction component of our multi-task learning framework to study the effectiveness of the network for this task. Figure 11 shows qualitative examples of predictions from the test set. It can be seen that we can get a fairly accurate prediction of the IR, and the difference between the response in each channel is also captured.

Quantitatively, we evaluate this by calculating the estimation error for the RT_{60} metric for these predicted IRs (Table 6). For the nearest neighbor baseline, error is calculated with the IR of the closest frame in the training set based on cosine similarity of visual features. Our method has a lower mean error and a smaller standard deviation of errors compared to the nearest neighbor baseline. The low error indicates the ability to predict an IR is reasonably effective and can lead to perceptually sound IRs. We also evaluate this task quantitatively by formulating a classification task of predicting the RT_{60} metric. We discretize the range of the RT_{60} into 10 classes, each with roughly equal number of samples. The classifier has a test accuracy of 61.5% which demonstrates the network's ability to estimate the RT_{60} range from the visual frame quite well.

Method	Mean Err.	Std.Dev Err.
Nearest Neighbor	0.212	0.083
Our Method	**0.182**	**0.064**

Table 6: The error in RT_{60} metric for the RIR. For both, lower values imply better performance

6 Conclusion

We presented a multi-task approach to learn geometry-aware visual features for mono to binaural audio conversion in videos. Our method exploits the inherent room and object geometry and spatial information encoded in the visual frames to generate rich binaural audio. We also generated a large-scale video dataset with binaural audio in photo-realistic environments to better understand and learn the relation between visuals and binaural audio. We also augment an in-the-wild 360 video dataset with pseudo-binaural sound and accompanying normal field-of-view video. Our state-of-the-art results on three datasets demonstrate the efficacy of our proposed formulation.

Despite encouraging results on these three datasets, they have different pros and cons in terms of scale, diversity, and realism as discussed

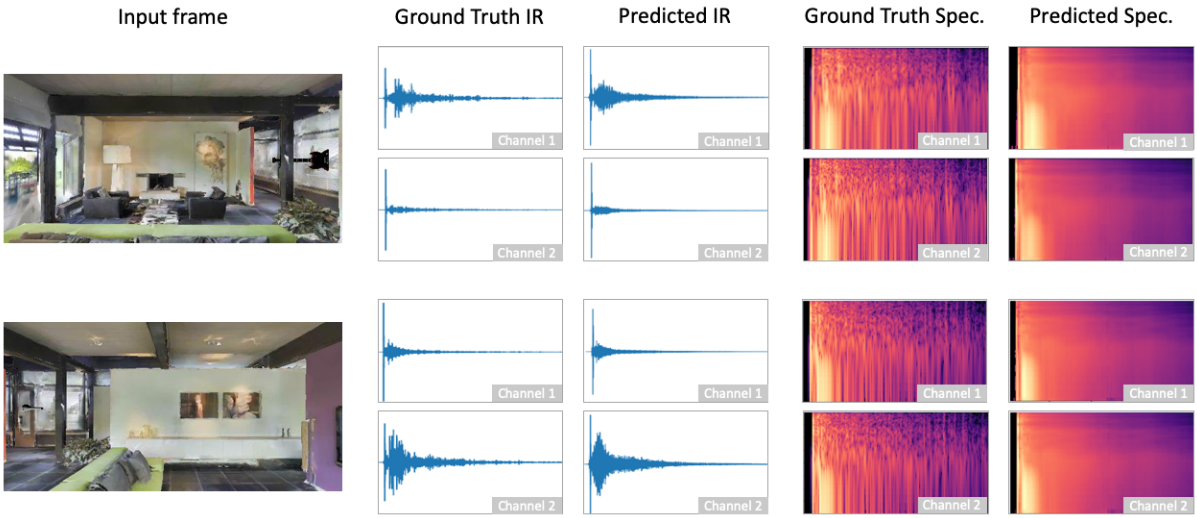


Fig. 11: IR Prediction: The first column is the input frame to the encoder. The second column depicts the ground truth IR for the frame and the fourth column is the corresponding spectrogram of this IR. The third and fifth columns show the predicted IR waveform and spectrogram, respectively. This predicted IR waveform is estimated from the spectrogram generated by our network.

in Sec. 4. It would be interesting future work to either create a crowdsourcing interface for building a large-scale realistic dataset of users wearing binaural microphones performing diverse everyday activities, or a new pipeline that can more intelligently integrate the strengths of these three existing datasets for training models that can generalize to in-the-wild videos in novel domains. Furthermore, we plan to explore how semantic models of object categories' sounds could benefit the spatialization task. We also plan to study the impact of explicitly performing object localisation to improve scene understanding and incorporate that in the binauralization task.

References

Afouras, T., Chung, J.S., Zisserman, A. (2019). My lips are concealed: Audio-visual speech enhancement through obstructions. *ICASSP.*

Arandjelovic, R., & Zisserman, A. (2017). Look, listen and learn. *ICCV.*

Arandjelović, R., & Zisserman, A. (2018). Objects that sound. *ECCV.*

Aytar, Y., Vondrick, C., Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. *NeurIPS.*

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., ... Zhang, Y. (2017). Matterport3d: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV).* (MatterPort3D dataset license available at: http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf)

Chen, C., Al-Halah, Z., Grauman, K. (2021). Semantic audio-visual navigation. *CVPR.*

Chen, C., Gao, R., Calamia, P., Grauman, K. (2022). Visual acoustic matching. *CVPR.*

Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., ... Grauman, K. (2020). Soundspaces: Audio-visual navigation in 3d environments. *ECCV.*

Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S.K., Grauman, K. (2020). Learning to set waypoints for audio-visual navigation. *ICLR.*

Chen, P., Zhang, Y., Tan, M., Xiao, H., Huang, D., Gan, C. (2020). Generating visually aligned sound from videos. *IEEE TIP*.

Christensen, J.H., Hornauer, S., Stella, X.Y. (2020). Batvision: Learning to see 3d spatial layout with two ears. *ICRA*.

Chung, J.S., Senior, A., Vinyals, O., Zisserman, A. (2017). Lip reading sentences in the wild. *CVPR*.

Dean, V., Tulsiani, S., Gupta, A. (2020). See, hear, explore: Curiosity via audio-visual association. *NeurIPS*.

Engel, J., Agrawal, K.K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. *ICLR*.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., ... Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *SIGGRAPH*.

Font, F., Roma, G., Serra, X. (2013). Freesound technical demo. *Proceedings of the 21st ACM International Conference on Multimedia*.

Gabbay, A., Shamir, A., Peleg, S. (2018). Visual speech enhancement. *INTERSPEECH*.

Gan, C., Huang, D., Chen, P., Tenenbaum, J.B., Torralba, A. (2020). Foley music: Learning to generate music from videos. *ECCV*.

Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A. (2020). Music gesture for visual sound separation. *CVPR*.

Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B. (2020). Look, listen, and act: Towards audio-visual embodied navigation. *ICRA*.

Gao, R., Chen, C., Al-Halah, Z., Schissler, C., Grauman, K. (2020). Visualechoes: Spatial image representation learning through echolocation. *ECCV*.

Gao, R., Feris, R., Grauman, K. (2018). Learning to separate object sounds by watching unlabeled video. *ECCV*.

Gao, R., & Grauman, K. (2019a). 2.5d visual sound. *CVPR*.

Gao, R., & Grauman, K. (2019b). Co-separating sounds of visual objects. *ICCV*.

Gao, R., & Grauman, K. (2021). Visualvoice: Audio-visual speech separation with cross-modal consistency. *CVPR*.

Gao, R., Oh, T.-H., Grauman, K., Torresani, L. (2020). Listen to look: Action recognition by previewing audio. *CVPR*.

Garg, R., Gao, R., Grauman, K. (2021). Geometry-aware multi-task learning for binaural audio generation from video. *BMVC*.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *CVPR*.

Hu, D., Li, X., et al. (2016). Temporal multimodal learning in audiovisual speech recognition. *CVPR*.

Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., ... Dou, D. (2020). Discriminative sounding objects localization via self-supervised audiovisual matching. *NeurIPS*.

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

Korbar, B., Tran, D., Torresani, L. (2018). Co-training of audio and video representations from self-supervised temporal synchronization. *NeurIPS*.

Lu, Y.-D., Lee, H.-Y., Tseng, H.-Y., Yang, M.-H. (2019). Self-supervised audio spatialization with correspondence classifier. *ICIP*.

Majumder, S., Al-Halah, Z., Grauman, K. (2021). Move2Hear: Active audio-visual source separation. *ICCV.*

Majumder, S., & Grauman, K. (2022). Active audio-visual separation of dynamic sound sources. *ECCV.*

Morgado, P., Li, Y., Nvasconcelos, N. (2020). Learning representations from audio-visual spatial alignment. *NeurIPS.*

Morgado, P., Vasconcelos, N., Langlois, T., Wang, O. (2018). Self-supervised generation of spatial audio for 360° video. *NeurIPS.*

Murphy, D.T., & Shelley, S. (2010). Openair: An interactive auralization web resource and database. *Audio Engineering Society Convention 129.*

Owens, A., & Efros, A.A. (2018). Audio-visual scene analysis with self-supervised multisensory features. *ECCV.*

Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T. (2016). Visually indicated sounds. *CVPR.*

Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A. (2016). Ambient sound provides supervision for visual learning. *ECCV.*

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *NeurIPS.*

Perraudin, N., Balazs, P., Søndergaard, P.L. (2013). A fast griffin-lim algorithm. *WASPAA.*

Purushwalkam, S., Gari, S.V.A., Ithapu, V.K., Schissler, C., Robinson, P., Gupta, A., Grauman, K. (2021). Audio-visual floorplan reconstruction. *ICCV.*

Rayleigh, L. (1875). On our perception of the direction of a source of sound. *Proceedings of the Musical Association.*

Richard, A., Markovic, D., Gebru, I.D., Krenn, S., Butler, G., de la Torre, F., Sheikh, Y. (2021). Neural synthesis of binaural speech from mono audio. *ICLR.*

Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention.*

Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., Torralba, A. (2019). Self-supervised audio-visual co-segmentation. *ICASSP.*

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., ... others (2019). Habitat: A platform for embodied ai research. *ICCV.*

Schissler, C., Loftin, C., Manocha, D. (2017). Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics.*

Schroeder, M.R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America.*

Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., So Kweon, I. (2018). Learning to localize sound source in visual scenes. *CVPR.*

Tang, Z., Bryan, N.J., Li, D., Langlois, T.R., Manocha, D. (2020). Scene-aware audio rendering via deep acoustic analysis. *IEEE Transactions on Visualization and Computer Graphics.*

Tian, Y., Li, D., Xu, C. (2020). Unified multisensory perception: Weakly-supervised audio-visual video parsing. *ECCV.*

Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C. (2018). Audio-visual event localization in unconstrained videos. *ECCV.*

Tzinis, E., Wisdom, S., Jansen, A., Hershey, S., Remez, T., Ellis, D.P., Hershey, J.R. (2021). Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *ICLR.*

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *JMLR.*

Wu, Y., Zhu, L., Yan, Y., Yang, Y. (2019). Dual attention matching for audio-visual event localization. *ICCV.*

Xu, X., Dai, B., Lin, D. (2019). Recursive visual sound separation using minus-plus net. *ICCV.*

Xu, X., Zhou, H., Liu, Z., Dai, B., Wang, X., Lin, D. (2021). Visually informed binaural audio generation without binaural audios. *CVPR.*

Yang, K., Russell, B., Salamon, J. (2020). Telling left from right: Learning spatial correspondence of sight and sound. *CVPR.*

Yu, J., Zhang, S.-X., Wu, J., Ghorbani, S., Wu, B., Kang, S., ... Yu, D. (2020). Audio-visual recognition of overlapped speech for the lrs2 dataset. *ICASSP.*

Zaunschirm, M., Schörkhuber, C., Höldrich, R. (2018). Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *The Journal of the Acoustical Society of America*, *143*(6), 3616–3627.

Zhao, H., Gan, C., Ma, W.-C., Torralba, A. (2019). The sound of motions. *ICCV.*

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A. (2018). The sound of pixels. *ECCV.*

Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X. (2019). Talking face generation by adversarially disentangled audio-visual representation. *AAAI.*

Zhou, H., Xu, X., Lin, D., Wang, X., Liu, Z. (2020). Sep-stereo: Visually guided stereophonic audio generation by associating source separation. *ECCV.*

Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L. (2018). Visual to sound: Generating natural sound for videos in the wild. *CVPR.*