

Geometry-Aware Multi-Task Learning for Binaural Audio Generation from Video

Rishabh Garg¹
rishabh@cs.utexas.edu

¹The University of Texas at Austin

Ruohan Gao²
rhgao@cs.stanford.edu

²Stanford University

Kristen Grauman^{1,3}
grauman@cs.utexas.edu

³Facebook AI Research

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 a new large video dataset with realistic binaural audio simulated for real-world scanned environments. On two datasets, we demonstrate the efficacy of our method, which achieves state-of-the-art results.

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 [BS]. 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.

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

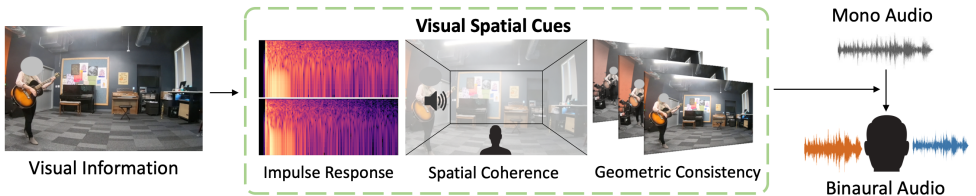


Figure 1: To generate accurate binaural audio from monaural audio, the visuals provide significant cues that can be learnt jointly with audio prediction. Our approach learns to extract spatial information (e.g., the guitar player is on the left), geometric consistency of the position of the sound sources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

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 [2, 51, 59]. 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 SimBinaural, a large-scale dataset of simulated videos with binaural sound in photo-realistic 3D indoor scene environments. This new dataset facilitates both learning and quantitative evaluation, allows us to explore the impact of particular parameters in a controlled manner, and even benefits learning in real videos. Finally, we show the efficacy of our method via extensive experiments in generating realistic binaural audio, achieving state-of-the-art results.

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.* [30] use 360 videos from YouTube to predict first order ambisonic sound useful for 360 viewing, while Lu *et al.* [29] use a self-supervised audio spatialization network using visual frames and optical flow. Whereas [29] uses correspondence to learn audio synthesizer ratio masks, which does not necessitate understanding of sound making objects, 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 [39].

More closely related to our problem, the 2.5D visual sound approach by Gao and Grauman generates binaural audio from video [27]. Building on those ideas, Zhou *et al.* [59] propose an associative pyramid network (APNet) architecture to fuse the modalities and jointly train on audio spatialization and source separation task. Concurrent to our work, Xu *et al.* [53] 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 video dataset with acoustically accurate binaural information (which will be shared publicly). We outperform the existing methods and show that the new dataset 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 [43], estimate reverberation time and equalization of the room using an actual 3D model of a room [46], and learn audio-visual correspondence from video [54]. Chen *et al.* [0] 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 [8, 9, 14, 20, 30]. Other work predicts depth maps using spatial audio [11] and learns representations via interaction using echoes recorded in indoor 3D simulated environments [24]. 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 [3, 5, 35]. Audio-visual data is leveraged for audio-visual speech recognition [12, 26, 55, 58], audio-visual event localization [47, 48, 50], sound source localization [9, 27, 40, 45, 47, 56], self-supervised representation learning [24, 28, 32, 33, 35], generating sounds from video [10, 18, 34, 60], and audio-visual source separation for speech [11, 5, 13, 16, 17, 33], music [19, 21, 52, 56, 57], and objects [21, 23, 49]. 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.

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

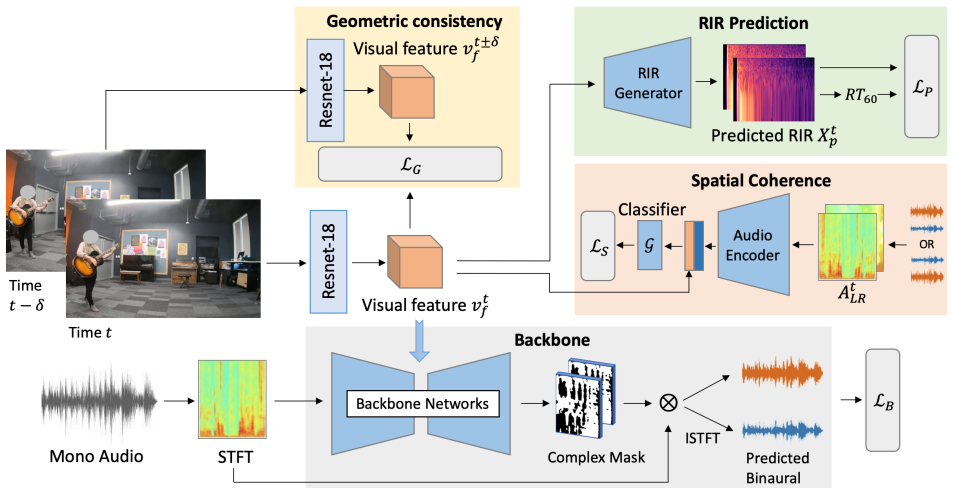


Figure 2: Proposed network. The network takes the visual frames and monoaural audio as input. The ResNet-18 visual features v_f^t are trained in a multi-task setting. The features v_f^t are used to directly predict the RIR via a decoder (top right). Audio features from binaural audio, which might have 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.

setting (Section 3.2). Next we describe the training and inference method (Section 3.3), and finally we describe the proposed SimBinaural dataset (Section 3.4).

3.1 Problem Formulation

Our objective is to map the monoaural 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 $\{v^1 \dots v^T\}$ and monoaural audio a_M^t , we aim to predict a two channel binaural audio output $\{a_L^t, a_R^t\}$. Whereas a single-channel audio a_M^t lacks spatial characteristics, two-channel binaural audio $\{a_L^t, a_R^t\}$ conveys two distinct waveforms to the left and right ears separately and hence provides spatial effects to the listener. By coupling the monoaural 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 $\{\mathcal{A}_L^t, \mathcal{A}_R^t\}$ from the input mono spectrogram \mathcal{A}_M^t , where $\mathcal{A}_X^t = \text{STFT}(a_X^t)$, conditioned on visual features v_f^t 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}_M^t to binaural ones. During training, the mono audio is obtained by averaging the two channels $a_M^t = (a_L^t + a_R^t)/2$ 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 [24]. 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_D^t , which, multiplied with the original audio spectrogram \mathcal{A}_M^t , gives the predicted difference spectrogram $\mathcal{A}_{D(pred)}^t = M_D^t \cdot \mathcal{A}_M^t$. The true difference spectrogram of the training input \mathcal{A}_D^t is the STFT of $a_L^t - a_R^t$. We minimize the distance between these two spectrograms: $\|\mathcal{A}_D^t - \mathcal{A}_{D(pred)}^t\|_2^2$. We also predict the two channels via two complex masks M_L^t and M_R^t , one for each channel, to obtain the predicted channel spectrograms $\mathcal{A}_{L(pred)}^t$ and $\mathcal{A}_{R(pred)}^t$ like above. This gives us the overall backbone objective:

$$\mathcal{L}_B = \|\mathcal{A}_D^t - \mathcal{A}_{D(pred)}^t\|_2^2 + \left\{ \|\mathcal{A}_L^t - \mathcal{A}_{L(pred)}^t\|_2^2 + \|\mathcal{A}_R^t - \mathcal{A}_{R(pred)}^t\|_2^2 \right\}. \quad (1)$$

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 $\mathcal{A}_{LR} = \{\mathcal{A}_L^t, \mathcal{A}_R^t\}$ 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, and we create misaligned samples by flipping the two channels in the ground truth audio to get $\mathcal{A}_{LR} = \{\mathcal{A}_R^t, \mathcal{A}_L^t\}$. 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 [24]. 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 $\mathcal{X}_{(pred)}^t$. We minimize the euclidean distance between the predicted RIR $\mathcal{X}_{(pred)}^t$, and the ground truth \mathcal{X}_{gt}^t . Additionally, we obtain the RIR waveform from the predicted spectrogram $\mathcal{X}_{(pred)}^t$ via the Griffin-Lim algorithm [49, 50] 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:

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

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:

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

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:

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

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 $a_{D(pred)}^t$ and two-channel audio $\{a_L^t, a_R^t\}$ via an inverse Short-Time Fourier Transformation (ISTFT) operation.

3.4 SimBinaural Dataset

We experiment with both real world video (FAIR-Play [47]) and video from scanned environments with high quality simulated audio. For the latter, to facilitate large-scale experimentation—and to augment learning from real videos—we create a new dataset called SimBinaural of

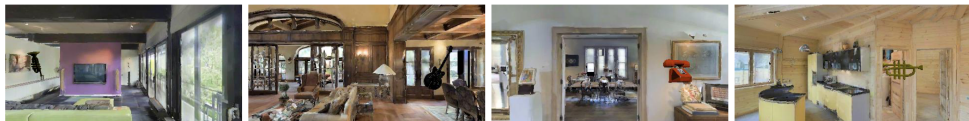


Figure 3: Example frames from the SimBinaural dataset.

Dataset	#Videos	Length (hrs)	#Rooms	RIR
FAIR-Play [22]	1,871	5.2	1	No
SimBinaural	21,737	116.1	1,020	Yes

Table 1: A comparison of the data in FAIR-Play and the large scale data we generated.

simulated videos in photo-realistic 3D indoor scene environments.¹ The generated videos, totalling 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 SoundSpaces² audio simulations [20] together with the Habitat simulator [21], we create realistic videos with binaural sounds for publicly available 3D environments in Matterport3D [6]. See Fig. 3 and Supp. video. Our resulting dataset is much larger and more diverse than the widely used FAIR-Play dataset [22] which is real video 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 guitar, violin, flute etc. and other sound-making objects like phones and clocks into the scene. 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 [20] room impulse response with an anechoic audio waveform (e.g., a guitar playing for an inserted guitar 3D object). Using this setup, we capture videos with simulated binaural sound. 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. Please see Supp. for details.

4 Experiments

We validate our approach on both FAIR-Play [22] (an existing real video benchmark) and our new SimBinaural dataset. We compare to the following baselines:

- **Flipped-Visual:** We flip the visual frame horizontally to provide incorrect visual information while testing. The other settings are the same as our method.
- **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 [22]:** 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 [54]:** 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.

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

²SoundSpaces [20] 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 of 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.

	FAIR-Play		SimBinaural			
	STFT	ENV	Scene-Split		Position-Split	
			STFT	ENV	STFT	ENV
Mono-Mono	1.215	0.157	1.356	0.163	1.348	0.168
Audio-Only	1.102	0.145	0.973	0.135	0.932	0.130
Flipped-Visual	1.134	0.152	1.082	0.142	1.075	0.141
Mono2Binaural [22]	0.927	0.142	0.874	0.129	0.805	0.124
APNet [59]	0.904	0.138	0.857	0.127	0.773	0.122
Backbone+IR Pred	n/a	n/a	0.801	0.124	0.713	0.117
Backbone+Spatial	0.873	0.134	0.837	0.126	0.756	0.120
Backbone+Geom	0.874	0.135	0.828	0.125	0.731	0.118
Our Full Model	0.869	0.134	0.795	0.123	0.691	0.116

Table 2: Binaural audio prediction errors on the FAIR-Play and SimBinaural datasets. For both metrics, lower is better.

- **PseudoBinaural** [53]: 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 [22, 61, 59]: 1) **STFT Distance**, the euclidean distance between the predicted and ground truth STFT spectrograms, which directly measures how accurate is our produced spectrogram, 2) **Envelope Distance** (ENV) which measures the euclidean distance between the envelopes of the predicted raw audio signal and the ground truth and can further capture the perceptual similarity.

Implementation details All networks are written in PyTorch [66]. The backbone network is based upon the networks used for 2.5D visual sound [22] and APNet [59]. The audio network consists of a U-Net [40] type architecture while the RIR generator is adapted from GANSynth [15]. To preprocess both datasets, we follow the standard steps from [22]. We resample all the audio to 16kHz and for training the backbone, we use 0.63s clips of the 10s audio and the corresponding frame. Frames are extracted at 10fps. The visual frames are randomly cropped to 448×224 . For testing, we use a sliding window of 0.1s to compute the binaural audio for all methods. Please see Supp. for more details.

SimBinaural results We evaluate on two data splits: 1) *Scene-Split*, where the train and test set have disjoint scenes from Matterport3D [6] 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 (right) 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.

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 in [53]. We compare our method with Augment-PseudoBinaural [53]³ 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.

User study 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 set of the two datasets. 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. 4, users preferred our method both for the accuracy of the direction of sound (left) and binaural audio quality (right).

Visualization Figure 5 shows the t-SNE projections [50] 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 [59] (center) are randomly distributed. Simultaneously, our features also accurately capture the angle of the object from the center (right). Fig. 6 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.

5 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

Method	STFT	ENV
APNet [59]	1.291	0.162
PseudoBinaural [53]	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.

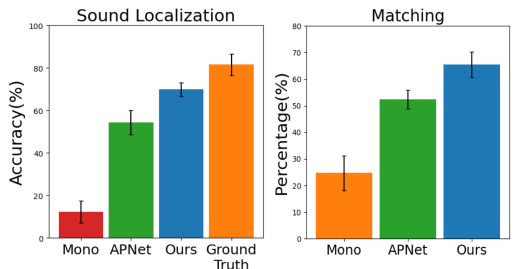


Figure 4: User study results. See text for details.

³The pre-trained model provided by PseudoBinaural [53] is trained on a different split instead of the standard split from [2] and hence it is not directly comparable in Table 2. We evaluate on the new split in Table 3.

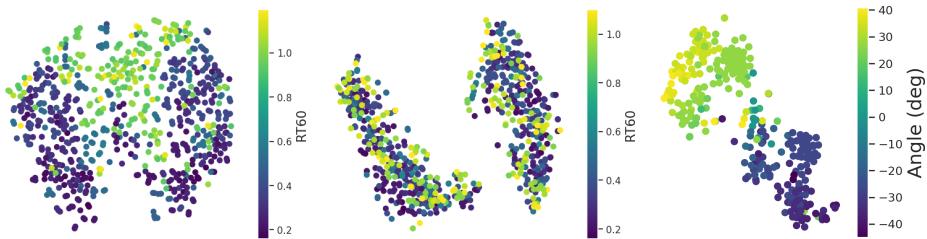


Figure 5: t-SNE of visual features colored by RT_{60} for our method (left) and APNet (center); and colored by angle of the object from the center (right).

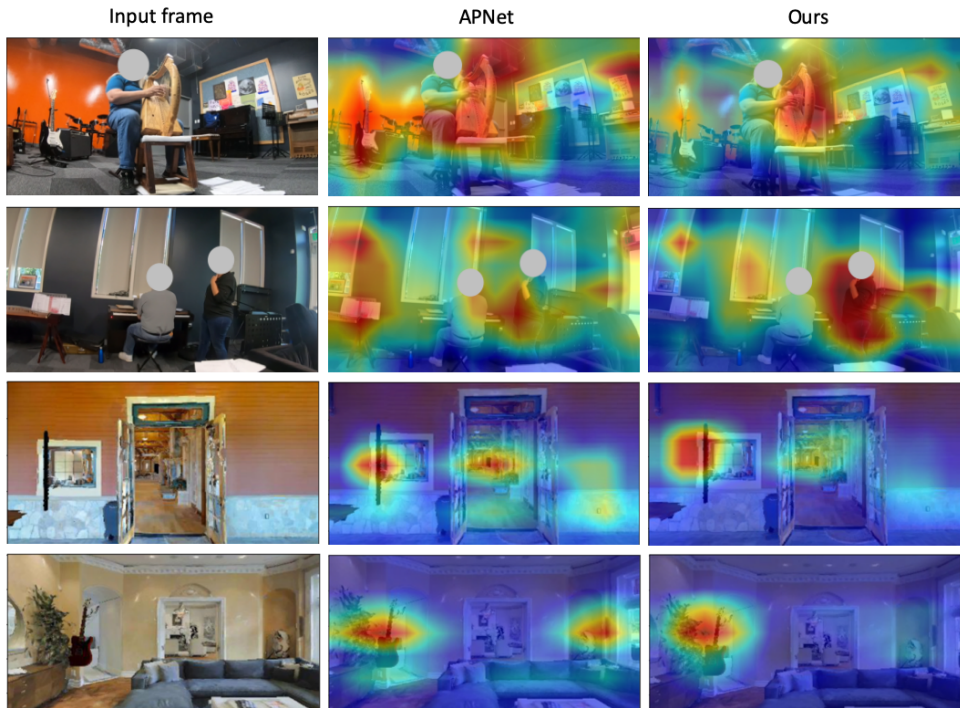


Figure 6: Qualitative visualization of the activation maps for the visual network for APNet [59] and ours. We can see that while the activation maps for APNet [59] are diffused and focusing on non-essential parts like objects in the background, our method focuses more on the object/region producing the sound and its location.

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. This dataset will be made publicly available to support further research in this direction. Our state-of-the-art results on two datasets demonstrate the efficacy of our proposed formulation. In future work we plan to explore how semantic models of object categories’ sounds could benefit the spatialization task.

Acknowledgements Thanks to Changan Chen for help with experiments, Tushar Nagarajan for feedback on paper drafts, and the UT Austin vision group for helpful discussions. UT Austin is supported by NSF CNS 2119115 and a gift from Google. Ruohan Gao was supported by a Google PhD Fellowship.

References

- [1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, 2018.
- [2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. In *ICASSP*, 2019.
- [3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017.
- [4] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018.
- [5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016.
- [6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. MatterPort3D dataset license available at: http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf.
- [7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020.
- [8] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2020.
- [9] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, 2021.
- [10] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE TIP*, 2020.
- [11] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *ICRA*, 2020.
- [12] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, 2017.
- [13] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. Facefilter: Audio-visual speech separation using still images. In *INTERSPEECH*, 2020.
- [14] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. In *NeurIPS*, 2020.

- [15] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. In *ICLR*, 2019.
- [16] 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.
- [17] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. In *INTERSPEECH*, 2018.
- [18] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *ECCV*, 2020.
- [19] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020.
- [20] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020.
- [21] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019.
- [22] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019.
- [23] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.
- [24] Ruohan Gao, Changan Chen, Ziad Al-Halab, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020.
- [25] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.
- [26] Di Hu, Xuelong Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *CVPR*, 2016.
- [27] 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.
- [28] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. In *NeurIPS*, 2018.
- [29] Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Self-supervised audio spatialization with correspondence classifier. In *ICIP*, 2019.
- [30] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2Hear: Active audio-visual source separation. In *ICCV*, 2021.
- [31] Pedro Morgado, Nono Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. In *NeurIPS*, 2018.

- [32] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *NeurIPS*, 2020.
- [33] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.
- [34] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016.
- [35] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016.
- [36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019.
- [37] Nathanaël Perraudin, Peter Balazs, and Peter L Søndergaard. A fast griffin-lim algorithm. In *WASPAA*, 2013.
- [38] Lord Rayleigh. On our perception of the direction of a source of sound. *Proceedings of the Musical Association*, 1875.
- [39] 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 *ICLR*, 2021.
- [40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.
- [41] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP*, 2019.
- [42] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019.
- [43] Carl Schissler, Christian Loftin, and Dinesh Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 2017.
- [44] Manfred R Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 1965.
- [45] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018.
- [46] Zhenyu Tang, Nicholas J Bryan, Dingzeyu Li, Timothy R Langlois, and Dinesh Manocha. Scene-aware audio rendering via deep acoustic analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

- [47] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.
- [48] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020.
- [49] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *ICLR*, 2021.
- [50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.
- [51] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019.
- [52] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *ICCV*, 2019.
- [53] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *CVPR*, 2021.
- [54] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *CVPR*, 2020.
- [55] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP*, 2020.
- [56] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.
- [57] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019.
- [58] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019.
- [59] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020.
- [60] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018.