Localize to Binauralize: Audio Spatialization from Visual Sound Source Localization

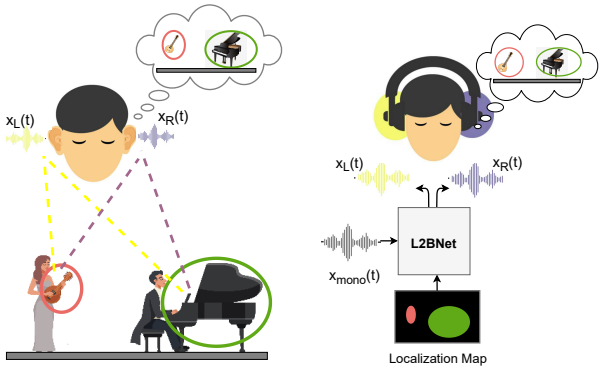
Kranthi Kumar Rachavarapu Aakanksha Vignesh Sundaresha Rajagopalan A. N.
Indian Institute of Technology Madras, India
{kranthi.rachavarapu, aakankshajha30, vigneshsundaresh}@gmail.com raju@ee.iitm.ac.in

Abstract

Videos with binaural audios provide immersive viewing experience by enabling 3D sound sensation. Recent works attempt to generate binaural audio in a multimodal learning framework using large quantities of videos with accompanying binaural audio. In contrast, we attempt a more challenging problem – synthesizing binaural audios for a video with monaural audio in a weakly semi-supervised setting. Our key idea is that any down-stream task that can be solved only using binaural audios can be used to provide proxy supervision for binaural audio generation, thereby reducing the reliance on explicit supervision. In this work, as a proxy-task for weak supervision, we use Sound Source Localization with only audio. We design a two-stage architecture called Localize-to-Binauralize Network (L2BNet). The first stage of L2BNet is a Stereo Generation (SG) network employed to generate two-stream audio from monaural audio using visual frame information as guidance. In the second stage, an Audio Localization (AL) network is designed to use the synthesized two-stream audio to localize sound sources in visual frames. The entire network is trained end-to-end so that the AL network provides necessary supervision for the SG network. We experimentally show that our weakly-supervised framework generates two-stream audio containing binaural cues. Through user study, we further validate that our proposed approach generates binaural-quality audio using as little as 10% of explicit binaural supervision data for the SG network.

1. Introduction

The perception of movement is primarily guided by rich visual cues in animals. This can be attributed to its evolutionary advantages over other modalities like sound. However, audio contains localization information that can be exploited, albeit not as richly as the visual modality. For example, consider a situation where you are talking to a friend on the sidewalk. The road is busy. Even if you are not facing the road, you are aware of vehicles zooming past. If you listen carefully, you can also figure out the direction of movement. Fig. 1(a) illustrates a similar phenomenon for live musical performances.



(a) (b)

Figure 1. (a) Binaural cues enable us to localize different instruments in live musical performances even with our eyes closed. (b) We use saliency maps to model localization information and use it to convert monaural audio to binaural audio.

According to [18], there are two major factors that allow us to spatialize using just the sounds we hear, namely, Interaural Time Difference (ITD) and Interaural Level Difference(ILD). ITD is the time difference between the same sound reaching each ear, and ILD is the amplitude difference between them. In the commonly used recording setups, a monoaural track is recorded using a single microphone. It aggregates sound signals from various sources and consequently loses such difference information. Binaural recordings use two microphones embedded inside a dummy head-model with realistic ear pinna to model closely the sound that would have fallen on a person's eardrums had they been present on site and preserves the localization information.

Despite being introduced many years ago, binaural audio recordings are relatively scarce, especially for videos. This can be attributed primarily to the fact that special equipment with multiple microphones is required for recording

it. Devices with multiple microphones are expensive and creating custom recording hardware, like the dummy-head mentioned earlier, is both expensive and involved. As such, there is a need to develop computational approaches that can make binaural audio more accessible as it can have a far-reaching impact. When accompanied by binaural audio tracks, videos provide a more immersive 3D experience for the viewers by enabling sound spatialization. Such experiences are of high interest to AR/VR enthusiasts and audiophiles. For people with visual impairments, being able to localize using the binaural audio track would enrich the experience of *listening to movies* and could possibly contribute towards making the cinematic experience more inclusive.

Recent works [5, 31, 14], attempt this problem of generating stereo audio in a self-supervised learning setup using videos with accompanying stereo audio. Another recent work [32], has attempted to utilize the abundantly available videos with monaural audio to enhance the quality of stereo sound generated. However, both these works require significant amounts of video recordings with binaural audio and creating such datasets is cumbersome.

As part of this work, we investigate how well we can localize an object based only on information extracted from binaural audio. We subsequently attempt to leverage this localization capability to solve the inverse problem of generating stereo tracks from monaural audio with minimal supervision. Through this work, we seek to establish that the efficient and commonly available object localization networks can be used to provide weak supervision for stereo-audio generation task while reducing the amount of binaural recordings needed for self-supervision. Fig. 1(b) gives an overview of our main idea.

Building on the idea that the complementary nature and natural synchronization of the video and audio stream are sufficient for stereo-sound generation given real binaural audio supervision, we try to minimize the amount of binaural recordings needed for the task. To this end, we solve the proxy-downstream task of sound source localization to provide the necessary weak supervision for binaural audio generation. This choice was made based on the observation that stereo-audio alone can be successfully used to localize a sound-making object in an accompanying video [4] while using only monaural audio fails. As such, if the audio being generated is able to localize well, we can conclude that it belongs to the subset of audios that contain localization information. However, we cannot be sure that this mapping between stereo audio and localization is unique. To force the network to generate only real binaural audio and not spurious solutions, we experiment by providing small amounts of explicit supervision to the network. Results show that as little as 10% of the total binaural recordings are enough to generate binaural-quality audios. The key novelties of our work can be listed as below-

- We propose an end to end model to convert a monaural audio accompanying a video to stereo audio using localization as the weak supervision.
- We also show that guiding the stereo audio generation task using localization helps reduce the amount of binaural recordings needed for learning to 10% of that needed by the present state of the art.

2. Related Works

Sound Source Localization. Sound source localization (SSL) is a widely researched task in the audio processing community and several methods exist which perform SSL using recorded *binaural* audio clips alone. Such methods primarily use ILD and ITD to estimate the elevation and azimuth angles of the sound sources [1, 3, 13, 29, 27, 20, 26, 10]. Of late, researchers have moved towards leveraging the visual cues from videos to address a problem called audio-visual sound source localization. The task is to localize visible objects generating sound by learning the audio visual correspondence between the *monaural* audio and the video frames[7, 15, 2, 24, 23]. Drawing inspiration from human experience, where one can localize based on the sounds one hears, [4] shows that tracking of sounding objects is possible even in the absence of good visual cues by leveraging SSL in videos with stereo audio recordings.

Monaural to Binaural audio conversion. There are several recent works that use visual guidance to generate stereo audio [5, 31, 8, 12, 32, 11, 14]. These works are supervised and use ground truth stereo or binaural tracks for supervision. In [14], 360° video and spatial audio are used as ground truth to localize sound sources. In [5], a new dataset is created comprising of videos with their corresponding binaural audio recordings, which is subsequently used to train a U-Net to generate the binaural spectrogram from mixed monaural audio input. An attempt to leverage the easily available monoaural audio clips for domain-specific pre-training has been tried to enhance the quality of stereo audio generated [32].

3. Method

Our primary goal is to generate the binaural audio stream $\{x_L(t), x_R(t)\}$ corresponding to any video $V = \{I_1 \dots I_T\}$ with monaural audio $x_{mono}(t)$. Existing methods approach this problem using a fully supervised setup and require large amounts of videos with binaural recordings. We attempt to solve this problem with *minimal* explicit supervision so as to reduce the amount of such recordings required significantly. The main idea is to introduce a down-stream task that can be constrained such that only

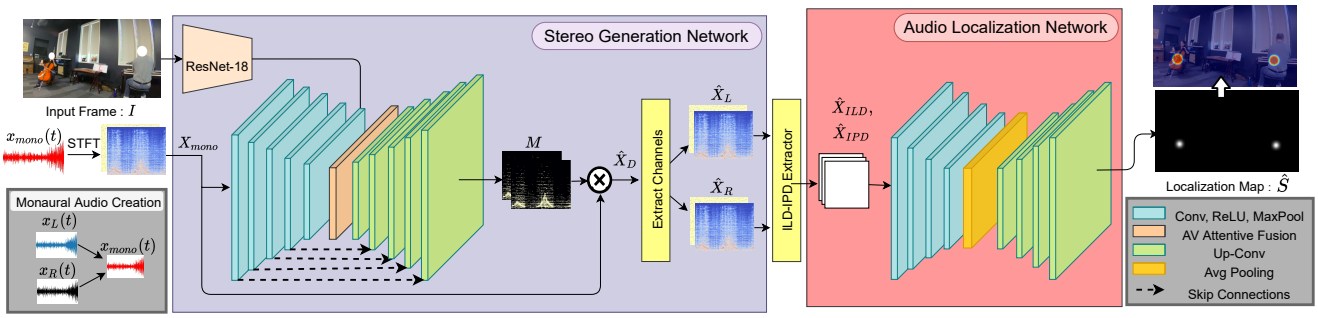


Figure 2. Architecture of our proposed L2BNet comprising of a Stereo Generation Network and an Audio Localization network with an ILD-IPD Extractor.

binaural audio can solve it and use it to provide proxy-supervision for binaural audio generation. We use *Sound Source Localization* as the down-stream task for our setup.

We propose a two-stage architecture called Localize-to-Binauralize network (*L2BNet*). An overview of our method is shown in Fig. 2. The proposed *L2BNet* consists of two networks: (a) Stereo-Generation (SG) network and (b) Audio-Localization (AL) network. The SG network takes the monaural audio $x_{mono}(t)$ as input and uses the features extracted from the corresponding visual frame I to generate a two-stream audio $\{\hat{x}_L(t), \hat{x}_R(t)\}$. The AL network takes as input a two-stream audio and uses it to localize sound sources on a visual frame. These two networks are described in detail next.

3.1. Stereo Generation (SG) Network

This network takes monaural audio $x_{mono}(t)$ and visual frame I as inputs and produces a two-stream audio $\{\hat{x}_L(t), \hat{x}_R(t)\}$, which at the end of the learning stage should perceptually sound like binaural audio. We train the SG network to generate the difference audio signal, $x_D(t)$, between the two audio streams since this was found to be more effective than learning the left and right spectrograms directly [5]. Subsequently, the two-stream audio can be recovered from the difference audio as follows:

$$\hat{x}_L(t) = x_{mono}(t) - \frac{x_D(t)}{2}, \;\; \hat{x}_R(t) = x_{mono}(t) + \frac{x_D(t)}{2}. \quad (1)$$

The SG network consists of two subnetworks - (a) Visual Subnetwork and (b) Audio Subnetwork. Our stereo generation network closely follows the architecture of [5] but with necessary modifications as described below.

Visual Subnetwork: This is a pretrained convolutional neural network employed to extract visual features from the input frame. We use pretrained ResNet-18 [6] trained on ImageNet [22] and extract features from the final convolutional layer. This is done to best preserve the spatial information of the objects in the input image which is essential to provide cues for binaural audio generation. This network takes as input an image and generates a visual representation $A_v \in \mathbb{R}^{C_v \times W \times H}$, which is a C_v dimensional feature map on $W \times H$ spatial grid.

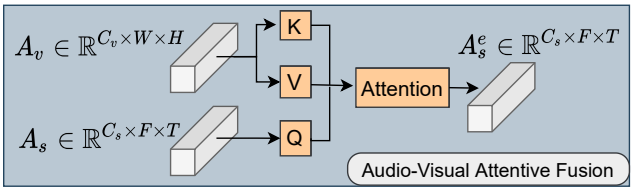


Figure 3. Attentive Feature Fusion

Audio Subnetwork: This has a U-net [21] like architecture which has been widely used for many audio-visual tasks [15, 28, 9, 5]. We create the monaural audio $x_{mono}(t)$ by mixing the input binaural audios $\{x_L(t), x_R(t)\}$ as $x_{mono}(t) = \frac{x_L(t)+x_R(t)}{2}$, and compute the corresponding spectrogram $X_{mono} \in \mathbb{C}^{F_s \times T_s}$ using Short Term Fourier Transform (STFT), where F_s and T_s are the frequency and time resolutions respectively, of STFT. The real and imaginary parts of this input spectrogram X_{mono} are stacked together and fed through a sequence of convolutional layers to obtain the latent representation $A_s \in \mathbb{R}^{C_s \times F \times T}$, which is a C_s dimensional feature map on $F \times T$ dimensional time-frequency grid. We perform an attentive feature fusion (described below) of the audio features A_s and visual features A_v to enhance the audio feature representation. These features are then passed through a sequence of up-convolutions to finally obtain a complex mask M. The spectrogram of the audio difference between channels is computed using this mask by multiplying with the input monaural audio as $X_D = MX_{mono}$. We then apply inverse STFT to obtain the audio difference signal, $x_D(t)$ and estimate the two-channel audio output using Equation 1.

Audio-Visual Feature Fusion: The goal of the feature fusion step is to infuse the complementary information available in the visual frame about the scene configuration into the audio features so that the predicted binaural audio agrees with the object placement in the frame. To achieve this, we

perform attentive feature fusion [25] between the audio features, $A_s \in \mathbb{R}^{C_s \times FT}$, and visual features, $A_v \in \mathbb{R}^{C_v \times WH}$, in the following manner:

$$Q = \mathbf{W_q} A_s, \quad K = \mathbf{W_k} A_v, \quad V = \mathbf{W_v} A_v, \tag{2}$$

$$A_s^e = A_s + V^T \, softmax(K^T \, Q), \tag{3}$$

where $\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v}$ are linear projection layers that operate on visual and audio features. The feature fusion described in Equation 3, learns to dynamically focus on relevant visual features A_v based on the audio feature representation A_s to get an enhanced audio feature representation A_s^e.

3.2. Audio-Localization (AL) Network

The sound source localization problem is well studied in literature [2, 23, 17]. These methods use both audio and visual cues to localize the sound creating objects in the image where image provides the necessary grounding. In contrast to these approaches, our audio localization network performs sound-source localization using *audio-alone*. Instead of regressing for the azimuth and elevation angles [30, 26] or bounding box [4], we predict a localization map on a visual frame akin to saliency prediction. This is because, in this work, we are only interested in localizing sound-sources on the visual frame than its size/distance.

Our AL network is an auto-encoder that takes binaural features as input and produces the sound source localization by predicting a soft score indicating the probability of each pixel being the sound source. The input binaural features are stacked together and are given as input to the encoder. The encoder, which consists of a sequence of convolutional and pooling layers, generates a low-resolution feature map $z \in \mathbb{R}^{D \times F \times T}$. The F, T dimensional information is collapsed by average-pooling to get $\hat{z} \in \mathbb{R}^{D \times 1 \times 1}$. This feature map acts as input to the decoder, which consists of a sequence of up-convolutional layers, to generate the final localization map $\hat{S} \in \mathbb{R}^{W \times H}$. This localization map depicts saliency regions for the sound-sources on the visual frame I corresponding to the input binaural audio. See supp. for network architecture details.

We experimented with different forms of binaural features as input to the audio localization network. For initial experimentation, we used the binaural audio spectrograms as the input. We then modified the spectrogram to obtain the binaural cues, the Interaural Phase Difference (IPD) and the Interaural Level Difference (ILD), using the expressions below and giving them input to the AL network.

$$X_{ILD} = 20 \log \left| \frac{X_L}{X_R} \right|, \quad X_{IPD} = \angle \left(\frac{X_L}{X_R} \right), \tag{4}$$

where X_L, X_R are the STFT of $x_L(t)$ and $x_R(t)$ respectively, and $|X|$, $\angle X$ are respectively the magnitude and phase of the spectrogram X. Note that $X_{ILD} \in \mathbb{R}^{F \times T}$ and $X_{IPD} \in \mathbb{R}^{F \times T}$ have the same size as the input binaural spectrogram. This was done keeping in mind that research attributes the ability of humans to localize sound sources to binaural cues such as the IPD and ILD. Therefore, extracting IPD and ILD features using Equation 4 captures the relevant binaural cues for better sound-source localization.

3.3. Weakly-Supervised (WS) Training Setup

Initially, the Stereo-Generation network and the Audio-Localization network were trained together using only localization loss as weak supervision(WS). Note that the AL network takes as input the explicit binaural cues (ILD and IPD) extracted from the synthesized two-stream audio of the SG network and performs sound source localization. This forces the SG network to predict only such two-stream audios that contain good binaural cues, else the localization task fails. While this WS framework enhances the binaural cues in the generated audio, it cannot ensure good overall quality. The synthesized two-stream audio could contain good binaural cues but still be *noisy* and have artefacts.

In addition, there is ill-posedness associated with ambiguity in Head-Related Transfer Functions (HRTFs). For instance, we show in our experiments that the same level of localization accuracy can be achieved using binaural audios in any of the following two settings: $\{x_L(t), x_R(t)\}$ or $\{\alpha x_L(t), \beta x_R(t)\}$, where α, β are constants. Intuitively, these two settings correspond to the same recording setup with two different HRTFs. Hence, the SG network can learn any such mapping which gives good localization accuracy but may not produce consistent binaural audio. Thus, weak supervision is necessary but not sufficient.

This weak supervision and extreme ill-posedness of the binaural audio generation task makes it imperative to introduce additional constraints to guide the SG network to generate good quality binaural audios corresponding to a particular HRTF setting. We achieve this by using a few binaural videos to provide explicit supervision and guide the SG network in an end-to-end learning framework, resulting in a *Weakly and Semi-Supervised* setup.

3.4. Loss function and Training

In our *Weakly-Supervised* (WS) learning setup, the Audio Localization network alone provides supervision for the Stereo Generation network. The loss for this WS setup is,

$$\mathcal{L}_{WS} = ||S - \hat{S}||_2^2 - \lambda ||M||_2^2 \tag{5}$$

where S is the ground truth (GT) saliency map for localization, $\hat{S}$ defined as $\hat{S} = AL\left(SG\left(X_{mono}, I;\ \theta_{SG}\right);\ \theta_{AL}\right)$ is the predicted localization map, and θ_{SG}, θ_{AL} are the learnable parameters. λ is the regularization parameter on the predicted mask, M, of SG network.

For the proposed *Weakly and Semi-Supervised* (WSS) learning setup, in addition to proxy-supervision from sound

source localization, we also use few binaural recordings to provide minimal explicit supervision to SG network. Let p be the percentage of samples in the dataset used for explicit supervision. We train the L2BNet in an alternating two-stage learning setup. In Stage-1, both SG network and AL network are trained independently for one epoch using only the $p\%$ explicit supervision data with the following loss,

$$\mathcal{L}_{WSS}^{AL} = ||S - AL\left(X_L, X_R;\ \theta_{AL}\right)||_2^2, \quad (6)$$

$$\mathcal{L}_{WSS}^{SG} = ||X_D - SG\left(X_{mono}, I;\ \theta_{SG}\right)||_2^2. \quad (7)$$

We then use the SG network with updated weights to predict binaural audios for all the remaining $(100-p)\%$ monaural videos in the training dataset and call them *pseudo-binaurals*. In Stage-2, the entire L2BNet is trained end-to-end for one epoch, using *pseudo-binaurals* along with *weak-supervision* from SSL task using the following loss,

$$\mathcal{L}_{WSS} = \mathcal{L}_{WS} + \alpha\mathcal{L}_{WSS}^{SG-pseudo}, \quad (8)$$

where $\mathcal{L}_{WSS}^{SG-pseudo} = ||X_D^{pseudo} - SG\left(X_{mono}, I;\ \theta_{SG}\right)||_2^2$, α is the regularization parameter, and $\mathcal{L}_{WS}$ is the same as Equation 5.

This proposed two-stage training disassociates binaural audio generation into two sub-tasks: (i) binaural cue infusion into SG network through sound source localization from AL network and (ii) explicitly optimizing for the audio quality of SG network. Furthermore, using explicit $p\%$ binaural supervision along with weighted loss ensures that our model learns two-stream channels with an audio structure similar to GT binaurals. Finally, we opted for *ILD/ITD features* to extract stronger binaural cues for localization as compared to binaural audios. All these measures guide the network and mitigate spurious learning.

4. Experiments

4.1. Datasets

We conduct the experiments on two datasets with musical instrument recordings, each with varying sound sources and locations: FAIR-Play [5] and YTMusic [16].

FAIR-Play [5] - This dataset consists of 1871 videos of musical instruments along with binaural audios recorded with a dummy-head setup. A pre-trained Yolo object detector [19] is used to detect humans/musical instruments in the videos, and a visual saliency map is generated around the centroid of the bounding box. We use the same train-test split proposed by the authors of [5] in our experiments.

YTMusic [16] - This dataset consists of recordings of musical performances as 360° videos with ambisonic sounds. It originally contained 397 videos as YouTube links, out of which 287 are still found available. Since this dataset contains ambisonic audio recordings, a binaural decoder was used to convert them to binaural audio. In particular, we use the HRTF from NH2 subject in the ARI HRTF dataset[1] to perform decoding. We manually annotate the bounding boxes for musical instruments and generate visual saliency maps around each of their centroids.

4.2. Implementation details and Evaluation

Networks are trained to generate binaural audio at 16kHz from input monaural audio processed at 16kHz and video at 10fps. Each training sample consists of 0.63s of monaural audio and the corresponding RGB frame and the saliency map at the centre of the clip as input, both of which are used to predict the two-stream audio, which is then used to predict the saliency map. The monaural audio segments are normalized. The STFT is computed with 512 frequency bins and 64 time bins. We augment the data by flipping the visual frame along the vertical axis and/or swapping the corresponding two-stream binaural audio. In all the experiments, we set α, λ to $0.2, 0.01$ respectively. All the models were trained with Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a learning rate of 10^{-5} on RTX 2080Ti GPU with a batch size of 32. During testing, monaural audio is processed with an overlapping window of 0.05 sec.

Evaluation and Metrics - We evaluate the predicted binaural audios using two standard metrics: (1) *STFT Distance* and (2) *Envelope Distance* similar to [5, 16]. STFT distance quantifies the frequency domain fidelity between the ground truth and predicted audios, while Envelope distance compares audios for consistency in phase. We evaluate the quality of the predicted object localization map using *F-measure* and *Mean Average Error (MAE)* as the predicted localization map is similar to visual saliency.

4.3. Sound Source Localization using Audio

The goal of this experiment is to verify the claim that sound sources can be localized using audio only if it contains some binaural cues. For this, we employ the Audio-Localization network described in Section 3.2, which takes (only) audio as input and learns to localize the sound sources on a spatial grid. Note that we represent the direction of arrival for different sound sources in a visual frame as a saliency map. The AL network has access only to the localization saliency map and not the visual frame. We experiment with the following forms of audio representations as input to the AL network:

Monaural: The average of GT binaural audio $x_{mono} = \frac{x_L + x_R}{2}$.

Binaural-Difference audio: The difference between two binaural audio channels $x_D = (x_L - x_R)$.

Binaural-mixed: We create a new two-channel audio as a weighted combination of original binaural audio $\{\alpha x_L + \beta x_R, (1-\alpha)x_L + (1-\beta)x_R\}, 0 \leq \alpha, \beta \leq 1$.

Binaural audio: The GT binaural audio $\{x_L, x_R\}$.

[1] HRTF available at http://www.kfs.oeaw.ac.at/hrtf

Table 1. Quantitative evaluation of Audio-based Visual Sound Source Localization task. ↑ Higher is better. ↓ Lower is better.

Audio Representation	F-measure ↑	MAE ↓
Monaural (x_{mono})	0.096	0.1766
Binaural-Difference (x_D)	0.184	0.0805
Binaural-mixed ($\{x_L^{mix}, x_R^{mix}\}$)	0.334	0.0309
Binaural ($\{x_L, x_R\}$)	0.380	0.0281
ILD & ITD ($\{x_{ILD}, x_{IPD}\}$)	**0.394**	**0.0183**

ITD and ILD: Binaural cues (ILD and IPD) of the input binaural audio $\{x_L, x_R\}$ extracted using Equation 4.

For each audio representation mentioned above, the AL network was trained on FAIR-Play dataset with MSE loss on the localization saliency maps till convergence. The quantative results of this Audio-based Sound Source Localization task are reported in Table 1. The *ILD and ITD* representation performs the best among all. This can be attributed to the fact that it captures the explicit binaural cues necessary for sound source localization. The *binaural* audio also performs equally well indicating that the localization is independent of the channel order. The *binaural audio difference* only works for the single source setup but fails when there are multiple sounding objects. The *monaural audio* representation fails to localize and performs poorly as the object spatial information is collapsed in monaural audio.

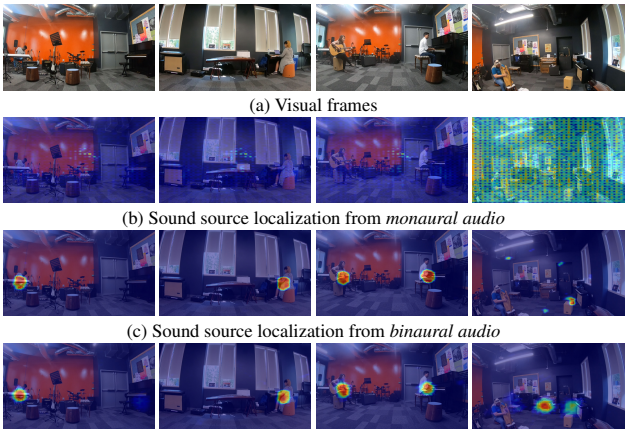
Figure 4 shows visual comparisons of the predicted localization map. The *monaural audio* representation is unable to localize the sources and predicts equal probability for all the locations. The *binaural* and *ILD & ITD* representations accurately localize the sound sources. Last column of Figure 4 shows an example where all representations fail to localize. However, *ILD & ITD* representation consistency performs better. Hence, achieving good performance on the sound source localization task requires some level of binaural information in the input audio; otherwise, the performance will be very poor. Refer to supp. for more qualitative comparisons.

4.4. Localization guided Binaural Audio Synthesis

Having established that binaural cues are necessary for sound source localization, in this section, we look into the problem of utilizing sound source localization for the task of binaural audio synthesis. First, we discuss our experiments with only *weak supervision* in the form of sound source location on the visual frame. This is then followed by *weak and semi-supervision* setup where we use some minimal explicit supervision in addition to weak supervision.

4.4.1 Binaural audio generation in Weakly Supervised Learning Setup

In this setup, our primary goal is to investigate whether the proxy down-stream task of sound source localization provides sufficient supervision to the SG network to infuse binaural cues in the synthesized two-stream audio. For this experiment, we train the L2BNet in an end-to-end learning setup with supervision coming from the localization loss alone. Note that only the SG network has access to the visual frame. The AL network has access only to the predicted two-channel audio generated by the SG network through the binaural feature extractor as shown in Fig. 2. We train this setup on the FAIR-Play dataset using WS loss described in Equation 5.



(a) Visual frames

(b) Sound source localization from *monaural audio*

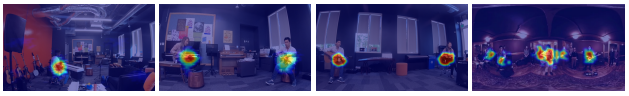
(c) Sound source localization from *binaural audio*

(d) Sound source localization from *ILD and ITD features*

Figure 4. Visual comparisons of Audio-based Visual Sound Source Localization task using various input audio forms.

Fig. 5 (a) shows the sound source localization results for this weakly-supervised setup. It can be observed that the proposed method localizes sound sources well, even in multiple instruments setup case. This indicates that the predicted two-channel audio $\{\hat{x}_L, \hat{x}_R\}$ indeed has binaural quality, which facilitates the localization. Refer to supp. for more qualitative comparisons

The quantitative evaluation of predicted binaural audio $\{\hat{x}_L, \hat{x}_R\}$ is not straight forward because as observed in Section 4.3, both *Binaural* and *Binaural-mixed* perform equally well on the localization task. Hence, the mapping between localization and binaural audio is not one-to-one. This means that the network trained in a weakly supervised setting can learn to predict binaural audios corresponding to any arbitrarily consistent HRTF while performing well on the localization task. As such, quantitatively comparing such audio against a ground truth binaural audio recorded with a fixed HRTF does not make sense. For lack of a better metric, despite the arguments presented, we report the correlations between the predicted two-channel audio and the binaural recordings from the FAIR-Play dataset. The *STFT / Envelope distance* of the predicted two-channel audio ($\{\hat{x}_L, \hat{x}_R\}$) with respect to ground truth is $3.219/0.231$ whereas its flipped version ($\{\hat{x}_R, \hat{x}_L\}$) has the corresponding metrics as $3.984/0.262$. This indicates that the predicted two-channel audio has binaural features and hence



(a) (b)

Figure 5. Visual comparisons of sound source localization task in (a) *Weakly Supervised* learning setup and (b) *Weakly Semi-Supervised* learning setup.

correlates with the ground truth binaural audios. Though the predicted binaural audio has perceptible binaural cues, it also has artefacts such as noise/muffled sound. This is because the proposed *WS* learning framework is designed to enhance the binaural cues in the predicted audio but does not ensure good audio quality.

4.4.2 Binaural Audio Generation in Weakly and Semi-Supervised Setup

The *weakly-supervised* setup with only sound source localization providing supervision to the SG network generates audio with noise and artefacts. It does not address the ambiguity in HRTF either. To resolve these issues, we introduce a few ground truth (GT) binaural video samples as data while training our L2BNet. In addition to resolving the issues of the WS setup, this also makes comparisons with baselines more tractable. Our objective in this *Weakly Semi-Supervised* (WSS) learning setup is to use a few GT binaural audios as explicit supervision for the SG network so that the predicted two-stream audio corresponds to the recording setup of the audio used for supervision.

For this, L2BNet is trained end-to-end using alternating training process and loss functions described in Section 3.4. We train it on FAIR-Play and YT-Music datasets separately for varying amounts of GT binaural samples as explicit supervision along with weak-supervision from SSL task.

The results of quantitative evaluation for generated binaural audios using this WSS learning setup are reported in Table 2. The performance of our L2BNet-WSS framework is lower than fully supervised approaches of [5, 32]. This can be explained by the fact that we are indirectly optimizing for the binaural audios through localization task and use very limited amount of supervision. For fair comparison, we create a baseline where only the SG network is trained with identical data, which we refer to as *Ours-SG*. This is done to establish that the same number of samples used by our L2BNet, when used in a completely supervised training setup, is not enough to learn the binaural mapping. Note that our SG network has some modifications over the backbone architecture of [5] and hence serves as a better baseline. When only 10% of the GT binaural samples are used for supervision, the baseline method (*Ours-SG 10%*) fails to perform any meaningful binauralization as reflected by the STFT and Envelope distance on comparing with monaural audio. On the other hand, when we provide weak supervision, along with the 10% GT binaural samples, to the same network with our Audio-based Sound Source Localization task (*L2BNet-WSS 10%*), the proposed method learns to generate two-stream audio with binaural cues as indicated by better quantitative metrics. This can be attributed to the fact that the proposed WSS framework aids in learning the discriminative binaural features that ensure better performance on the localization task, while agreeing with the recording setup of binaural audios used for explicit supervision. On increasing the amount of supervision to 30%, our approach (*L2BNet-WSS 30%*) performs much better than the baseline (*Ours-SG 30%*). Detailed experiments for varying amounts of explicit supervision data are given in Section 4.6. In Figure 5 (b), we show the sound source localization results of this WSS approach where it is able to localize the sound sources well. See supp. for more qualitative comparisons. The video results are available on our project page[2].

Table 2. Quantitative comparisons of the proposed *Weakly and Semi-Supervised* approach with various baseline methods on FAIR-Play and YT-Music using *STFT* and *Envelope* Distance. $F/S/W$ indicate Full/Semi/Weak supervision.

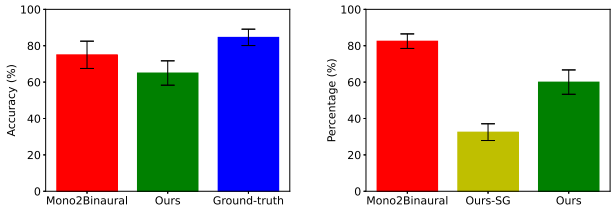
				Fair-Play		YT-Music	
	F	S	W	STFT	ENV	STFT	ENV
Mono				1.195	0.156	3.075	0.241
Mono2Binaural [5]	✓			0.951	0.141	1.346	0.179
Sep-Stereo [32]	✓			0.879	0.135	1.051	0.145
Ours-SG (10 %)		✓		1.188	0.156	2.156	0.203
Ours-SG (30 %)		✓		1.109	0.151	1.855	0.192
L2BNet-WSS (10 %)		✓	✓	1.121	0.151	1.908	0.195
L2BNet-WSS (30 %)		✓	✓	1.028	0.148	1.816	0.189

4.5. User Study

We conduct user studies to validate our claim that the proposed *Weakly Semi-Supervised* approach generates good quality binaural audios with as little as 10% of binaural audios as explicit supervision. We designed two experiments and used 13 test samples with single and multiple sources. A total of 20 users participated, and we used the results corresponding to our L2B-WSS (10%) for both experiments.

User Study 1 - Binaural Cue Perception: Here, we examine whether our generated binaural audio contains sufficient binaural cues to perceive the correct direction of sound sources. Each user listens to audio samples selected randomly from *GT binaural*, *Mono2Binaural* or our method, and is asked to pick the direction of sound source as *left*, *right* or *centre*. The results are shown in Fig. 6 (a) where we can see that the users were able to accurately perceive the direction of sound source with 85% accuracy for *GT binaural* samples, 75% for *Mono2Binaural*, and 65% for our approach. This indicates that our method generates audio

[2] https://github.com/KranthiKumarR/Localize-to-Binauralize



(a) (b)

Figure 6. (a) Results from User Study 1 showing the accuracy of predicted directions. (b) Results from User Study 2 depicting the perceived closeness to GT binaural.

with perceptible binaural cues. One-sided ANOVA analysis on these results yield an F-statistic of 8.42 with a p-value of 0.003, indicating that the results are statistically significant.

User Study 2 - Binaural Audio Quality: We conduct this study to examine the quality of our synthesized binaural audios. The participants are shown two videos corresponding to a single sample, one with GT binaural audio and the other with monaural audio. Then they are shown the same sample with predicted binaural audio from our method, *Mono2Binaural* method or *Ours-SG10%* method and asked which of the previous two videos is the third video closer to based on audio spatialization quality. The results are shown in Figure 6 (b). Samples corresponding to *Mono2Binaural* method and *Ours-SG 10%* were scored close to GT binaural 82.5% and 32.5% times on average, while our method received an average score of 60% indicating the enhanced binaural quality of our generated audio.

These user studies further validate that our proposed *Weakly Semi-Supervised* learning framework successfully infuses perceptible binaural cues into monaural audios with as little as 10% ground truth binaural audios as supervision.

4.6. Ablations

Are ILD and IPD features necessary? Given that both *binaural* audio and *ILD-IPD* features are equally good for localizing sound sources (from Section 4.3), we investigate which of the two representations is best suited for our *weak supervision* setup. For this, we retrain our L2BNet WSS 10% setup by removing the ILD-IPD feature extractor and directly feeding the predicted two-stream audio. The quantitative values obtained are reported in Table 3 and *ILD-IPD feature* based method performs better as it encodes the discriminative binaural features more explicitly.

Importance of Attention: We retrain our model L2BNet-WSS (10%) *with* and *without* attention. The results are reported in Table 3, where we can observe a marginal improvement in performance when attention is employed.

Performance with varying percentages of supervision: We performe experiments on L2BNet with 10%, 15%, 30%, 50% and 100% GT binaurals in the *WSS* framework and report the results in Figure 7. As baseline, we train SG

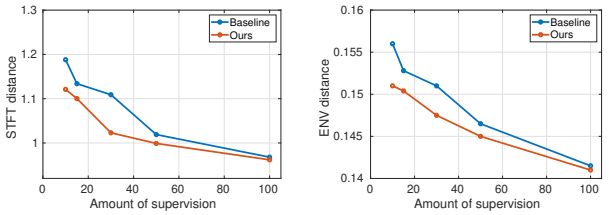


Figure 7. Ablation Study to examine the variation in the performance of the proposed *Weakly and Semi-Supervised* approach with varying percentages of supervision.

Table 3. Analysis of the performance of the proposed *WSS* approach with/without *ILD & IPD* features and *Attention*.

AL Network Input		Attention	
Binaural	ILD & IPD	Without	With
1.143/0.152	1.121/0.151	1.121/0.151	1.162/0.155

network with same amount of explicit supervision data. With higher percentage of explicit supervision ($> 30\%$), the SG network learns to generate good quality binaural audios, which directly translates to a good performance on the sound source localization task. Since the explicit binaural audios used for supervision already contain sufficient binaural cues which are successfully being learnt by the SG network with such high amounts of explicit supervision, the AL network may not add any more binaural cues to the already good-enough predicted binaural audio. Thus our method does not improve in performance over the baseline method. When the explicit binaural supervision is minimal ($\leq 30\%$), the proposed L2BNet-WSS scheme outperforms, as the SSL task significantly enhances the binaural cues when ground truth binaurals are limited.

5. Conclusion

We examined the problem of synthesizing binaural audios from videos with monaural audio with very binaural audio supervision. To address this problem, we proposed a framework called Localize-to-Binauralize which reduces the amount of supervision required by leveraging the weak-supervision from sound source localization to enhance binaural cues in the audio. Through our experiments, we show that our framework generates two-stream audio having binaural quality using as little as 10% of explicit supervision data. The user study performed further indicates that the two-stream audio synthesized using our proposed L2BNet in a *Weakly Semi-Supervised* setup has sufficient binaural cues for better hearing experience as well and better sound source perception.

Acknowledgement: Support from Institute of Eminence (IoE) project No. SB20210832EEMHRD005001 is gratefully acknowledged.

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018. 2

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 2, 4

[3] Antoine Deleforge, Radu Horaud, Yoav Y Schechner, and Laurent Girin. Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):718–731, 2015. 2

[4] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7053–7062, 2019. 2, 4

[5] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2, 3, 5, 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. *Advances in neural information processing systems*, 12:813–819, 1999. 2

[8] Haikun Huang, Michael Solah, Dingzeyu Li, and Lap-Fai Yu. Audible panorama: Automatic spatial audio generation for panorama imagery. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–11, 2019. 2

[9] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11):1767–1779, 2019. 3

[10] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 88–95. IEEE, 2005. 2

[11] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2

[12] Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Self-supervised audio spatialization with correspondence classifier. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3347–3351. IEEE, 2019. 2

[13] Michael I Mandel, Ron J Weiss, and Daniel PW Ellis. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394, 2009. 2

[14] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*, pages 362–372, 2018. 2

[15] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2, 3

[16] Timothy Langlois Pedro Morgado, Nuno Vasconcelos and Oliver Wang. Self-supervised generation of spatial audio for 360deg video. In *Neural Information Processing Systems (NIPS)*, 2018. 5

[17] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 292–308, 2020. 4

[18] Lord Rayleigh. On our perception of the direction of a source of sound. *Proceedings of the Musical Association*, 2:75–84, 1875. 1

[19] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[20] Michael Risoud, J-N Hanson, Fanny Gauvrit, C Renard, P-E Lemesre, N-X Bonne, and Christophe Vincent. Sound source localization. *European annals of otorhinolaryngology, head and neck diseases*, 135(4):259–264, 2018. 2

[21] 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*, pages 234–241. Springer, 2015. 3

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[23] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 2, 4

[24] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. 4

[26] Juan Manuel Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors*, 18(10):3418, 2018. 2, 4

[27] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang. Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):178–188, 2018. 2

[28] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 3

[29] John Woodruff and DeLiang Wang. Binaural localization of multiple sources in reverberant and noisy environments.

IEEE Transactions on Audio, Speech, and Language Processing, 20(5):1503–1512, 2012. 2

[30] Nelson Yalta, Kazuhiro Nakadai, and Tetsuya Ogata. Sound source localization using deep learning models. *Journal of Robotics and Mechatronics*, 29(1):37–48, 2017. 4

[31] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 2

[32] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. pages 52–69, 2020. 2, 7