2.5D Visual Sound

Ruohan Gao*
The University of Texas at Austin
rhgao@cs.utexas.edu

Kristen Grauman
Facebook AI Research
grauman@fb.com†

Abstract

Binaural audio provides a listener with 3D sound sensation, allowing a rich perceptual experience of the scene. However, binaural recordings are scarcely available and require nontrivial expertise and equipment to obtain. We propose to convert common monaural audio into binaural audio by leveraging video. The key idea is that visual frames reveal significant spatial cues that, while explicitly lacking in the accompanying single-channel audio, are strongly linked to it. Our multi-modal approach recovers this link from unlabeled video. We devise a deep convolutional neural network that learns to decode the monaural (single-channel) soundtrack into its binaural counterpart by injecting visual information about object and scene configurations. We call the resulting output 2.5D visual sound—*the visual stream helps "lift" the flat single channel audio into spatialized sound. In addition to sound generation, we show the self-supervised representation learned by our network benefits audio-visual source separation. Our video results:* `http://vision.cs.utexas.edu/projects/2.5D_visual_sound/`

1. Introduction

Multi-modal perception is essential to capture the richness of real-world sensory data and environments. People perceive the world by combining a number of simultaneous sensory streams, among which the visual and audio streams are paramount. In particular, both audio and visual data convey significant *spatial* information. We see where objects are and how the room is laid out. We also hear them: sound-emitting objects indicate their location, and sound reverberations reveal the room's main surfaces, materials, and dimensions. Similarly, as in the famous cocktail party scenario, while having a conversation at a noisy party, one can hear another voice calling out and turn to face it. The two senses naturally work in concert to interpret spatial signals.
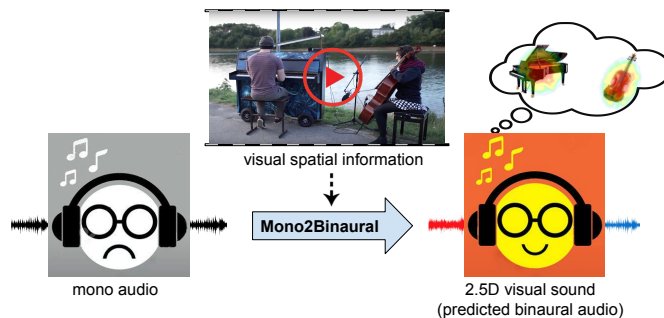

Figure 1: Binaural audio creates a 3D soundscape for listeners, but such recordings remain rare. The proposed approach infers *2.5D visual sound* by injecting the spatial information contained in the video frames accompanying a typical monaural audio stream.

The human auditory system uses *two* ears to extract individual sound sources from a complex mixture. The duplex theory proposed by Lord Rayleigh says that sound source locations are mainly determined by time differences between the sounds reaching each ear (Interaural Time Difference, ITD) and differences in sound level entering the ears (Interaural Level Difference, ILD) [37]. Accordingly, to mimic human hearing, *binaural audio* is usually recorded using two microphones attached to the two ears of a dummy head (see Fig. 2). The rig's two microphones, their spacing, and the physical shape of the ears are all significant for approximating how humans receive sound signals. As a result, when playing binaural audio through headphones, listeners feel the *3D sound sensation* of being in the place where the recording was made and can easily localize the sounds. The immersive spatial sound is valuable for audiophiles, AR/VR applications, and social video sharers alike.

However, binaural recordings are difficult to obtain in daily life due to the high price of the recording device and the required expertise. Consumer-level cameras typically only record monaural audio with a single microphone, or stereo audio recorded using two microphones with arbitrary arrangement and without physical representation of the pinna (outer ear). We contend that for both machines

---

*Work done during an internship at Facebook AI Research.

†On leave from The University of Texas at Austin (grauman@cs.utexas.edu).

and people, monaural or even stereo auditory input has very limited dimension. Monaural audio collapses all independent audio streams to the same spatial point, and the listener cannot sense the spatial locations of the sound sources.

Our key insight is that video accompanying monaural audio has the potential to unlock spatial sound, lifting a flat audio signal into what we call "2.5D visual sound". Although a single channel audio track alone does not encode any spatial information, its accompanying visual frames do contain object and scene configurations. For example, as shown in Fig. 1, we observe from the video frame that a man is playing the piano on the left and a man is playing the cello on the right. Although we cannot sense the locations of the sound sources by listening to the mono recording, we can nonetheless anticipate what we *would* hear if we were personally in the scene by inference from the visual frames.

We introduce an approach to realize this intuition. Given unlabeled training video, we devise a MONO2BINAURAL deep convolutional neural network to convert monaural audio to binaural audio by injecting the spatial cues embedded in the visual frames. Our encoder-decoder style network takes a mixed single-channel audio and its accompanying visual frames as input to perform joint audio-visual analysis, and attempts to predict a two-channel binaural audio that agrees with the spatial configurations in the video. When listening to the predicted binaural audio—the 2.5D visual sound—listeners can then feel the locations of the sound sources as they are displayed in the video.

Moreover, we show that apart from binaural audio generation, the MONO2BINAURAL conversion process can also benefit audio-visual source separation, a key challenge in audio-visual analysis. State-of-the-art systems [13, 52, 31, 1, 10] aim to separate a mixed monaural audio recording into its component sound sources, and thus far they rely solely on the spatial cues evident in the visual stream. We show that the proposed audio-visual binauralization can *self-supervise* representation learning to elicit spatial signals relevant to separation from the audio stream as well. Critically, gaining this new learning signal requires neither semantic annotations nor single-source data preparation, only the same unlabeled binaural training video.

Our main contributions are threefold: Firstly, we propose to convert monaural audio to binaural audio by leveraging video frames, and we design a MONO2BINAURAL deep network to achieve that goal; Secondly, we collect FAIR-Play, a 5.2 hour video dataset with binaural audio—the first dataset of its kind to facilitate research in both the audio and vision communities; Thirdly, we propose to perform audio-visual source separation on predicted binaural audio, and show that it provides a useful self-supervised representation for the separation task. We validate our approach on four challenging datasets spanning a variety of sound sources (e.g., instruments, street scenes, travel, sports).

2. Related Work

Generating Sounds from Video Recent work explores ways to generate audio conditioned on "silent" video. Material properties are revealed by the sounds objects make when hit with a drumstick, and can be used to synthesize new sounds from silent videos [32]. Recurrent networks [53] or conditional generative adversarial networks [7] can generate audio for input video frames, while powerful simulators can synthesize audio-visual data for 3D shapes [51]. Rather than generate audio from scratch, our task entails converting an input one-channel audio to two-channel binaural audio guided by the visual frames.

Only limited prior work considers video-based audio spatialization [26, 28]. The system of [26] synthesizes sound from a speaker in a room as a function of viewing angle, but assumes access to an acoustic impulse recorded in the specific room of interest, which restricts practical use, e.g., for novel "off-the-shelf" videos. Concurrent work to ours [28] generates *ambisonics* (audio for the full viewing sphere) given 360° video and its mono audio. In contrast, we focus on normal field of view (NFOV) video and binaural audio. We show that directly predicting binaural audio creates better 3D sound sensations for listeners without being restricted to 360° videos. Moreover, while the end goal of [28] is audio spatialization, we also demonstrate that our MONO2BINAURAL conversion process aids audio-visual source separation.

Audio(-Visual) Source Separation Audio-only source separation has been extensively studied in the signal processing literature. "Blind" separation tackles the case where only a single channel is available [42, 43, 46, 19]. Separation becomes easier when multiple channels are observed using multiple microphones [29, 49, 9] or binaural audio [48, 8, 50]. Inspired by this, we transform mono to binaural by observing video, and then leverage the resulting representation to improve audio-visual separation.

Audio-visual source separation also has a rich history, with methods exploring mutual information [11], subspace analysis [41, 35], matrix factorization [34, 39, 13], and correlated onsets [6, 25]. Recent methods leverage deep learning for audio-visual separation of speech [10, 31, 1, 12], musical instruments [52], and other objects [13]. New tasks are also emerging, such as learning to separate on- and off-screen sounds [31], learning object sound models from unlabeled video [13], or predicting sounds per pixel [52]. All these methods exploit mono audio cues to perform audio-visual source separation, whereas we propose to predict binaural cues to enhance separation. Furthermore, different from the task of localizing pixels responsible for a given sound [21, 18, 54, 3, 52, 40, 44], our goal is to perform binaural audio synthesis.

Self-Supervised Learning Self-supervised learning exploits labels freely available in the structure of the data, and audio-visual data offers a wealth of such tasks. Recent work explores self-supervision for visual [33, 2] and audio [4] feature learning, cross-modal representations [5], and audio-visual alignment [31, 23, 16]. Our MONO2BINAURAL formulation is also self-supervised, but unlike any of the above, we use visual frames to supervise audio spatialization, while also learning better sound representations for audio-visual source separation.

3. Approach

Our approach learns to map monaural audio to binaural audio via video. In the following, we first describe our binaural audio video dataset (Sec. 3.1). Then we present our MONO2BINAURAL formulation (Sec. 3.2), and our network and training procedure to solve it (Sec. 3.3). Finally we introduce our approach to leverage inferred binaural sound to perform audio-visual source separation (Sec. 3.4).

3.1. FAIR-Play Data Collection

Training our method requires binaural audio and accompanying video. Since no large public video datasets contain binaural audio, we collect a new dataset we call FAIR-Play with a custom rig. As shown in Fig. 2, we assembled a rig consisting of a 3Dio Free Space XLR binaural microphone, a GoPro HERO6 Black camera, and a Tascam DR-60D recorder as the audio pre-amplifier. We mounted the GoPro camera on top of the 3Dio binaural microphone to mimic a person's embodiment for seeing and hearing, respectively. The 3Dio binaural microphone records binaural audio, and the GoPro camera records videos at 30fps with stereo audio. We simultaneously record from both devices so the streams are roughly aligned.

Note that both the ear shaped housing (pinnae) for the microphones and their spatial separation are significant; professional binaural mics like 3Dio simulate the physical manner in which humans receive sound. In contrast, stereo sound is captured by two mics with an arbitrary separation that varies across capture devices (phones, cameras), and so lacks the spatial nuances of binaural. The limit of binaural capture, however, is that a single rig inherently assumes a single *head-related transfer function*, whereas individuals have slight variations due to inter-person anatomical differences. Personalizing head-related transfer functions is an area of active research [20, 45].

We captured videos with our custom rig in a large music room (about 1,000 square feet). Our intent was to capture a variety of sound making objects in a variety of spatial contexts, by assembling different combinations of instruments and people in the room. The room contains various instruments including cello, guitar, drum, ukelele, harp, piano, trumpet, upright bass, and banjo. We recruited 20 volun-

Figure 2: Binaural rig and data collection in a music room.

teers to play and recorded them in solo, duet, and multi-player performances. We post-process the raw data into 10s clips. In the end, our FAIR-Play[1] dataset consists of 1,871 short clips of musical performances, totaling 5.2 hours. In experiments we use both the music data as well as ambisonics datasets [28] for street scenes and YouTube videos of sports, travel, etc. (cf. Sec. 4).

3.2. Mono2Binaural Formulation

Binaural cues let us infer the location of sound sources. The interaural time difference (ITD) and the interaural level difference (ILD) play an essential role. ITD is caused by the difference in travel distances between the two ears. When a sound source is closer to one ear than the other, there is a time delay between the signals' arrival at the two ears. ILD is caused by a "shadowing" effect—a listener's head is large relative to certain wavelengths of sound, so it serves as a barrier, creating a shadow. The particular shape of the head, pinnae, and torso also act as a filter depending on the locations of the sound sources (distance, azimuth, and elevation). All these cues are missing in monaural audio, thus we cannot sense any spatial effect by listening to single-channel audio.

We denote the signal received at the left and right ears by $x^L(t)$ and $x^R(t)$, respectively. If we mix the two channels into a single channel $x^M(t) = x^L(t) + x^R(t)$, then all spatial information collapses. We can formulate a self-supervised task to take the mixed monaural signal $x^M(t)$ as input and split it into two separate channels $\tilde{x}^L(t)$ and $\tilde{x}^R(t)$, using the original $x^L(t)$, $x^R(t)$ as ground-truth during training. However, this is a highly under-constrained problem, as $x^M(t)$
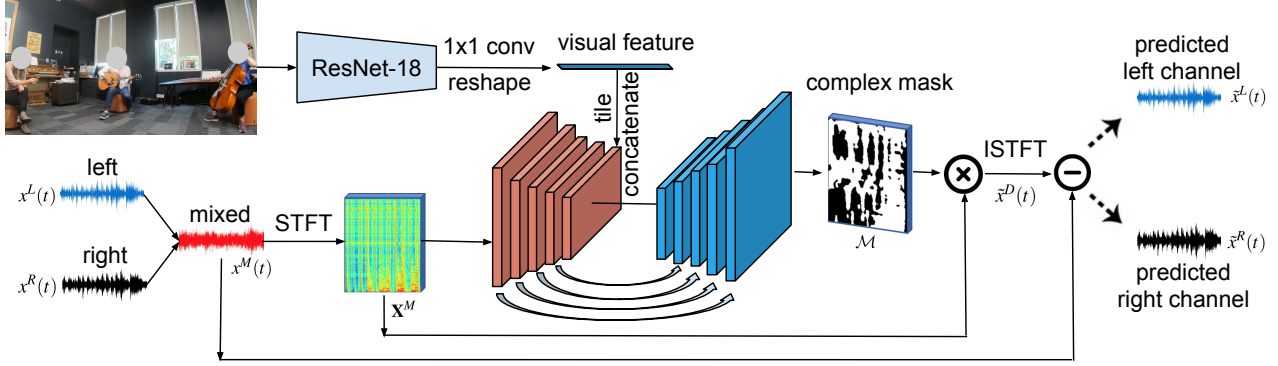
Figure 3: Our MONO2BINAURAL deep network takes a mixed monaural audio and its accompanying visual frame as input, and predicts a two-channel binaural audio output that satisfies the visual spatial configurations. An ImageNet pre-trained ResNet-18 network is used to extract visual features, and a U-NET is used to extract audio features and perform joint audio-visual analysis. We predict a complex mask for the audio difference signal, then combine it with the input mono audio to restore the left and right channels, respectively. At test time, the input is single-channel monaural audio.

lacks the necessary information to recover both channels. Our key idea is to guide the MONO2BINAURAL process with the accompanying video frames, from which *visual* spatial information can serve as supervision.

Instead of directly predicting the two channels, we predict the difference of the two channels:

$$x^D(t) = x^L(t) - x^R(t). \quad (1)$$

More specifically, we operate on the frequency domain and perform short-time Fourier transform (STFT) [15] on $x^M(t)$ to obtain the complex-valued spectrogram $\mathbf{X}^M$, and the objective is to predict the complex-valued spectrogram $\mathbf{X}^D$ for $x^D(t)$:

$$\mathbf{X}^M = \{\mathbf{X}^M_{t,f}\}^{T,F}_{t=1,f=1}, \quad \mathbf{X}^D = \{\mathbf{X}^D_{t,f}\}^{T,F}_{t=1,f=1}, \quad (2)$$

where t and f are the time frame and frequency bin indices, respectively, and T and F are the numbers of bins. Then we obtain the predicted difference signal $\tilde{x}^D(t)$ by the inverse short-time Fourier transform (ISTFT) [15] of $\mathbf{X}^D$. Finally, we recover both channels—the binaural audio output:

$$\tilde{x}^L(t) = \frac{x^M(t) + \tilde{x}^D(t)}{2}, \quad \tilde{x}^R(t) = \frac{x^M(t) - \tilde{x}^D(t)}{2}. \quad (3)$$

3.3. Mono2Binaural Network

Next we present our MONO2BINAURAL deep network to perform audio spatialization. The network takes the mono audio $x^M(t)$ and visual frames as input and predicts $x^D(t)$.

As shown in Fig. 3, we extract visual features from the center frame of the audio segment using ResNet-18 [17], which is pre-trained on ImageNet. The ResNet-18 network extracts per-frame features after the 4th ResNet block with size $(H/32) \times (W/32) \times C$, where H, W, C denote the frame and channel dimensions. We then pass the visual feature

through a 1×1 convolution layer to reduce the channel dimension, and flatten it into a single visual feature vector.

On the audio side, we adopt a U-NET [38] style architecture. The U-NET encoder-decoder network adopted here is ideal for our dense prediction task where the input and output have the same dimension. We mix the left and right channels of the binaural audio, and extract a sequence of STFT frames to generate an audio spectrogram $\mathbf{X}^M$. We use the complex spectrogram: each time-frequency bin contains the real and imaginary part of the corresponding complex spectrogram value. Then it is passed through a series of convolution layers to extract an audio feature of dimension $(T/32) \times (F/32) \times C$. We replicate the visual feature vector $(T/32) \times (F/32)$ times, tile them to match the audio feature dimension, and then concatenate the audio and visual feature maps along the channel dimension. Through the series of operations, each audio feature dimension is injected with the visual feature to perform joint audio-visual analysis.

Finally, we perform up-convolutions on the concatenated audio-visual feature map to generate a complex multiplicative spectrogram mask $\mathcal{M}$. In source separation tasks, spectrogram masks have proven better than alternatives such as direct prediction of spectrograms or raw waveforms [47]. Similarly, here we also adopt the idea of masking, but our goal is to mask the spectrogram of the mixed mono audio and predict the spectrogram of the difference signal, rather than perform separation. The real and imaginary components of the complex mask are separately estimated in the real domain. We add a sigmoid layer after the up-convolution layers to bound the complex mask values to [-1, 1], similar to [10]. The series of convolutions and up-convolutions maps the input mono spectrogram to a complex mask that encodes the predicted binaural audio.

Initially, we attempted to directly predict the left and right channels. However, we found that direct prediction

makes the network fall back on a "safe" but useless solution of copying and pasting the input audio, without reasoning with the visual features. Instead, predicting the difference signal forces the deep network to analyze the visual information and learn the subtle difference between the two channels, as required by the binaural audio target.

The spectrogram of the difference signal is then obtained by complex multiplying the input spectrogram with the predicted complex mask:

$$\tilde{\mathbf{X}}^D = \mathcal{M} \cdot \mathbf{X}^M. \tag{4}$$

We train our MONO2BINAURAL network using L2 loss to minimize the distance between the ground-truth complex spectrogram and the predicted one. Finally, using ISTFT, we obtain the predicted difference signal $\tilde{x}^D(t)$, through which we recover the two channels $\tilde{x}^L(t)$ and $\tilde{x}^R(t)$ as defined in Eq. 3. See supp. for network details.

At test time, the network is presented with monaural audio and a video frame and infers the binaural output, i.e., the 2.5D visual sound. To process a full video stream, each video is decomposed into many short audio segments. Video frames usually do not change much within such a short segment. We use a sliding window to perform spatialization segment by segment with a small hop size, and average predictions on overlapping parts. Thus, our method is able to handle moving sound sources and cameras.

Our approach expects a similar field of view (FoV) between training and testing, and assumes the microphone is near the camera. Our experiments demonstrate we can learn MONO2BINAURAL for both normal FoV and 360° video, and furthermore the same system can cope with mono inputs from variable hardware (e.g., YouTube videos).

3.4. Audio-Visual Source Separation

So far we have defined our MONO2BINAURAL approach to convert monaural audio to binaural audio by introducing visual spatial cues from video. Recall that we have two goals: to predict binaural audio for sound generation itself, and to explore its utility for audio-visual source separation.

Audio source separation is the problem of obtaining an estimate for each of the J sources s_j from the observed linear mixture $x(t) = \sum_{j=1}^{J} s_j(t)$. For binaural audio source separation, the problem is to obtain an estimate for each of the J sources s_j from the observed binaural mixture $x^L(t)$ and $x^R(t)$:

$$x^L(t) = \sum_{j=1}^{J} s_j^L(t), \qquad x^R(t) = \sum_{j=1}^{J} s_j^R(t), \tag{5}$$

where $s_j^L(t)$ and $s_j^R(t)$ are time-discrete signals received at the left ear and the right ear for each source, respectively.

Interfering sound sources are often located at different spatial positions in the physical space. Human listeners exploit the spatial information from the coordination of both ears to resolve sound ambiguity caused by multiple
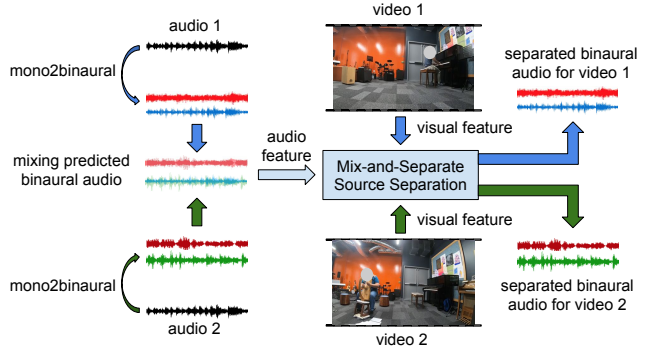

Figure 4: Mix-and-Separate [52, 31, 10]-inspired framework for audio-visual source separation. During training, we mix the binaural audio tracks for a pair of videos to generate a mixed audio input. The network learns to separate the sound for each video conditioned on their visual frames.

sources. This ability is greatly diminished when listening with only one ear, especially in reverberant environments [22]. Audio source separation by machine listeners is similarly handicapped, typically lacking access to binaural audio [52, 13, 31, 10]. However, we hypothesize that our MONO2BINAURAL *predicted* binaural audio can aid separation. Intuitively, by forcing the network to learn how to lift mono audio to binaural, its representation is encouraged to expose the very spatial cues that are valuable for source separation. Thus, even though the MONO2BINAURAL features see the same video as any other audio-visual separation method, they may better decode the latent spatial cues because of their binauralization "pre-training" task.

In particular, we expect two main effects. First, binaural audio embeds information about the spatial distribution of sound sources, which can act as a regularizer for separation. Second, binaural cues may be especially helpful in cases where sound sources have similar acoustic characteristics, since the spatial organization can reduce source ambiguities. Related regularization effects are observed in other vision tasks. For example, hallucinating motion enhances static-image action recognition [14], or predicting semantic segmentation informs depth estimation [27].

To implement a testbed for audio-visual source separation, we adopt the Mix-and-Separate idea [52, 31, 10]. We use the same base architecture as our MONO2BINAURAL network except that now the input to the network is a pair of training video clips. Fig. 4 illustrates the separation framework. We mix the sounds of the predicted binaural audio for the two videos to generate a complex audio input signal, and the learning objective is to separate the binaural audio for each video conditioned on their corresponding visual frames. Following [52], we only use spectrogram magnitude and predict a ratio mask for separation. Per-pixel L1 loss is used for training. See supp. for details.

4. Experiments

We validate our approach for generation and separation.

4.1. Datasets

We use four challenging datasets spanning a wide variety of sound sources, including musical instruments, street scenes, travel, and sports.

FAIR-Play Our new dataset consists of 1,871 10s clips of videos recorded in a music room (Fig. 2). The videos are paired with binaural audios of high quality recorded by a professional binaural microphone. We create 10 random splits by splitting the data into train/val/test splits of 1,497/187/187 clips, respectively.

REC-STREET A dataset collected by [28] using a Theta V 360° camera with TA-1 spatial audio microphone. It consists of 43 videos (3.5 hours) of outdoor street scenes.

YT-CLEAN This dataset contains in-the-wild 360° videos from YouTube crawled by [28] using queries related to spatial audio. It consists of 496 videos of a small number of super-imposed sources, such as people talking in a meeting room, outdoor sports, *etc*.

YT-MUSIC A dataset that consists of 397 YouTube videos of music performances collected by [28]. It is their most challenging dataset due to the large number of mixed sources (voices and instruments).

To our knowledge, FAIR-Play is the first dataset of its kind that contains videos of professional recorded binaural audio. For REC-STREET, YT-CLEAN and YT-MUSIC, we split the videos into 10s clips and divide them into train/val/test splits based on the provided split1. These datasets only contain ambisonics, so we use a binaural decoder to convert them to binaural audio. Specifically, we use the head related transfer function (HRTF) from NH2 subject in the ARI HRTF Dataset[2] to perform decoding. For our FAIR-Play dataset, half of the training data is used to train the MONO2BINAURAL network, and the other half is reserved for audio-visual source separation experiments.

4.2. Implementation Details

Both our MONO2BINAURAL and separation networks are in PyTorch. For all experiments, we resample the audio at 16kHz and STFT is computed using a Hann window of length 25ms, hop length of 10ms, and FFT size of 512. For MONO2BINAURAL training, we randomly sample audio segments of length 0.63s from each 10s audio clip. During testing, we use a sliding window with hop size 0.05s to binauralize 10s audio clips for both our method and baselines. For source separation experiments, we use similar network design and training/testing strategies. See supp. for details.

_____

[2]http://www.kfs.oeaw.ac.at/hrtf

4.3. Mono2Binaural Generation Accuracy

We evaluate the quality of our predicted binaural audio by using common metrics as well as two user studies. We compare to the following baselines:

- **Ambisonics [28]**: We use the pre-trained models provided by [28] to predict ambisonics. The models are trained on the same data as our method. Then we use the binaural decoder to convert the predicted ambisonics to binaural audio. This baseline is not available for the BINAURAL-MUSIC-ROOM dataset.

- **Audio-Only:** To determine if visual information is essential to perform MONO2BINAURAL conversion, we remove the visual stream and implement a baseline using only audio as input. All other settings are the same except that only audio features are passed to the up-convolution layers for binaural audio prediction.

- **Flipped-Visual:** During testing, we flip the accompanying visual frames of the mono audios to perform prediction using the wrong visual information.

- **Mono-Mono:** A straightforward baseline that copies the mixed monaural audio onto both channels to create a fake binaural audio.

We report two metrics: 1) **STFT Distance:** The euclidean distance between the ground-truth and predicted complex spectrograms of the left and right channels:

$$\mathcal{D}_{\{\text{STFT}\}} = ||\mathbf{X}^L - \tilde{\mathbf{X}}^L||_2 + ||\mathbf{X}^R - \tilde{\mathbf{X}}^R||_2.$$

2) **Envelope (ENV) Distance:** Direct comparison of raw waveforms may not capture perceptual similarity well. Following [28], we take the envelope of the signals, and measure the euclidean distance between the envelopes of the ground-truth left and right channels and the predicted signals. Let $E[x(t)]$ denote the envelope of signal $x(t)$. The envelope distance is defined as:

$$\mathcal{D}_{\{\text{ENV}\}} = ||E[x^L(t)] - E[\tilde{x}^L(t)]||_2 + ||E[x^R(t)] - E[\tilde{x}^R(t)]||_2.$$

Results. Table 1 shows the binaural generation results. Our method outperforms all baselines consistently on all four datasets. Our MONO2BINAURAL approach performs better than the Audio-Only baseline, indicating the visual stream is essential to guide conversion. Note that the Audio-Only baseline uses the same network design as our method, so it has reasonably good performance. Still, we find our method outperforms it most when object(s) are not simply located in the center. Flipped-Visual performs much worse, demonstrating that our network properly learns to localize sound sources to predict binaural audio correctly.

The Ambisonics [28] approach does not do as well. We hypothesize several reasons. The method predicts four channel ambisonics directly, which must be converted to

	FAIR-Play		REC-STREET		YT-CLEAN		YT-MUSIC	
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Ambisonics [28]	-	-	0.744	0.126	1.435	0.155	1.885	0.183
Audio-Only	0.966	0.141	0.590	0.114	1.065	0.131	1.553	0.167
Flipped-Visual	1.145	0.149	0.658	0.123	1.095	0.132	1.590	0.165
Mono-Mono	1.155	0.153	0.774	0.136	1.369	0.153	1.853	0.184
MONO2BINAURAL (Ours)	**0.836**	**0.132**	**0.565**	**0.109**	**1.027**	**0.130**	**1.451**	**0.156**

Table 1: Quantitative results of binaural audio prediction on four diverse datasets. We report the STFT distance and the envelope distance; lower is better. For FAIR-Play, we report the average results across 10 random splits. The results have a standard error of approximately 5×10^{-2} for STFT distance and 3×10^{-3} for ENV distance on average.
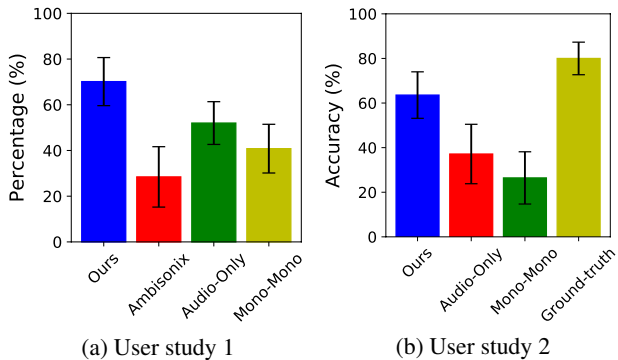

(a) User study 1 (b) User study 2

Figure 5: User studies to test how listeners perceive the predicted binaural audio.

binaural audio. While ambisonics have the advantage of being a more general audio representation that is ideal for 360° video, predicting ambisonics first and then decoding to binaural audio for deployment can introduce artifacts that make the binaural audio less realistic. Better head-related transfer functions could help to render more realistic binaural audio from ambisonics, but this remains active research [30, 24].[3] Furthermore, manually inspecting the results, we find that the decoded binaural audio by [28] conveys spatial sensation, but it is less accurate and stable than our method. Our approach directly formulates the audio spatialization problem in terms of the two-channel binaural audio that listeners ultimately hear, which yields better accuracy.

Our video results[4] show qualitative results including failure cases. Our system can fail when there are multiple objects of similar appearance, e.g. multiple human speakers. Our model incorrectly spatializes the audio, because the people are too visually similar. However, when there is only one human speaker amidst other sounds, it can successfully perform audio spatialization. Future work incorporating motion may benefit instance-level spatialization.

---

[3]We experimented with multiple ambisonics-binaural decoding solutions and report the best results for [28] in Table 1.

[4]http://vision.cs.utexas.edu/projects/2.5D_visual_sound/

User studies. Having quantified the advantage of our method in Table 1, we now report real user studies. To test how well the predicted binaural audio makes a listener feel the 3D sensation, we conduct two user studies.

For the first study, the participants listen to a 10s ground-truth binaural audio and see the visual frame. Then they listen to two predicted binaural audios generated by our method and a baseline (Ambisonics, Audio-Only, or Mono-Mono). After listening to each pair, participants are asked which of the two creates a better 3D sensation that matches the ground-truth binaural audio. We recruited 18 participants with normal hearing. Each listened to 45 pairs spanning all the datasets. Fig. 5a shows the results. We report the percentage of times each method is chosen as the preferred one. We can see that the binaural audio generated by our method creates a more realistic 3D sensation.

For the second user study, we ask participants to name the direction they hear a particular sound coming from. Using the FAIR-Play data, we randomly select 10 instrument video clips where some player is located in the left/center/right of the visual frames. We ask every participant to *only listen* to the ground-truth or predicted binaural audio from our method or a baseline, and then choose the direction the sound of a specified instrument is coming from. Note that for this study, we input real mono audio recorded by the GoPro mic for binaural audio prediction. Fig. 5b shows the results from the 18 participants. The true recorded binaural audio is of high quality, and the listeners can often easily perceive the correct direction. However, our predicted binaural audio also clearly conveys directionality. Compared to the baselines, ours presents listeners a much more accurate spatial audio experience.

4.4. Localizing the Sound Sources

Does the network attend to the locations of the sound sources when performing binauralization? As a byproduct of our MONO2BINAURAL training, we can use the network to perform sound source localization. We use a mask of size 32×32 to replace image regions with image mean values, and forward the masked frame through the network to predict binaural audio. Then we compute the loss, and repeat
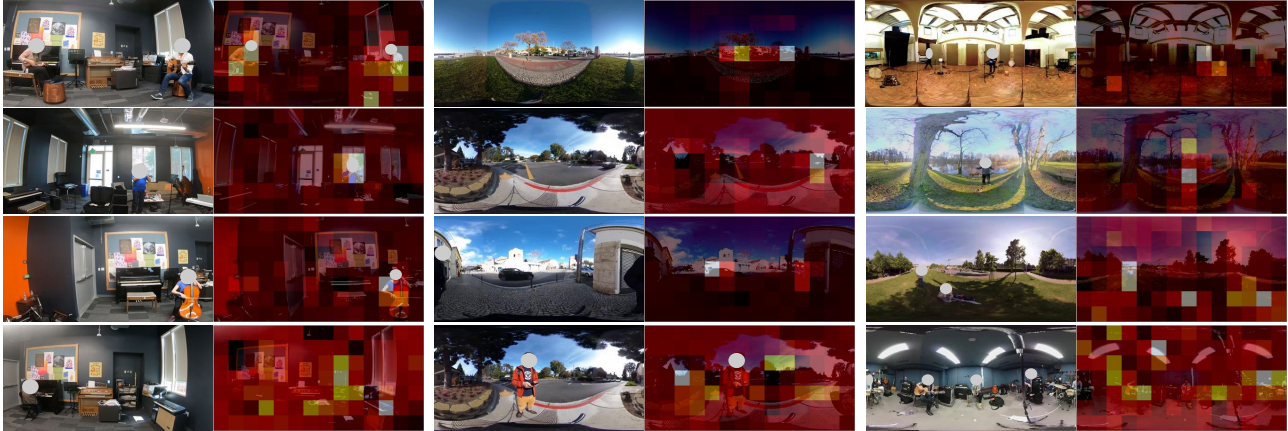
Figure 6: Visualizing the key regions the visual network focuses on when performing MONO2BINAURAL conversion. Each pair of images shows the frame accompanying the monaural audio (left) and the heatmap of the key regions overlaid (right).

	SDR	SIR	SAR
Mono	2.57	4.25	10.12
Mono-Mono	2.43	4.01	10.15
Predicted Binaural (Ours)	**3.01**	**5.03**	**10.24**
GT Binaural (upper bound)	3.25	5.32	10.60

Table 2: Audio-visual source separation results. SDR, SIR, SAR are reported in dB; higher is better.

by placing the mask at different locations of the frame. Finally, we highlight the regions which, when replaced, lead to the largest losses. They are considered the most important regions for MONO2BINAURAL conversion, and are expected to align with sound sources.

Fig. 6 shows examples. The highlighted key regions correlate quite well with sound sources. They are usually the instruments playing in the music room, the moving cars in street scenes, the place where an activity is going on, *etc*. The final row shows some failure cases. The model can be confused when there are multiple similar instruments in view, or silent or noisy scenes. Sound sources in YT-Clean and YT-Music are especially difficult to spatialize and localize due to diverse and/or large number of sound sources.

4.5. Audio-Visual Source Separation

Having demonstrated our predicted binaural audio creates a better 3D sensation, we now examine its impact on audio-visual source separation using the FAIR-Play dataset. The dataset contains object-level sounds of diverse sound making objects (instruments), which is well-suited for the Mix-and-Separate audio-visual source separation approach we adopt. We train on the held-out data of FAIR-Play, and test on 10 typical single-instrument video clips from the val/test set, with each representing one unique instrument in our dataset. We pairwise mix each video clip and perform separation, for a total of 45 test videos.

In addition to the ground truth binaural (upper bound) and the Mono-Mono baseline defined above, we compare to a **Mono** baseline that takes monaural audio as input and separates monaural audios for each source. Mono represents the current norm of performing audio-visual source separation using only single-channel audio [52, 13, 31]. We stress that all other aspects of the networks are the same, so that any differences in performance can be attributed to our binauralization self-supervision. To evaluate source separation quality, we use the widely used mir eval library [36], and the standard metrics: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). Table 2 shows the results. We obtain large gains by inferring binaural audio. The inferred binaural audio offers a more informative audio representation compared to the original monaural audio, leading to cleaner separation. See supp. video[4] for examples.

5. Conclusion

We presented an approach to convert single channel audio into binaural audio by leveraging object/scene configurations in the visual frames. The predicted 2.5D visual sound offers a more immersive audio experience. Our MONO2BINAURAL framework achieves state-of-the-art performance on audio spatialization. Moreover, using the predicted binaural audio as a better audio representation, we boost a modern model for audio-visual source separation. Generating binaural audio for off-the-shelf video can potentially close the gap between transporting audio and visual experiences, enabling new applications in VR/AR. As future work, we plan to explore ways to incorporate object localization and motion, and explicitly model scene sounds.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, 2018. 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 3

[3] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 3

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 3

[6] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *CVPR*, 2007. 2

[7] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *on Thematic Workshops of ACM Multimedia*, 2017. 2

[8] Antoine Deleforge and Radu Horaud. The cocktail party robot: Sound source separation and localisation with an active binaural head. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012. 2

[9] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010. 2

[10] 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. 2, 4, 5

[11] John W Fisher III, Trevor Darrell, William T Freeman, and Paul A Viola. Learning joint statistical models for audiovisual fusion and segregation. In *NIPS*, 2001. 2

[12] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. In *Interspeech*, 2018. 2

[13] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 2, 5, 8

[14] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow: Motion hallucination from static images for action recognition. In *CVPR*, 2018. 5

[15] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984. 4

[16] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[18] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NIPS*, 2000. 2

[19] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Deep learning for monaural speech separation. In *ICASSP*, 2014. 2

[20] Kazuhiro Iida, Yohji Ishii, and Shinsuke Nishioka. Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae. *The Journal of the Acoustical Society of America*, 2014. 3

[21] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *CVPR*, 2005. 2

[22] W Koenig. Subjective effects in binaural hearing. *The Journal of the Acoustical Society of America*, 1950. 5

[23] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. In *NIPS*, 2018. 3

[24] Matthias Kronlachner. Spatial transformations for the alteration of ambisonic recordings. *M. Thesis, University of Music and Performing Arts, Graz, Institute of Electronic Music and Acoustics*, 2014. 7

[25] Bochen Li, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *ICASSP*, 2017. 2

[26] Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng. Scene-aware audio for 360° videos. *SIGGRAPH*, 2018. 2

[27] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010. 5

[28] Pedro Morgado, Nono Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. *arXiv preprint arXiv:1809.02587*, 2018. 2, 3, 6, 7

[29] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi G Okuno, and Hiroaki Kitano. Real-time speaker localization and speech separation by audio-visual integration. In *IEEE International Conference on Robotics and Automation*, 2002. 2

[30] Markus Noisternig, Alois Sontacchi, Thomas Musil, and Robert Holdrich. A 3d ambisonic based binaural sound reproduction system. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003. 7

[31] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2, 3, 5, 8

[32] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016. 2

[33] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 3

[34] Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc QK Duong, Patrick Pérez, and Gaël Richard. Motion informed audio source separation. In *ICASSP*, 2017. 2

[35] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *ICASSP*, 2017. 2

[36] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and

C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, 2014. 8

[37] Lord Rayleigh. On our perception of the direction of a source of sound. *Proceedings of the Musical Association*, 1875. 1

[38] 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. 4

[39] Farnaz Sedighin, Massoud Babaie-Zadeh, Bertrand Rivet, and Christian Jutten. Two multimodal approaches for single microphone source separation. In *24th European Signal Processing Conference*, 2016. 2

[40] 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. 2

[41] Paris Smaragdis and Michael Casey. Audio/visual independent components. In *International Conference on Independent Component Analysis and Signal Separation*, 2003. 2

[42] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, 2007. 2

[43] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *12th International Conference on Digital Audio Effects*, 2009. 2

[44] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 2

[45] Edgar A Torres-Gallegos, Felipe Orduna-Bustamante, and Fernando Arámbula-Cosío. Personalization of head-related transfer functions (hrtf) based on automatic photo-anthropometry and inference from a database. *Applied Acoustics*, 2015. 3

[46] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 2007. 2

[47] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018. 4

[48] Ron J Weiss, Michael I Mandel, and Daniel PW Ellis. Source separation based on binaural cues and source model constraints. In *Ninth Annual Conference of the International Speech Communication Association*, 2008. 2

[49] Ozgur Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 2004. 2

[50] Xueliang Zhang and DeLiang Wang. Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM transactions on audio, speech, and language processing*, 2017. 2

[51] Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, and William T. Freeman. Generative modeling of audible shapes for object perception. In *ICCV*, 2017. 2

[52] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 2, 5, 8

[53] 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. 2

[54] A. Zunino, M. Crocco, S. Martelli, A. Trucco, A. Bue, and V. Murino. Seeing the sound: a new multimodal imaging device for computer vision. In *ICCV workshops*, 2015. 2