This WACV 2022 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Beyond Mono to Binaural: Generating Binaural Audio from Mono Audio with Depth and Cross Modal Attention

Kranti Kumar Parida[1] Siddharth Srivastava[2] Gaurav Sharma[1,3]
[1] IIT Kanpur [2] CDAC Noida [3] TensorTour Inc.
{kranti, grv}@cse.iitk.ac.in, siddharthsrivastava@cdac.in

Abstract

Binaural audio gives the listener an immersive experience and can enhance augmented and virtual reality. However, recording binaural audio requires specialized setup with a dummy human head having microphones in left and right ears. Such a recording setup is difficult to build and setup, therefore mono audio has become the preferred choice in common devices. To obtain the same impact as binaural audio, recent efforts have been directed towards lifting mono audio to binaural audio conditioned on the visual input from the scene. Such approaches have not used an important cue for the task: the distance of different sound producing objects from the microphones. In this work, we argue that depth map of the scene can act as a proxy for inducing distance information of different objects in the scene, for the task of audio binauralization. We propose a novel encoder-decoder architecture with a hierarchical attention mechanism to encode image, depth and audio feature jointly. We design the network on top of state-of-the-art transformer networks for image and depth representation. We show empirically that the proposed method outperforms state-of-the-art methods comfortably for two challenging public datasets FAIR-Play and MUSIC-Stereo. We also demonstrate with qualitative results that the method is able to focus on the right information required for the task. The qualitative results are available at our project page https://krantiparida.github.io/projects/bmonobinaural.html

1. Introduction

Humans can infer approximate location of different objects by hearing the sound they emit. This is possible because of two ears and the separation between them. Due to this separation there is a difference in sound waves received by both ears in terms of amplitude and time. These differences, known as Interaural Level Difference (ILD) and Interaural Time Difference (ITD), are exploited by the brain to infer spatial properties eg. position of the sound source [28]. Thus, while audio recorded with a single microphone loses all such characteristics whereas *binaural* audio recreates original sound more accurately and gives the listener a feeling of being in the recording place.

The recording setup for binaural audio requires two microphones placed inside a dummy human head's ears. Such setup is closer to human hearing as it accurately models the sound reflection around the head and within the folds of the ear. Since binaural recording requires a full size dummy head, it is too bulky to be integrated into standard devices such as cameras or smartphones. However, we could get high quality binaural audio using standard handheld devices if we could lift mono audio to binaural audio.

The aim of this work is to tackle this problem of audio binauralization: take a mono channel audio as input and predict the corresponding two channel binaural audio. In most prior approaches [11, 16, 39, 36], the visual information in the form of RGB image is used, in addition to the mono audio, to predict the binaural audio. The RGB image serves as an important side information for encoding appearance of the sound producing sources and their relative locations in the scene. But most existing approaches ignore other important information, like the distance of the source from the microphone or the geometry of the scene. In [29, 12] similar information in the form of explicit position and orientation of both source and receiver were fed along with the audio input. This improved the performance of the system as compared to using RGB images only. However, doing so requires specialized equipment to track position of the source(s) as well as the listener, which is infeasible in general. We address this by using depth features of the scene along with the visual appearance features as auxiliary signals in the process of audio binauralization. Further, image, depth and binaural audio have also been shown to be interrelated in [24]. However, unlike [24], where the authors used binaural echoes to improve the depth prediction, here we perform the reverse task of using the depth features to obtain binaural audio. Here, depth features can be considered as a proxy for encoding both position information of sources and geometry of the scene.

As opposed to the prior approaches [11, 39], we use visual transformer [26] instead of convolutional layers as the backbone for extracting both visual and depth features. We propose a carefully designed cross-modal attention network to better associate different audio components present in the sound to the location and depth or the corresponding objects in the scene. We also separate magnitude and phase losses for the predicted audio. We do this as both these losses are very different mathematical function and operate in different range and factorizing them makes the learning easier. We evaluate our approach on two challening public datasest for the task, ie. FAIR-Play and MUSIC-Stereo. We show that our approach outperforms the previous stat-of-the-art approaches quantitatively, and produces meaningful interpretable qualitative results.

2. Related Work

Audio-Visual Binaurlization: Recently the task of audio binauralization has been attempted in a data-drive fashion [11, 16, 19, 39] as compared to earlier approaches that use signal processing techniques [14, 31, 40]. All the signal processing approaches model the system in the form of a Linear Time Invariant (LTI) system. In most of the cases [14, 40], HRTFs are measured and then convolutions are performed with them to get the final binaural audio. More recently, data driven approaches have been tried for the task. In all the recent data driven approach some form of image information as auxiliary data have been used. In [11], the authors have used RGB frame as side information for binaural audio generation. In [16], the authors have used both the RGB frame and optical flow along with audio features for binaural audio generation. In the similar lines, the authors in [19], have used both RGB and optical flow for generating full First Order Ambisonics for 360-deg. videos. In [39], the authors approached the problem of audio binauralization in a multi-task setting by combining the task of source separation with it. In [25], the authors have improved the task of audio binaurlization by performing localization of sound sources in the image. In [12, 29], the authors performed binauralization of speech and noise signal played using a speaker by explicitly using the position and orientation of source and listener along with the audio features. A preliminary investigation of the usefulness of depth features for the task of binauralization is also available in [23].

Binaural audio and Depth: There is an inherent interplay between binaural audio and the depth of the objects in the scene. As our aim in the paper is to improve the task of bianauralization using depth information, the reverse task, i.e. improving depth prediction from binaural audio has also been attempted. We give here some of the recent works in this line. In [4], the depth map of the scene is estimated directly from received bianural echoes. In [9, 24], authors have used the received binaural echoes along with images to improve upon the task of depth prediction from images. In similar lines, the authors in [34] have solved both the audio spatialization and depth prediction in multi-task framework. Here, instead of echoes, a two channel audio directly from the sound source is used for both depth estimation of the scene and audio spatialization, where the number of audio channels are increased to eight from two.

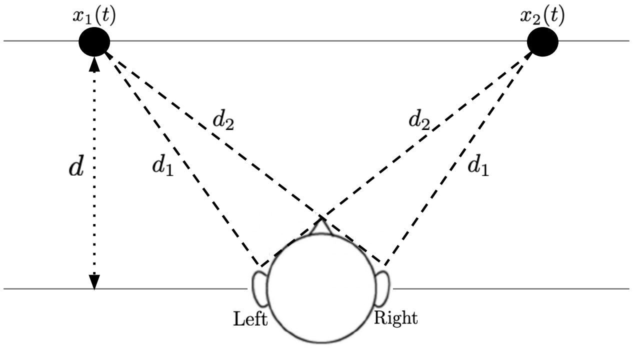

Figure 1. **Illustration of the concept.** $x_1(t)$ and $x_2(t)$ are two sound sources located at a distance d from the recording device. The sound received by the left and right ears will be different because of the head shape and the depth of the sound producing source, both in amplitude and time axes. Human brain exploits these differences for inferring the spatial information of the sources.

The depth of scene as an additional information has also been shown to be useful in other tasks such as image relighting [15], 3D pose estimation [35] etc.

Audio-Visual Learning: As our proposed work comes under the broad area of audio-visual processing, we give here some recent works in this area. In most of the works, semantic [2, 3, 20], temporal [21], and spatial correspondence [18, 37] between between both the modalities have been explored for learning features individually for each modality in a self-supervised manner. A separate stream of research have fused information from both audio and visual modality to improve upon tasks such as audio source separation [7, 8, 10, 38], sounding object localization [1, 13, 32], zero-shot learning [17, 22], saliency prediction [33].

3. Approach

Our task is to convert a mono channel audio, $x(t)$ to a binaural audio with $(y_l(t), y_r(t))$ as the left and right channels respectively. To achieve this, we design a transformer network based deep neural network with three input modalities, ie., RGB image, depth and mono channel audio. Using this multimodal network, we exploit inherent relationship between the two channels of audio and the sound source's distance and relative location in the scene.

Consider a simple case of two sound sources in the scene $x_1(t)$ and $x_2(t)$, at a depth d, with one in the extreme left of

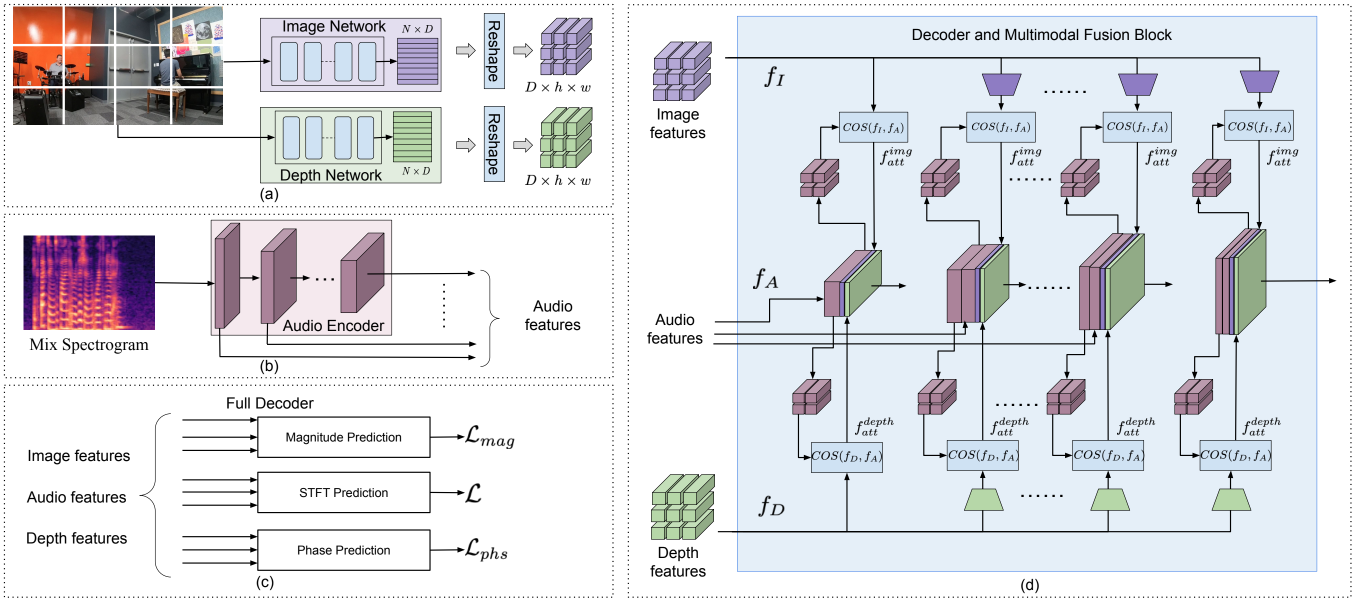

Figure 2. **Block diagram of proposed architecture.** The network takes mono audio and RGB image as input, and produces corresponding binaural audio consistent with the visual scene. (a) Image and Depth network input the same RGB image producing image and depth features, $f_{\mathbf{I}}$ and $f_{\mathbf{D}}$ respectively. Similarly, (b) audio encoder inputs mono audio producing audio features $f_{\mathbf{A}}$. (c) The image, audio and depth features are fed into individual decoder blocks having individual subnetworks to predict magnitude, STFT and phase of the difference of both channels. Each decoder has the same architecture. The detailed architecture of one such decoder is shown in (d). In each decoder, cross modal attention is computed for both image-audio and depth-audio at each layer. Both the attention outputs are then concatenated with audio features to obtain the final predicted binaural audio. We also use skip connections to concatenate features from audio encoder layer to each layer of the decoder.

visual field and other in the extreme right as in Fig.1. The mono audio, $x(t)$, is the combination of two sources, i.e. $x(t) = x_1(t) + x_2(t)$. Now, when the sound is received at each ear, there will be a time difference between the arrival of $x_1(t)$ and $x_2(t)$. For the left ear, $x_1(t)$ will arrive earlier than $x_2(t)$ and the reverse for the right ear. Assuming no reflecting and absorbing materials in the scene, the direct sound received at left and right ear can be modelled as

$$y_l(t) = \alpha_1 x_1(t - t_1) + \alpha_2 x_2(t - t_2) \quad (1)$$

$$y_r(t) = \alpha_1 x_1(t - t_2) + \alpha_2 x_2(t - t_1), \quad (2)$$

where, t_1, t_2 are the time delays with $t_1 < t_2$, and α_1, α_2 are the amplitude scaling factors. The time delays are symmetric wrt both ears because of the symmetric placement of sound sources. Let the distance for the left and right ear be d_1 and d_2 for source x_1. Similarly, the distance for right sound source will be d_2 and d_1 for left and right ear. There is a direct relationship between t_1 and d_1, i.e. $t_i = \frac{d_i}{v_s} \forall i = 1, 2$, where v_s is the velocity of sound in air. The amplitude scaling factor, α_1, α_2, also have a direct relationship with the distance as the wave attenuates more as it travels longer distance. The ITD and ILD described in the previous section are due to t_1, t_2 and α_1, α_2 respectively.

Hence, for the network to predict realistic binaural audio it should effectively model ITD and ILD. This depends upon the relative arrangement of sound sources and its distance from the recording device. Taking note of this fact, we use both depth and image features of the underlying visual scene to infuse depth and position information of different sound sources in the prediction process. To achieve this, we propose a network consisting of carefully designed cross-modal attention mechanism to associate features from RGB, mono channel audio and depth.

Following prior works [11, 39, 36], we use mix of both channels, $x_m(t) = x_l(t) + x_r(t)$ as input. The mixing of both channels looses spatial properties, and hence is a mono audio signal. For the output instead of predicting the individual left and right channels, we predict the difference between them, ie. $x_o(t) = x_l(t) - x_r(t)$. Finally, we perform simple arithmetic manipulation to get back the individual signals, where, $\hat{x}_l = \frac{x_m + \hat{x}_o}{2}$ and $\hat{x}_r = \frac{x_m - \hat{x}_o}{2}$. We use data in frequency domain by performing STFT on the time domain signal. We represent the STFT of input as $\mathbf{A} = \mathcal{F}(x_m) \in \mathbb{R}^{2 \times F \times T}$ and STFT of output as $\mathbf{O} = \mathcal{F}(x_o) \in \mathbb{R}^{2 \times F \times T}$. Further, we obtain the magnitude ($\mathbf{O}_{mag}$) and phase ($\mathbf{O}_{phs}$) of the complex output signal, where $\mathbf{O}_{mag} = \|\mathbf{O}\|_2$ and $\mathbf{O}_{phs} = \tan^{-1}(\frac{Re(\mathbf{O})}{Im(\mathbf{O})})$.

3.1. Overall Architecture

We show the overall architecture of our approach in Fig. 2. The network consists of (i) audio encoder network, (ii) image network, (iii) depth network and (iv) audio decoder network. The audio encoder network is a convolutional network that takes the mono audio as input and gives audio features as output. Both the image and depth network are self-attention transformer networks. Both the networks have the same architecture and take RGB image as input to produce image and depth features respectively. The audio decoder network further has three different sub network, where one outputs directly the complex STFT of the difference of both channels and other two predicts the magnitude and phase of the difference independently. We perform cross-modal attention both for image-audio and depth-audio features separately at each layer of all the subnetworks of audio decoder network. The output of the network is the difference between the right and left channel audios. This difference prediction impedes the networks tendency to copy the same audio to both the channels, as observed first by [11]. We now describe each component below.

3.2. Image and Depth Network

For extracting image and depth features from RGB image, $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, we use the recently proposed vision transformer (ViT) [6] backbone. We use ViT-Large architecture consisting of 24 attention blocks for both image and depth features. Following [26], we obtain features from four different layers of the transformer network, i.e. $l \in \{6, 12, 18, 24\}$, to get information at the varying level of details. Finally to align the number of channels of depth, image and audio networks, we perform a 1×1 convolution on features from each of the four layers resulting in d channels. We then concatenate features from all the four layers and obtain image and depth features $f_{\mathbf{I}} \in \mathbb{R}^{4d\times h\times w}$ and $f_{\mathbf{D}} \in \mathbb{R}^{4d\times h\times w}$ respectively, where $h \times w$ represents total number of patches in the image. We then use both the features as input for hierarchical attention calculation in the network. We initialize the image and depth network with ImageNet [5] and MIX6 [27] respectively.

3.3. Audio Encoder Network

Similar to [11], our audio encoder network consists of a UNet [30] style convolutional encoder architecture. We convert the time domain audio signal $x_m(t)$ into a STFT representation and concatenate both the real and imaginary to be fed as input to audio network, i.e. $\mathbf{A} \in \mathbb{R}^{2\times F\times T}$, where F and T are the no. of time-frequency bins in STFT. We then pass it through successive layers of convolutions. Finally, we obtain audio features, $f_{\mathbf{A}} \in \mathbb{R}^{d_A\times f\times t}$ as the output of audio encoder.

3.4. Audio Decoder and Multimodal Fusion

Audio Decoder. We adapt and build upon the audio decoder proposed in [11]. The decoder is further divided into three subnetworks and all share the same architecture whereas the output is different for each of the network. Each of the subnetwork in the decoder consists of 5 fractionally strided convolutional layers, which increases the dimension of input tensor successively at each layer. Each of the network takes the input from all three modalities. The first subnetwork, Mag subnetwork, predicts the magnitude of STFT of the difference signal, $|\tilde{\mathbf{O}}| \in \mathbb{R}^{1\times F\times T}$. The second network, STFT subnetwork, directly predicts the STFT of difference signal and produces a mask, $\mathbf{M} \in \mathbb{R}^{2\times F\times T}$, with values in the range $[-1, 1]$. We obtain the final output to be the difference between right and left channel audio, $\tilde{\mathbf{O}} \in \mathbb{R}^{2\times F\times T}$. We predict it by multiplying the mask, $\mathbf{M}$ with the mixed input signal, $\mathbf{A}$, i.e. $\tilde{\mathbf{O}} = \mathbf{M} \cdot \mathbf{A}$. The final subnetwork, Phs subnetwork, predicts the phase of the difference signal. Similar to STFT subnetwork, the Phs subnetwork predicts a mask with values in the range $[-1, 1]$ for each of the time-frequency bin in the spectrogram, i.e. $\mathbf{M}_p \in \mathbb{R}^{1\times F\times T}$. As the phase of any signal lie in the range $[-\pi, \pi]$, we multiply M_p with π to get the predicted phase, i.e. $\text{Phs}(\tilde{\mathbf{O}}) = \pi \times M_p$. At the time of prediction, we use the STFT subnetwork for obtaining the final output.

Multimodal Fusion. We perform similar multimodal fusion for all three subnetworks. The fusion operation combines the information from all three inputs at different scales. We show the fusion approach in the right side of Fig. 2. The image and depth features are already extracted from different layers. For audio, each time-frequency bin in the feature representation can be considered as an unique audio concept.Each of the audio concepts can act as basic building blocks of different audio sources present in the scene. These audio concepts represent different characteristics sound of the source. This characteristic sound can contain frequency variation within the source and a single audio concept can also contribute to multiple sound sources. E.g. the audio concept for a 7 string *guitar* can be the sound produced by each of the strings. Similarly an audio concept for *acoustic guitar* can also be shared by *electric or classical guitar* or by any similar sounding object such as *piano* or *saxophone*. So, our goal in multimodal fusion is to effectively associate different audio concepts to different object regions. Similar to the time-frequency bin in the audio representation, each co-ordinate in the spatial domain of the visual/depth features corresponds to certain region in the image. If the region contains a sounding object then the corresponding audio component should be weighted and also the depth value in the region should be used for the final output. We calculate two attention maps (i) between the image and audio features, and (ii) between the depth and audio features for fusing the information.

We design the network such that the output channels of audio encoder network is equal to the output channels of image and depth network, i.e. $d_I = d_D = d_A = 4d$. The attention is calculated between every pair of points.

$$f_{att}^{img}(i,j,k,l) = \frac{f_{\mathbf{I}}(:,i,j)^T f_{\mathbf{A}}(:,k,l)}{\sqrt{\|f_{\mathbf{I}}(:,i,j)\|_2^2}\sqrt{\|f_{\mathbf{A}}(:,k,l)\|_2^2}} \forall i,j,k,l \tag{3}$$

$$f_{att}^{depth}(i,j,k,l) = \frac{f_{\mathbf{D}}(:,i,j)^T f_{\mathbf{A}}(:,k,l)}{\sqrt{\|f_{\mathbf{D}}(:,i,j)\|_2^2}\sqrt{\|f_{\mathbf{A}}(:,k,l)\|_2^2}} \forall i,j,k,l \tag{4}$$

where, $f_{att}^{img}, f_{att}^{depth} \in \mathbb{R}^{h\times w\times f\times t}$ are the image-audio and depth-audio attention respectively. We then resize the 4D attention tensors into a 3D tensors such that the resulting attention maps are of size $[(h \times w) \times f_i \times t_i]$, where f_i, t_i are the input spatial feature dimension in the i^{th} layer of decoder. For the first decoder layer f_i, t_i is exactly equal to the dimension of the audio encoder output. After resizing operation the number of channel dimension in the feature map corresponds to all the distinct regions in the image. We interpret this as the attention weight in all the regions of the image for the particular audio concept. We obtain final attention map at i^{th} decoder level by concatenating both of them over the spatial axis.

$$f_{att}^{i} = \text{Concat}(\text{Resize}(f_{att}^{img}, f_{att}^{depth})) \tag{5}$$

Next, we concatenate the attention map with the audio features and feed the result to the next layer of the decoder. We also use skip connection at each decoder layer and concatenate features from corresponding encoder layer except for the first layer of decoder. As audio network increases the feature dimension in each level, it also increases the time-frequency bins in the feature representation and hence the finer details in the audio comes successively with each decoding layer. In order to account for the coarse to fine representation of the audio we add the image and depth features similar to first layer to obtain the attention map. To perform the attention calculation at each layer, the feature channel of the image and depth should align with the channels of audio features. To perform the alignment we use a one-layer neural network followed by GELU non-linearity to make the feature dimensions of both the modalities equal. For matching the channels of audio and image feature, the one-layer neural network used for every layer has weights of dimension $[d_I, d_i]$ and $[d_D, d_i]$ for image and depth features respectively. Please note that for first layer of decoder, i.e. $i = 1$, we do not use one-layer network as the channels are already aligned.

3.5. Loss Function and Training

We use three individual losses for each of the subnetworks in the decoder. For the STFT subnetwork, following earlier works [11, 39], we use an L2 loss between the ground truth and network output, given as

$$\mathcal{L}(\hat{\mathbf{O}}, \mathbf{O}) = \|\hat{\mathbf{O}} - \mathbf{O}\|_2^2 \tag{6}$$

where, $\hat{\mathbf{O}}, \mathbf{O}$ are ground truth and predicted difference between the left and right channel audio. Similar to the STFT subnetwork, we also minimize the L2 loss for magnitude and phase, given as

$$\mathcal{L}_{mag}(\hat{\mathbf{O}}_{mag}, \mathbf{O}_{mag}) = \|\hat{\mathbf{O}}_{mag} - \mathbf{O}_{mag}\|_2^2 \tag{7}$$

$$\mathcal{L}_{phs}(\hat{\mathbf{O}}_{phs}, \mathbf{O}_{phs}) = \|\hat{\mathbf{O}}_{phs} - \mathbf{O}_{phs}\|_2^2 \tag{8}$$

where, $\hat{\mathbf{O}}_{mag}, \hat{\mathbf{O}}_{phs}$ are the predicted magnitude and phase obtained from the respective network, $\mathbf{O}_{mag}, \mathbf{O}_{phs}$ are the magnitude and phase of the ground truth signal. It is to be noted here that minimizing STFT loss in eq. 6 also implicitly minimizes magnitude and phase. We have added individual magnitude and phase loss possibly penalizing each term twice as this was found to be helpful in prior work [29]. Further to enforce the reconstruction of magnitude and phase is correct, we add a reconstruction loss. Here, we reconstruct back the real and imaginary part of the spectrogram and force it to be closer to the ground truth. We calculate the reconstruction loss $\mathcal{L}_{rec}$ by estimating the real and imaginary part of the spectrogram using $\hat{\mathbf{O}}_{mag}\text{Cos}(\hat{\mathbf{O}}_{phs})$ and $\hat{\mathbf{O}}_{mag}\text{Sin}(\hat{\mathbf{O}}_{phs)}$ respectively. We obtain the reconstructed STFT, $\hat{\mathbf{O}}$ by concatenating both the real and imaginary channels. The reconstruction loss is the L2 loss between the reconstructed STFT, $\hat{\mathbf{O}}$ and original STFT $\mathbf{O}$.

$$\mathcal{L}_{rec} = \|\hat{\mathbf{O}}_{mag} e^{i\hat{\mathbf{O}}_{phs}} - \mathbf{O}\|_2^2 \tag{9}$$

The final loss function used for training is weighted combination of all the losses, given as

$$\mathcal{L}_{tot} = \mathcal{L} + \alpha_{mag}\mathcal{L}_{mag} + \alpha_{phs}\mathcal{L}_{phs} + \alpha_{rec}\mathcal{L}_{rec} \tag{10}$$

where, $\alpha_{mag}, \alpha_{phs}, \alpha_{rec}$ are the hyperparameter denoting weights of individual loss and are set empirically. We train the whole network in an end-to-end manner.

4. Experiments

Dataset. We report results on two dataset FAIR-Play and MUSIC-Stereo. For the FAIR-Play dataset [11], we use the five new splits as proposed in [36] for our experiments. The dataset of Music-Stereo was proposed in [36] by combining two existing datasets MUSIC-21 and MUSIC-duet proposed in [38] originally for the task of source separation. As the youtube IDs for MUSIC-Stereo is not publicly available, we select the binaural videos only from MUSIC-21 and MUSIC-duet as mentioned in [36]. In order to select the binaural videos from both the dataset, we calculate sum of the difference of left and right channel audio and then set a threshold of 0.001 for selecting videos with binaural audio. We considered those that have sum of difference more than 0.001 between both channels as binaural and discarded the rest. We obtained 713 unique videos to have binaural audio.

Modality	STFT (↓)	ENV (↓)	Mag (↓)	Phs (↓)	SNR (↑)
audio	1.337	0.166	2.674	1.560	5.01
+image	1.332	0.161	2.665	1.499	5.102
+depth	1.334	0.165	2.668	1.553	5.036
+image+depth	**1.158**	**0.155**	**2.316**	**1.487**	**5.670**

Table 1. **Audio binauralization by combining different modalities.** Using audio only (audio), audio with image features (+image), audio with depth features (+depth) and combination of audio, image and depth features (+image+depth). ↓ and ↑ indicates lower is better and higher is better respectively.

We then divide the videos into 80-10-10 into train, validation and test. Following the setting of [36], we split the videos into 10 second clips and we obtain a total of 15026 clips, around 8x more than FAIR-Play. For input data representation and preprocessing, we follow the same settings as described in earlier works [11, 39, 36].

Metrics. Following [36], we use five different metric for evaluation. STFT distance measures the squared L_2 distance between short time Fourier transform (STFT) of each channels for ground truth and predicted audio. Envelope (ENV) distance measures the L_2 difference of the envelope of ground truth and predicted audio for both channels, where we calculate the envelope from time-domain audio signal using Hilbert transform. Similar to STFT distance, we obtain Magnitude (Mag) distance by calculating the squared L_2 distance between the magnitude of STFT for both channels of ground truth and predicted audio. For Phase (Phs), we measure the L_1 difference between the phase of ground truth and predicted difference signal. We also report the Signal-to-noise ratio (SNR) for the predicted binaural audio, where signal refers to the ground truth binaural audio and noise refers to the distance between ground truth and prediction.

4.1. Ablation Study

In this section, we give the ablation results of our approach to demonstrate following points. (i) Impact of depth features over the image features. (ii)Contribution of each term of the loss function on the performance.

Impact of adding depth. To analyse the effectiveness of depth for the task of audio spatialization, we study the impact of each modality on the performance of the network. We add each of the modality (image, depth) one by one to the network and then combine all the modalities to verify the contribution of each modality. For a fair comparison, we use exactly the same transformer architecture with equal number of parameters for both image and depth. The results are shown in Tab. 1. We report the performance in the *split-1* of modified FAIR-PLAY dataset for all the models.

From Tab. 1, we observe that there is a improvement in performance for all the metric with image + mono audio as input over an audio only input (Tab. 1, row 1 vs row 2). The value of STFT, ENV, Mag, Phs decreases from $1.337, 0.166, 2.674, 1.560$ to $1.332, 0.161, 2.665, 1.499$ whereas the SNR value increases from 5.01 to 5.102. Similarly, we observe a decrease of STFT, ENV, Mag and Phs to $1.334, 0.165, 2.668, 1.553, 5.036$ from $1.337, 0.166, 2.674, 1.560$ and increase of SNR to 5.036 from 5.01 by using depth + mono audio as input as compared to audio input only (Tab 1, row 1 vs row 3). This shows that both the image and depth features are helpful towards a better audio binauralization. As both the image and depth backbone contain exactly same number of parameters, the improvement in performance can be attributed to the information encoded in it. Adding image results in a better performance in all the metrics over mono audio only input as compared to depth. This could be owed to the presence of semantic information in the RGB images in the form of appearance and relative location of different sound producing regions. Although adding depth information alone doesn't perform as good as to the approach of adding image information only but it performs better than the approach of using mono audio only as input. This is possibly due to the fact that depth input has relative distance information within the scene, and results in better binauralization as compared to mono audio input. From this observation, we hypothesize that combining depth with RGB will provide more contextual information leading to better localization of sound sources by the network and in turn better performance in binauralization task. This is also evident with the empirical performance of adding both depth and image features along with audio, which results in a significant improvement in performance in all the metrics. There is an an improvement of 13% for STFT, $\sim 6\%$ for ENV, $\sim 13\%$ for Mag, $\sim 5\%$ for phs and $\sim 13\%$ for SNR over mono audio input (Tab 1, row 1 vs row 4). This observation confirms that both image and depth information are helpful for the task of binauralization.

$\mathcal{L}$	$\mathcal{L}_{mag}$	$\mathcal{L}_{phs}$	$\mathcal{L}_{rec}$	STFT (↓)	ENV (↓)	Mag (↓)	Phs (↓)	SNR (↑)
✓	✗	✗	✗	1.206	0.158	2.413	1.488	5.418
✓	✓	✓	✗	1.185	0.157	2.401	1.481	5.497
✓	✗	✗	✓	1.190	0.158	2.411	1.487	5.435
✓	✓	✓	✓	**1.171**	**0.156**	**2.342**	**1.478**	**5.573**

Table 2. **Contribution of different losses on performance.** Performance after applying different combination of losses. We observe that adding all the losses gives the best performance. ↓ and ↑ indicates lower is better and higher is better respectively.

Contribution of different Losses. In order to get the contribution of individual losses in the final performance of the network, we add various combinations of the loss. We report the performance in *split-1* after 50 epochs in Tab. 2. From Tab. 2, we observe that all the losses contribute equally to the performance. We observe that adding both magnitude and phase loss improves all the metrics, i.e. STFT, ENV, Mag, Phs, SNR from $1.026, 0.158, 2.413, 1.488, 5.418$ to $1.185, 0.157, 2.401, 1.481, 5.497$ respectively (row 1 vs row2 in Tab. 2). This proves that minimizing explicit mag-

	FAIR-Play					MUSIC-Stereo				
Method	STFT ↓	ENV ↓	Mag ↓	Phs ↓	SNR ↑	STFT ↓	ENV ↓	Mag ↓	Phs ↓	SNR ↑
Mono-Mono [36]	1.024	0.145	2.049	1.571	4.968	1.014	0.144	2.027	1.568	7.858
Mono2Binaural [11, 36]	0.917	**0.137**	1.835	1.504	5.203	0.942	0.138	1.885	1.550	8.255
PseudoBinaural (w/o sep.) [36]	0.951	0.140	1.914	1.539	5.037	0.953	0.139	1.902	1.564	8.129
PseudoBinaural [36]	0.944	0.139	1.901	1.522	5.124	0.943	0.139	1.886	1.562	8.198
Ours	**0.909**	0.139	**1.819**	**1.479**	**6.397**	**0.670**	**0.108**	**1.340**	**1.538**	**10.754**
Sep-Stereo [39, 36]	0.906	0.136	1.811	1.495	5.221	0.929	0.135	1.803	1.544	8.306
Augment-PseudoBinaural [36]	0.878	0.134	1.768	1.467	5.316	0.891	0.132	1.762	1.539	8.419

Table 3. **Comparison with existing approaches** We report the results for existing approaches directly from [39]. ↓ indicates lower is better and ↑ indicates higher is better. The method in the last two rows uses atleast 2x more data than ours and solve both task of audio binauralization and source separation jointly. Our approach outperforms other approaches in similar setting in almost all the metric and also performs comparably to superior approaches mentioned in the last two rows of the table for FAIR-Play dataset and even outperforms for MUSIC-Stereo dataset.

nitude and phase loss improves the performance, also consistent with the observation reported in [29]. When adding only the reconstruction loss without magnitude and phase loss in comparison to adding explicit magnitude and phase loss performance drops to $1.190, 0.158, 2.411, 1.487, 5.435$ from $1.185, 0.157, 2.401, 1.481, 5.497$ for all the metrics i.e. STFT, ENV, Mag, Phs and SNR respectively. This observation shows that minimizing individual magnitude and phase loss is more relevant for the task as compared to using reconstruction loss. We conclude that as both the losses, i.e. magnitude and phase are calculated from very different mathematical function (magnitude is a quadratic function of real and imaginary value where as phase is a trigonometric function) and are also in very different value range, hence separating them into individual component helps in the training process. Finally adding all the losses gives the best performance in all the metrics with STFT, ENV, Mag, Phs and SNR of $1.171, 0.156, 2.342, 1.478$ and 5.573 respectively.

4.2. Comparison to state-of-the-art

Baseline and prior approaches. We compare our approach against various baselines and competitive approaches on both the datasets. The baseline of Mono-Mono is a simple approach where the input audio is copied for both the left and right channel. The existing approach of Mono2Binaural [11] uses only the image features along with audio features as the input to the decoder and a simple concatenation method for fusing both the features. In another existing approach of sep-stereo [39], a multi-task approach of binaural prediction and source separation is trained jointly using a single backbone. The data used for training is also atleast 2x more than the amount of data used for training our method or Mono2Binaural approach. Although this approach is not directly comparable to ours as it uses more data and also solves multiple tasks jointly, this can serve as an upper bound for us. In one of the recent self-supervised approach, PseudoBinaural [36] generated from mono audio were used for training instead of the recorded ones. There are a number of variants to this approach, in PseudoBinaural (w/o sep.) the generated audios were used for training the binaural prediction task only whereas in PseudoBinaural the generated binaural audio is used for training both binauralization and source separation like sep-stereo [36]. Finally, in Augment-PseudoBinaural both the real audios and the generated ones were used for training both binauralization and separation task jointly. This is the most superior method as it solves both tasks and also uses 2x more data as compared to sep-stereo and 4x more data than ours. Similar to sep-stereo this method is not directly comparable to ours, we report it here as an upperbound for our case.

Comparison on FAIR-Play dataset. We report the results on FAIR-Play dataset for different existing and baseline approaches along with the proposed approach in Tab.3. We report the results for all the methods by averaging over all the five splits in the modified FAIR-Play dataset. We report the results on baseline and all the prior approaches directly from [36]. We observe that our method outperforms all the existing methods that is trained for a single task and dataset similar to ours. We observe that our propose method obtains an performance improvement of $\sim 11\%, \sim 4\%, \sim 11\%, \sim 6\%$ and $\sim 29\%$ for STFT, ENV, Mag, Phs and SNR respectively over the baseline of Mono-Mono. Also, our method outperforms the best performing method in similar setting, Mono2Binaural in four metrics out of five. We obtain STFT and magnitude value for our approach to be 0.909 and 1.819 respectively as compared to 0.917 and 1.835 for Mono2Binaural. We also obtain the value of 1.479 and 6.397 for Phs and SNR outperforming Mono2Binaural as well. We also observe that the performance of our approach is also in the similar ball park of approaches that solves multiple tasks and uses multiple datasets. The performance of our approach in the three metric of STFT, ENV, Mag is only worse by $\sim 0.3\%, \sim 2\%, \sim 0.4\%$ for Sep-Stero [39] where as we outperform it on the rest two metrics Phs and SNR by $\sim 2\%$ and $\sim 18\%$ respectively. Finally, for the Augment-

Pseudo we are worse in four metric STFT, ENV, Mag by $\sim$ 3% and Phs by $\sim$ 0.8%. But we outperform even this method on SNR metric by $\sim$ 17%. This proves that our method is competitive enough for the task of binaurlization as it outperforms superior approaches in some of the metrics.

Comparison on MUSIC-Stereo dataset. We also report the performance of our approach along with baseline and existing approach in Tab. 3. Similar to the FAIR-Play dataset, all the results of prior approaches and baseline are reported directly from [36]. We observe that our method outperforms significantly both baseline and other approaches in similar setting. There is an improvement of $\sim$ 34%, $\sim$ 25%, $\sim$ 34%, $\sim$ 2% and $\sim$ 36% in the metrics of STFT, ENV, Mag, Phs and SNR respectively over the baseline. The proposed approach also outperforms the best method in similar setting, i.e. Mono2Binaural by $\sim$ 29%, $\sim$ 22%, $\sim$ 29%, $\sim$ 0.8% and $\sim$ 30% for all the metrics STFT, ENV, Mag, Phs and SNR respectively. Further, we observe that our method also significantly outperforms the mulit-task and larger data approaches, i.e. sep-stereo and Augment-PseudoBinaural, in all the five metrics as well. The Music-Stereo dataset contains diverse in the wild music videos from Youtube whereas FAIR-Play dataset contains videos where all of them are recorded inside the same recording room with very minimal variation in the background. Our proposed approach outperforms existing method by a higher margin in MUSIC-Stereo dataset as compared FAIR-Play dataset, which suggests that our method generalizes well to unconstrained setting.

4.3. Qualitative Results

We give qualitative results of visual and depth attention map obtained from all layers of decoder in Fig. 3. We provide the input image in the first row for comparison. From the visual attention map in first column, we observe that the attention values are spread out over the entire image in the first layer but in successive layers of 2,3, and 4 it produces high values only to the sound sources. We also observe that layer 3 produces high values for the source on the right side of the image whereas layer 4 produces high values for the source in the left side of the image. This region specific attention map can be considered as the inherent association between left and right audio channel with left and right regions of the image, which is important for an effective binauralization. For the depth attention maps, we observe that instead of attending to the sound source location, it looks at different structure of the rooms such as wall, ceiling and floor in layer 1, 2, and 3 of the decoder. From these attention maps, we make a general observation that the depth network infuses information about the geometry of the room resulting in better binauralization. We also provide predicted binaural results in our project page and request readers to listen to the videos to have a sense of the reconstruction.

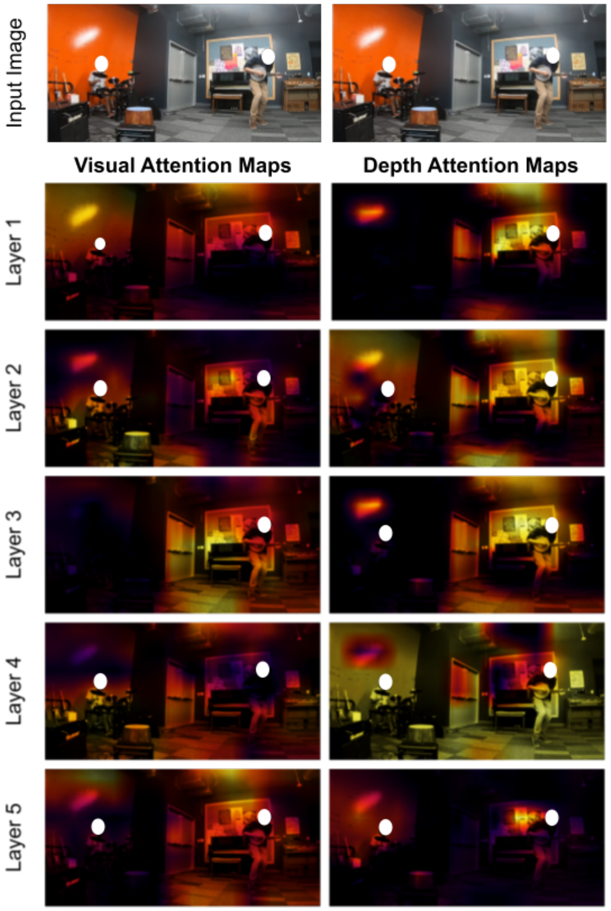

Figure 3. **Attention Map Visualization** Attention maps for both visual and depth channel at each decoder layer on FAIR-Play dataset. The first row shows the input image. We observe that visual attention map progressively attends to the sound producing regions in the image where as the depth attention maps attends to the structure of the room, i.e. wall, ceiling and floor.

5. Conclusion

We proposed an end-to-end trainable multi-modal transformer network with hierarchical multi-modal attention, for mono to binaural audio generation. We studied the impact of image and depth inputs along with their combinations on this task. We demonstrated that adding depth provides additional structural information which significantly improves audio binauralization quantitatively and aids in better source localization qualitatively, as visually analysed from attention maps. The proposed method obtains state-of-the-art results on two challenging datasets (FAIR-Play and MUSIC-Stereo) for the task.

Acknowledgment. Kranti Kumar Parida gratefully acknowledges support from the Visvesvaraya PhD Fellowship.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020.

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, pages 435–451, 2018.

[4] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *IEEE International Conference on Robotics and Automation*, pages 1581–1587, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[7] 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. *ACM Transactions on Graphics*, 2018.

[8] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[9] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pages 658–676. Springer, 2020.

[10] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *European Conference on Computer Vision*, pages 35–53, 2018.

[11] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019.

[12] Israel D Gebru, Dejan Marković, Alexander Richard, Steven Krenn, Gladstone A Butler, Fernando De la Torre, and Yaser Sheikh. Implicit hrtf modeling using temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3385–3389. IEEE, 2021.

[13] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33, 2020.

[14] HE Jianjun, Ee Leng Tan, Woon-Seng Gan, et al. Natural sound rendering for headphones: integration of signal processing techniques. *IEEE Signal Processing Magazine*, 32(2):100–113, 2015.

[15] Yang Liu, Alexandros Neophytou, Sunando Sengupta, and Eric Sommerlade. Relighting images in the wild with a self-supervised siamese auto-encoder. In *IEEE Winter Conference on Applications of Computer Vision*, pages 32–40, 2021.

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

[17] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *IEEE Winter Conference on Applications of Computer Vision*, 2021.

[18] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*, 2018.

[20] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[21] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, pages 631–648, 2018.

[22] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *IEEE Winter Conference on Applications of Computer Vision*, pages 3251–3260, 2020.

[23] Kranti Kumar Parida, Siddharth Srivastava, Neeraj Matiyali, and Gaurav Sharma. Depth infused binaural audio generation using hierarchical cross-modal attention. *arXiv preprint arXiv:2108.04906*, 2021.

[24] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond image to depth: Improving depth prediction using echoes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8268–8277, 2021.

[25] Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. Localize to binauralize: Audio spatialization from visual sound source localization. In *IEEE International Conference on Computer Vision*, pages 1930–1939, 2021.

[26] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE International Conference on Computer Vision*, pages 12179–12188, 2021.

[27] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

[29] Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Butler, Fernando Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2020.

[30] 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.

[31] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Ritta Väänänen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705, 1999.

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

[33] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4766–4776, 2020.

[34] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *European Conference on Computer Vision*, pages 638–655. Springer, 2020.

[35] Erwin Wu and Hideki Koike. Futurepose-mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1384–1392, 2019.

[36] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021.

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

[38] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *European conference on computer vision*, pages 570–586, 2018.

[39] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision*, pages 52–69. Springer, 2020.

[40] Dmitry N Zotkin, Ramani Duraiswami, and Larry S Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6(4):553–564, 2004.