The Thesis Committee for Rishabh Garg
Certifies that this is the approved version of the following Thesis:

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

APPROVED BY
SUPERVISING COMMITTEE:

Kristen Grauman, Supervisor

David Harwath

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

by

Rishabh Garg

Thesis

Presented to the Faculty of the Graduate School
of The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in Computer Sciences

The University of Texas at Austin

May 2021

Acknowledgments

I would like to express my profound gratitude to my advisor, Prof. Kristen Grauman. Her guidance, support and keen insights have been vital for the completion of this thesis and are a constant source of inspiration for me.

I would also like to thank Dr. David Harwath for his helpful comments and feedback.

I am also extremely fortunate and thankful to have Ruohan Gao as my mentor. Without his constant advice, readiness to answer all questions, and guidance at every step, this work would not have been possible.

Finally, I would like to thank my labmates for the helpful feedback and discussions, and my family and friends for their continuous support and encouragement.

Abstract

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

Rishabh Garg, M.S.Comp.Sci.

The University of Texas at Austin, 2021

Supervisor: Kristen Grauman

Human audio perception is inherently spatial and videos with binaural audio simulate the spatial experience by delivering different sounds to both ears. However, videos are typically recorded with mono audio and hence generally do not offer the rich audio experience of binaural audio. We propose an audio spatialization method that uses the visual information in videos to convert mono audio to binaural. We leverage the spatial and geometric information about the audio present in the visuals of the video to guide the learning process. We learn these geometry aware features in visuals in a multi-task manner to generate rich binaural audio. We also generate a large video dataset with binaural audio in photorealistic environments to better understand and evaluate the task. We demonstrate the efficacy of our method to generate better binaural audio by learning more spatially coherent visual features by extensive evaluation on two datasets.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human perception is complex and multi-sensory. We constantly capture sensory data from various sources to understand an environment, of which sight and sound are the vital components. Our brains continuously use both audio and visual information in conjunction to help us navigate the world and our surroundings, as these modalities inherently contain rich spatial information. For example, let's say we are in an environment where a person walks by us. Even if we close our eyes, our brains still use the sounds from the footsteps to give us a fair idea of the position of the person with respect to us as he walks away. This is possible because we hear *binaural* audio from our two ears, which is intrinsically spatial and helps us navigate in a 3D world and implicitly localize sound.

It is believed that this spatialization can occur because of the Interaural Level Difference, which is difference in the sound levels for each ear, and the Interaural Time Difference, which is the difference in time between the sounds reaching each ear [Rayleigh, 1875]. In addition to this, the shape of the head and pinna filtering effect can also affect how we perceive sound. In parallel, our visual perception complements this by locating the person and identifying that the sound we heard is from that particular source.

In addition to spatial cues that help us in locating the sound sources, audio also contains cues that provide context of the surroundings and locations. These are in the form of reflections and reverberations from the environment, helping us understand the orientation and material of walls and the room we are in. For example, due to these effects, we can intuitively interpret the difference in sound if the same audio is heard in a long corridor versus a large room, or a room with heavy carpets and drapes versus a room with smooth marble surfaces. The sounds therefore provide information about the geometry of the room and the materials in the surroundings as it propagates from the sound source at a location to the receiver.

Videos or other media with binaural audio similarly imitate that rich audio

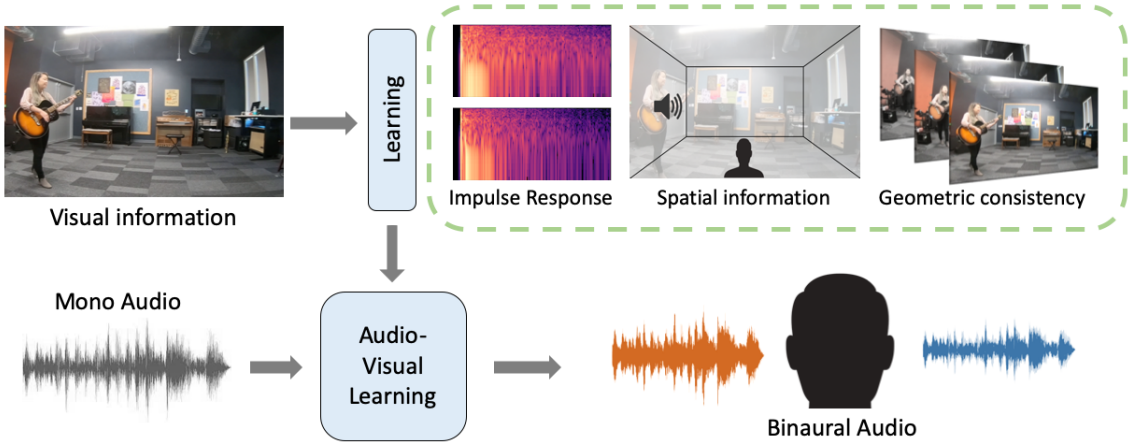

Figure 1.1: To generate accurate binaural audio from mono audio, the visuals provide significant cues that can be learnt jointly with audio prediction. We can learn to extract information like spatial information that the guitar player is on the left, geometric consistency of the position of the guitar player over time, and basic information about the binaural impulse response from the room.

experience for the user and make the media feel more real and immersive. With virtual reality and augmented reality applications becoming more ubiquitous, binaural audio is becoming more essential for real world applications. However, currently collecting binaural data is a challenge. Presently, such audio is collected either using an array of microphones or using dummies that imitate the human ears and head. The process is therefore less accessible, requiring special equipment, and more costly, leading to less availability of large amounts of such data. On the other hand, due to the ubiquity and ease of capture of videos from mobile devices, mono audio from a single microphone is readily available.

Recent work explores how monaural audio can be upgraded to binaural audio by leveraging the visual stream in videos [Morgado et al., 2018, Gao and Grauman, 2019a, Zhou et al., 2020]. Mono audio contains little spatial information on its own so it is not possible to get spatial information without some context. The premise in prior work is to provide the context to generate rich binaural audio from mono sources via visual information from videos, since videos have visual frames with associated corresponding audio. Just as audio data provides hints about the physical surroundings, visual data from videos also has vital information about sound making objects and their locations, as well as the room and environment they

are in and the object configurations. For instance, if we have a video of a person playing a guitar on the left side of the frame, while the mono audio alone is not sufficient, using the visual frames we can form a reasonable guess of the audio we would hear if we were at the camera location, which is that the audio from the guitar would be more prominent in the left ear. Past efforts have been focused on using the visual features to implicitly infer some spatial characteristics to lift a single audio channel to two spatial channels. However, the visuals provide more context than simple generic visual features.

In this work, we go beyond using basic visual features and guide the binauralization process via multi-task learning to look for geometric cues of the environment that dictate how a listener receives the sound in the real world.In particular, we propose three ways to incorporate geometric audio-visual cues: (1) predicting the *impulse response*, (2) predicting the visual stream's coherence with the spatial location of the sound, and (3) the consistency in geometry of the objects across time. See Figure 1.1.

First, we consider impulse response prediction to regularize learning. An impulse response (IR) is a concise acoustic description of the environment capturing the ways in which a sound wave will interact with the room and materials, as it propagates from a sound source to a listener at a particular location. It encapsulates all geometric information required for perfect reconstruction of binaural audio. The visual frames convey a lot of information like the layout of the room and the sound source with respect to the receiver, which form the basis of the IR. We predict this IR from a single frame in a multi-task setting with the binauralization task, thereby encoding this information into a latent geometric code to generate better binaural audio.

Second, we account for audio-visual spatial coherence. The agreement of audio with the visuals forms a key part of our perception of a video. Hence, we also want to ensure that the audio predicted by our model understands the relation between the frames on the one hand, and the left and right channel audio on the other hand. To achieve this, we constrain the visual features such that they have *spatially coherent* information, i.e., they can understand the difference when the audio is aligned with the visuals and when it is is not.

The third component of the proposed multi-task objective captures geometric

consistency of objects over time. Current methods treat the audio and visual frame pairs from videos as independent samples. However, since these are videos, objects do not generally have dramatic instantaneous change in their layout or the geometry of the scene. This implies that the visual features representing these characteristics that we utilise for guiding binauralization should have some consistency across short periods of time. This spatio-temporal consistency means that the geometric arrangement of the key visual components is fairly consistent across time, and hence the resulting audio-visuals are too.

In addition to our novel multi-task approach, a secondary contribution of this work is a large-scale simulated dataset, SimBinaural, to support binauralization research. Binaural audio and impulse response collection in the real world is expensive and challenging, requiring special equipment. It is limited by the fact that it differs for each physical space and source-receiver position pair within that space. Thus even if it is captured, the variety in the data is severely limited to a few positions and rooms. To facilitate and understand the relation between the visuals and the room impulse response for learning better geometric features, we create a large-scale dataset of simulated videos in photo-realistic 3D indoor scene environments which resemble real world audio recordings in a room. This dataset facilitates both learning and quantitative evaluation, and allows us to explore the impact of particular parameters (e.g., distance of source to receiver in a video) in a controlled manner.

To recap, the main contributions of this work are as follows. We provide a novel way to improve binauralization from mono audio using videos by learning better audio-visual representations. We guide the visual network to encode information that captures spatial and temporal cues to perform more accurate and richer sound. Finally, we demonstrate the efficacy of our method by achieving state-of-the-art results on two datasets.

Chapter 2

Related Work

The joint learning of audio and visual information has been studied for a variety of different tasks for a long time, and has received more attention recently with the wide availability of large amounts of video data.

2.1 Audio-Visual Learning

In recent years, researchers have tried to perform cross-modal learning on different audio-visual tasks, particularly from videos, to understand the natural synchronisation between visuals and the audio [Aytar et al., 2016, Arandjelovic and Zisserman, 2017, Owens et al., 2016b]. Audio-visual data can convey varied types of information which has been leveraged in a multitude of ways for tasks like audio-visual navigation [Gan et al., 2020c, Chen et al., 2020b, Chen et al., 2020c], multi-modal action recognition [Kazakos et al., 2019, Gao et al., 2020b, Long et al., 2018, Lee et al., 2021, Nagrani et al., 2020], audio-visual speech recognition [Hu et al., 2016, Chung et al., 2017, Zhou et al., 2019a, Yu et al., 2020], audio-visual event localisation [Tian et al., 2018, Tian et al., 2020, Wu et al., 2019], self-supervised representation learning [Owens et al., 2016b, Owens and Efros, 2018, Korbar et al., 2018, Gao et al., 2020a, Morgado et al., 2020], audio inpainting [Zhou et al., 2019b], and generating sounds from video [Owens et al., 2016a, Zhou et al., 2018, Gan et al., 2020a, Chen et al., 2020d].

The audio-visual source separation task has been well studied. Researchers have explored audio-visual separation applied to different areas like separation of speech [Ephrat et al., 2018, Owens and Efros, 2018, Afouras et al., 2018, Gabbay et al., 2018, Afouras et al., 2019, Chung et al., 2020], music [Zhao et al., 2018, Zhao et al., 2019, Xu et al., 2019, Gan et al., 2020b, Gao and Grauman, 2019b], and objects [Gao et al., 2018, Gao and Grauman, 2019b, Tzinis et al., 2021]. These tasks generally try to isolate a particular sound from an audio clip with a mixture of sounds, using

cues from visuals and audio. In contrast, we do not isolate particular sounds from a mixture, but perform a different task to produce binaural two channel audio from a mono audio clip using the visuals.

There have been several methods which have been proposed to explicitly identify sound-making regions from visuals and localize the pixels in the image/video [Kidron et al., 2005, Hershey and Movellan, 2000, Arandjelović and Zisserman, 2018, Zhao et al., 2018, Senocak et al., 2018, Tian et al., 2018, Hu et al., 2020, Rouditchenko et al., 2019]. Unlike these methods, we do not aim to localise sounds within a frame but rather ensure that some information about the spatial positioning is learnt by the visual features.

2.2 Audio Spatialization

Mono audio by itself does not have enough information for spatialization to two channel binaural audio, as discussed. To overcome this, recent works have proposed using video frames to provide a kind of self-supervision to implicitly infer the relative positions of sound-making objects. They formulate the problem as an upmixing task from mono to binaural using the visual information. [Morgado et al., 2018] use 360 videos from YouTube to predict first order ambisonic sound useful for 360 viewing. [Lu et al., 2019] use a self-supervised audio spatialization network using visual frames and optical flow. They incorporate a spatial correspondence classifier as an auxiliary loss by classifying if the visuals correspond to the audio channels or are swapped, and they test on a YouTube dataset they collected. While [Morgado et al., 2018] are limited to ambisonics and do not predict binaural audio or use normal field of view videos, [Lu et al., 2019] use optical flow to help the task.

More closely related to our problem, the first work to generate binaural audio from video is the 2.5D visual sound work by [Gao and Grauman, 2019a]. They collected a binaural video dataset FAIR-Play for the task. They generated binaural audio using mono audio conditioned on a visual frame. The visual features can contribute to the audio generation by providing context of objects. The visual information is added via simple concatenation to the audio feature. This has inspired a line of similar work. [Zhou et al., 2020] pose the sound source separation task as an extreme case of creating binaural audio. They propose an associative

pyramid network (APNet) architecture to fuse the modalities and jointly train on audio spatialization and source separation task. Building on these methods, we introduce additional supervision to help guide the visual features so that they encode more spatial and geometric information inherent in videos.

More recently, [Richard et al., 2021] tackle the problem without using the visuals. For their method, they render binaural audio waveform output directly which is conditioned on the ground truth relative position and orientation of the listener with respect to the source. Different from us, their method is limited to speech synthesis and requires knowing the actual physical position and orientation of the source and receiver as opposed to inferring them from video.

Concurrent to our work, [Xu et al., 2021] proposed to generate binaural audio for training from mono audio, by using video crops and spherical harmonics to map audio to specific locations. In contrast to their method, we generate a large scale realistic video dataset which can provide a large amount of accurate binaural information which can further improve the results.

2.3 Audio and Geometry/3D Spaces

Recent works have also tried to exploit the complementary nature of audio and the characteristics of the environment in which it is heard or recorded. [Schissler et al., 2017] estimate the acoustic properties of materials in a room by adjusting the materials so that a virtual sound simulation in the environment matches the actual acoustic impulse response from the room. Similarly [Tang et al., 2020] use an actual 3D model of a room generated via an app. Using this 3D model and an audio recording in the room, they estimate reverberation time and equalization of the room from audio and compute material characteristics for audio rendering in the room. [Yang et al., 2020] learn audio-visual correspondence by classifying if the video's left-right audio channels have been flipped and use this as a pretext task for other audio-visual downstream tasks. For this objective they also collect the YouTube-ASMR-300K dataset of ASMR videos from YouTube with spatial audio.

Binaural audio has also been recently used to achieve different objectives and reason about the 3D environment. [Chen et al., 2020b] introduce the SoundSpaces audio platform to perform audio-visual navigation in scanned 3D environments,

using binaural audio to guide policy learning. Ongoing work continues to explore audio-visual navigation models for embodied agents [Gan et al., 2020c, Chen et al., 2020c, Chen et al., 2020a]. [Christensen et al., 2020] predict depth maps using spatial audio. They emit short chirps from a speaker and record the echoes using a dummy ear stereo microphone setup which is then used to infer depth maps. [Gao et al., 2020a] perform representation learning of visual information via interaction using echoes recorded in indoor 3D simulated environments. They demonstrate the utility of the features learnt in this fashion, and, unlike prior work, do so in the absence of audio input and for a number of downstream tasks.

In contrast to any of these works, we are interested in a different problem of generating accurate spatial binaural sound from videos. We do not use it for navigation or to explicitly estimate information about the environment. Rather, the output of our model is spatial sound to provide a human listener with an immersive audio-visual experience.

Chapter 3

Approach

Our objective is to generate binaural audio from videos with mono audio. In this section, we first describe the overall model and task (Section 3.1), and then we present our networks along with the learning framework for the proposed multi-task setting (Section 3.2).

3.1 Model formulation

Our approach has three main components: the *backbone* for converting mono audio to binaural by injecting the visual information, the *geometric consistency* module which ensures that we maintain the geometric information consistency for the features learnt, and an *IR prediction* module that predicts the room impulse response directly from the frames.

Primary task Our primary objective is to map a given mono sound and video to spatial audio. The reason for the spatial effect of sound is the cues in the two-channel binaural audio: the difference in time when the sound is heard in the left and right ear (Interaural Time Difference) and the difference in levels of the sound in the left and right ear (Interaural Level Difference) [Rayleigh, 1875]. These differences are interpreted to estimate the location of the sound. Therefore when we have a single-channel audio a_M^t, it does not have any spatial characteristics. Binaural audio has two channels $\{a_L^t, a_R^t\}$ to convey the audio to the left and right ear separately and provides spatial effects to the listener.

While the binaural audio provides supervision to recover the two channels from a_M^t, a_M^t alone contains inadequate information to infer the spatialization. Hence, we condition the mono audio on the visuals of the video. More specifically, we transfer the audios into the time-frequency domain using the Short-Time Fourier Transformation (STFT). We aim to predict the binaural audio spectrogram $\{\mathcal{A}_L^t, \mathcal{A}_R^t\}$ from the input mono spectrogram $\mathcal{A}_M^t$, where $\mathcal{A}_X^t = \text{STFT}(a_X^t)$. We extract visual

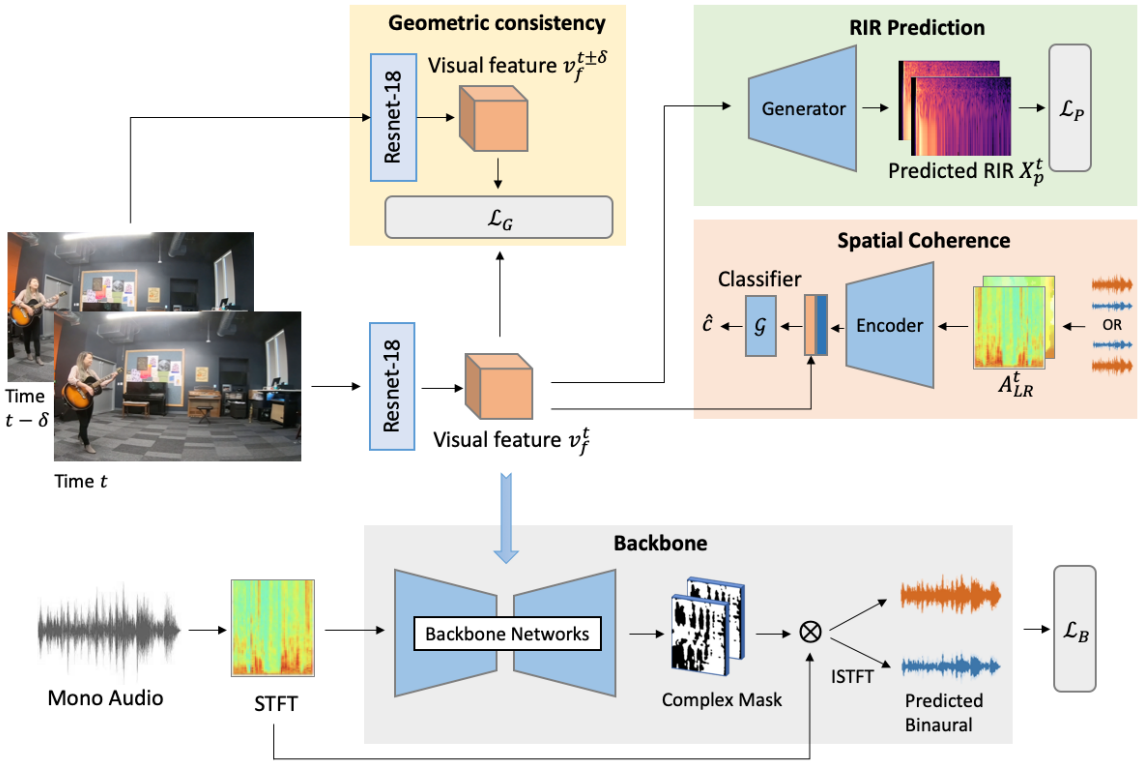

Figure 3.1: Overall network: The overall network takes the visual frames and mono audio as input. The Resnet-18 visual features v_f^t are trained in a multi-task setting. The features v_f^t are used to directly predict the IR via a generator (top right). The binaural audio, which might have flipped channels, is used to get audio features, which combined with v_f^t, are used to train a spatial coherence classifier $\mathcal{G}$ (middle right). Two temporally adjacent frames are also used to ensure geometric consistency (top center). The features v_f^t are jointly trained with the backbone network to predict the final binaural audio.

features v_f^t from the video frames to act as the additional cues we need for the task.

Consistency module Having defined the basic task, we next overview the proposed consistency module, which is comprised of two parts: geometric consistency and spatial coherence.

Geometric consistency: Since the videos are continuous samples over time rather than individual frames, we want that the visual frames have spatio-temporal geometric consistency as a regularizer. Since the position of the source of sound and the position of the camera—as well as the physical environment where the video is recorded—do not typically change instantaneously in videos, there is a

natural coherence between the sound in a video observed at two points that are temporally close. Since visual features are used to condition our binaural prediction, we encourage our visual features to learn a latent representation that is coherent across short intervals of time. The visual features v_f^t and $v_f^{t\pm\delta}$ should be relatively similar to each other to produce audio with fairly similar spatial effects. This enforces temporal consistency on the visuals so that they have similar geometric effects.

Spatial coherence: Since we are predicting binaural audio, we want to ensure that the predicted audio understands which channel is left and right with respect to the visual information. This is crucial to achieve the proper spatial effect while watching videos, as the audio needs to match the seen visuals. We incorporate this in our model by using a classifier to identify if the visuals are aligned with the audio or if the audio does not match. We create misaligned audio by flipping the two channels in the ground truth audio, so the features learn to find the cues in the visual frames which dictate where we hear the sound from and identify whether it is flipped according to the frames.

IR prediction module The third and final component of our multi-task model trains the visual features to be predictive of the room impulse response (IR). An impulse response gives a compact description of the initial direct sound, the early reflections from the surfaces of the room, and a reverberant tail from the subsequent higher order reflections between the source and receiver. These reverberations can be characterised by metrics like the reverberation time RT_{60}. This is the time it takes the energy of the impulse to decay 60dB and can be calculated from the energy decay curve of the IR [Schroeder, 1965]. Since we want our audio-visual feature to be a latent representation of the geometry of the room and the source-receiver position pair, we introduce an auxiliary task to predict the room impulse response (IR) directly from the visual frames. Given a video feature v_f^t, we use a generator to directly estimate the corresponding IR spectrogram $\mathcal{X}_p^t$ and compare it to the ground-truth spectrogram $\mathcal{X}_{gt}^t$. In addition to directly predicting the full IR, we also calculate the RT_{60} metrics of the predicted wave to help learn the impulse responses better.

3.2 Networks and Learning Framework

Next we define the neural networks and specific loss functions we use to train our approach for the objectives described above.

Backbone network The backbone network is used for the baseline task of converting mono to binaural audio and is based upon the networks used for 2.5D visual sound [Gao and Grauman, 2019a]. The audio network consists of a U-Net [Ronneberger et al., 2015] type architecture. It comprises an encoder and a decoder connected via skip connections. The input audio to the network is the STFT spectrogram of the mono audio A_M^t. During training, the mono audio is obtained by taking the mean of the two channels of the ground truth binaural audio $a_m^t = (a_L^t + a_R^t)/2$. The visual network consists of a Resnet-18 [He et al., 2016] to extract visual features v_f^t. These are reduced in dimension, and then tiled and concatenated with the output of the audio encoder to fuse the information from the audio and visual streams. This network does not directly predict the two channels, but instead predicts the *difference* of the two channels. This helps it reason better about the distinction of the channels and not collapse to the easy case of predicting the same output for both channels. Since it is hard to predict the STFT directly, the output of the backbone network is a complex mask M_D^t, which is then multiplied by the original audio spectrogram $\mathcal{A}_M^t$ to get the predicted difference spectrogram:

$$\mathcal{A}_{D(pred)}^t = M_D^t \cdot \mathcal{A}_M^t.$$

The spectrogram difference of the input $\mathcal{A}_D^t$ is computed as the STFT of $a_L^t - a_R^t$. We minimize the distance between these two spectrogram denoted as

$$\|\mathcal{A}_D^t - \mathcal{A}_{D(pred)}^t\|_2^2.$$

In parallel, we also predict the left and right channels directly via an APNet network [Zhou et al., 2020]. It consists of a decoder that predicts two complex masks M_L^t and M_R^t, one for each channel. This decoder fuses the visual features v_f^t at each layer of the decoder and the two masks are used to obtain the predicted channel spectrograms $\mathcal{A}_{L(pred)}^t$ and $\mathcal{A}_{R(pred)}^t$ like above. We again minimize the L2

distance to each channel denoted as

$$\|\mathcal{A}_L^t - \mathcal{A}_{L(pred)}^t\|_2^2 + \|\mathcal{A}_R^t - \mathcal{A}_{R(pred)}^t\|_2^2.$$

This gives us the overall backbone loss:

$$\mathcal{L}_B = \|\mathcal{A}_D^t - \mathcal{A}_{D(pred)}^t\|_2^2 + \left\{\|\mathcal{A}_L^t - \mathcal{A}_{L(pred)}^t\|_2^2 + \|\mathcal{A}_R^t - \mathcal{A}_{R(pred)}^t\|_2^2.\right\} \tag{3.1}$$

For generating audio during test time, the spectrogram $\mathcal{A}_{D(pred)}^t$ is then used to obtain the predicted difference signal $a_{D(pred)}^t$ via an inverse Short-Time Fourier Transformation (ISTFT) operation. The two-channel audio $\{a_L^t, a_R^t\}$ is recovered using input mono audio a_m^t as $a_L^t = (a_m^t + a_{D(pred)}^t)/2$ and $a_R^t = (a_m^t - a_{D(pred)}^t)/2$.

Consistency module For the spatial consistency loss, to ensure that the visual features v_f^t and $v_f^{t\pm\delta}$ are relatively similar, we use an L2 loss directly on the video features with a margin to allow some leeway:

$$\mathcal{L}_S = \max(\|v_f^t - v_f^{t\pm\delta}\| - \alpha, 0), \tag{3.2}$$

where α is the margin we want to allow between two visual features. This ensures that similar visuals should be represented with similar features and the margin allows room for dissimilarity for the changes due to time. This serves to act as a regularizer since visual features guide the learning of the final audio. If the underlying visual features are similar, the predicted audio is conditioned by coherent visuals and therefore more spatially consistent.

We further encourage the visual features to have geometric understanding of the relative positions of the sound source and receiver. To achieve this, we predict if the visual frame and the audio features are aligned or not. We build a classifier $\mathcal{G}$ which takes the binaural audio as input $\mathcal{A}_{LR} = \{\mathcal{A}_L^t, \mathcal{A}_R^t\}$ and combines it with the visual features v_f^t to classify if the audio heard matches the visuals seen. To misalign the visuals and the audio during training, we simply flip the two audio channels with 50% probability to get $\mathcal{A}_{RL} = \{\mathcal{A}_R^t, \mathcal{A}_L^t\}$. To recognise if the audio is flipped, the visual features are forced to reason about information of the relative positions of the sound sources. For the audio we use an encoder similar to the backbone network, and after combining the features, reduce the feature dimensions followed

by a fully connected layer to predict an indicator variable $\hat{c}$ which denotes if the audio is flipped or not. We then calculate the binary cross entropy (BCE) loss on the classifier's prediction of whether the audio is flipped or not $c = \mathcal{G}(\mathcal{A}_{LR}, v_f^t)$ and the actual indicator $\hat{c}$, yielding the geometric consistency loss:

$$\mathcal{L}_G = \text{BCE}(\mathcal{G}(\mathcal{A}_{LR}, v_f^t),\ \hat{c}). \tag{3.3}$$

IR prediction module An impulse response is a binaural signal that captures the acoustic response of an environment to an audio stimulus. We predict it directly from the visual frames. We convert the impulse response $\{r_L, r_R\}$ to the frequency domain using the STFT and obtain magnitude spectrograms $\mathcal{X}_{gt}$ for each channel. The IR prediction network consists of a decoder which takes the computed features v_f^t from the visual frame as input and performs upconvolutions to obtain a predicted magnitude spectrogram $\mathcal{X}_{(pred)}^t$. We want to minimize the difference between the predicted IR $\mathcal{X}_{(pred)}^t$ and the ground truth $\mathcal{X}_{gt}^t$. Therefore we minimize the L2 loss between the two spectrograms. In addition, using the predicted spectrogram we obtain the original IR waveform using the Griffin-Lim algorithm [Griffin and Lim, 1984, Perraudin et al., 2013] and compute the $RT_{60(pred)}$ from $\mathcal{X}_{(pred)}^t$. We compute the L1 distance between the RT_{60} of the predicted IR $RT_{60(pred)}$, and the ground truth IR $RT_{60(gt)}$. Thus the overall IR prediction loss is:

$$\mathcal{L}_P = \|\mathcal{X}_{(pred)}^t - \mathcal{X}_{gt}^t\|_2^2 + |RT_{60(pred)} - RT_{60(gt)}|. \tag{3.4}$$

Overall objective: Therefore the overall multi-task loss is a combination of these losses:

$$\mathcal{L} = \lambda_B \mathcal{L}_B + \lambda_S \mathcal{L}_S + \lambda_G \mathcal{L}_G + \lambda_P \mathcal{L}_P \tag{3.5}$$

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

Chapter 4

Experimental Setup

In this section, we first describe the datasets we use and give details of the dataset generated (Section 4.1) and then we discuss the implementation details for the training and testing (Section 4.2).

4.1 Datasets

FAIR-Play: This dataset was collected by [Gao and Grauman, 2019a] for the purpose of this task. It consists of video recordings of people playing various instruments in a music room captured with a binaural microphone rig. The dataset consists of 1,871 10-second clips, and the split into train/val/test is provided by the authors. We follow the same protocols provided for our method as well.

SimBinaural: Binaural audio and impulse response collection in the real world is expensive and challenging, requiring special equipment. IR collection is also limited by the fact that it differs for each physical space and source-receiver position pair within that space. Thus even if an IR is captured, the variety in the data is severely limited to a few positions and rooms. To facilitate and understand the relation between the visuals and the audio for learning better geometric features, we create a dataset called SimBinaural of simulated videos in photo-realistic 3D indoor scene environments. The generated videos resemble real-world audio recordings sampled from 1,020 distinct rooms in 80 distinct environments (each environment is a multi-room home).

Using the SoundSpaces audio simulations [Chen et al., 2020b] together with the Habitat simulator [Savva et al., 2019], we created realistic videos with binaural sounds for publicly available 3D environments in Matterport3D [Chang et al., 2017]. Habitat is an open-source 3D simulator that allows fast rendering for multiple datasets including Matterport3D and Replica [Straub et al., 2019]. For our data, we use 80 environments comprised of diverse indoor environments including real-

Figure 4.1: The first row displays example frames from videos in FAIR-Play [Gao and Grauman, 2019a] while the second row shows examples from the newly introduced SimBinaural dataset.

Dataset	#Videos	Length (hrs)	#Rooms	IR
FAIR-Play [Gao and Grauman, 2019a]	1,871	5.2	1	No
SimBinaural	107,280	903.7	1,020	Yes

Table 4.1: A comparison of the data in FAIR-Play and the large scale data we generated.

world homes with 3D meshes and image scans. SoundSpaces [Chen et al., 2020b] is a dataset with precomputed room impulse responses obtained with geometrical acoustic simulations for the two 3D datasets (Matterport3D and Replica). They augment Habitat to allow insertion of arbitrary sound sources in an array of real-world scanned environments by providing impulse responses for source and receiver position pairs. These impulse responses are provided for position pairs that are densely sampled from the environment.

We want the videos we generate to have audio emitting from plausible sound-making objects visible in the video. Since the simulator does not have objects that emit sound, we explicitly insert different 3D models of various instruments like guitar, violin, flute etc. and other sound-making objects into the scene. Each kind of object has multiple different models of that class for diversity, so it does not associate a sound with a particular 3D model only. We placed more than 30 objects from 13 classes roughly evenly.

To generate realistic binaural sound in the environment, we use the SoundSpaces [Chen et al., 2020b] room impulse responses, which are a function of room geometry, materials, and the sound source location. We choose a sound source location to

place a 3D object, and a receiver (camera+microphone) location where the sound is heard. Using the IR for the appropriate position at which we place the object and the receiver, we convolve it with an audio waveform that is plausible from the source location (e.g., a guitar playing for an inserted guitar 3D object). This results in binaural audio for the receiver as if it is coming from the source object. We use sounds recorded in anechoic environments so that there is no existing reverberations to affect the data. The sounds are obtained from a copyright-free internet source [fre, 2021] and OpenAIR data [Murphy and Shelley, 2010] to form a set of 127 different sound clips spanning the 13 distinct object categories.

Finally we want to capture videos with visuals and audio just like in the real world. We place an agent equipped with a camera and binaural microphone at the receiver location from above. The source and receiver locations are chosen so that they are in the same room. Additionally, we use ray tracing to ensure that the object is in view of the agent and turn the camera towards the object. The source positions are densely sampled from the environment to have all possible source positions in all the 3D environments.

Using this setting for generating audio-visuals, we create videos by moving the agent, and therefore the camera and microphones, around the room in different trajectories for each video. For a particular video and trajectory, we use a fixed source position and the agent traverses a random path. It is important to note that while the source is fixed, the trajectories and camera orientations are chosen such that they source is always in view, and there are no obstacles. This also means that the view of the object is not the same throughout the video; it changes as the camera moves and rotates, so we get diverse orientations of the object and positions within a video frame, for each video. For each position that the camera moves to, we compute the audio heard at that location using the IR.

The agent remains at one position for 5 seconds before moving to the next location smoothly. We generate the videos at 5 frames per second and the length of the trajectories vary from 2 to 40 across different videos. At each position in the trajectory, there is a small translational motion of the camera as well. The average length of the videos in the dataset is 30.3s and the median length is 20s. We generated over 100K videos, of which a subset (about 20%) is used for training and testing.

4.2 Implementation details

All the networks for training were written in PyTorch [Paszke et al., 2019]. For preprocessing both datasets, we followed the standard preprocessing steps from [Gao and Grauman, 2019a]. We resampled all the audio to 16kHz and computed the STFT using a FFT size of 512, window size of 400, and hop length of 160. For training the backbone, we use 0.63s clips of the 10s audio and use the corresponding frame. The frames are extracted from the videos at 10fps. The visual frames are randomly cropped to 448 $\times$ 224. For testing, we use a sliding window of 0.1s to compute the binaural audio for all methods.

The backbone networks are same as [Gao and Grauman, 2019a] and [Zhou et al., 2020]. The U-Net consists of 5 convolution layers for downsampling and 5 upconvolution layers in the upsampling network. They include skip connections for the layers with the same feature size. The APNet consists of 3 upsampling layers which combine the visual feature v_f^t with the upconvolution layers of the U-Net. The visual network is a ResNet-18 [He et al., 2016] with the pooling and fully connected layers removed. The encoder for spatial coherence follows the same architecture as the encoder of the U-Net for the audio feature extraction. The classifier combines the audio and visual features and has a fully connected layer to predict the outcome. The generator network is adapted from GANSynth [Engel et al., 2019], modified to fit the required dimensions of the audio spectrogram.

For training, all baselines are evaluated with the same parameters for fairness. We use the Adam optimizer [Kingma and Ba, 2015] and a batch size of 64. The initial learning rates are 0.001 for the audio and fusion networks, and 0.0001 for all other networks. We trained the FAIR-Play dataset for 1000 epochs and SimBinaural for 100 epochs. The δ for choice of frame is set to 1s and the λ's used are set based on validation set performance to $\lambda_B = 10, \lambda_S = 1, \lambda_G = 0.01, \lambda_P = 1$.

Chapter 5

Results

In this section we present the results of our proposed method. We evaluate the performance using two standard metrics as used by [Gao and Grauman, 2019a, Morgado et al., 2018, Zhou et al., 2020].

STFT Distance: This metric is the Euclidean distance of the predicted STFT spectrograms of the left and the right channel to the ground truth spectrograms.

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

This directly measures how good a spectrogram we produce, which is the objective we are training on.

Envelope Distance: While the previous metric measures the similarity of the spectrograms, we would like to measure similarity of raw audio. However, comparing raw audio waveforms directly may not be informative about the actual perceptual similarity of the audio. Hence, following prior work, we measure the Euclidean distance between the envelopes of the predict signal and the ground truth for each channel by calculating envelopes of the audio signal. Let the envelope of a signal a_L^t be E_L^t, then the metric is given by

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

5.1 Evaluation results

In this section, we first present a case study on the IR prediction module (Section 5.1.1). We then describe the baselines we compare to (Section 5.1.2), followed by result of experimental evaluations on the SimBinaural dataset (Section 5.1.3) and the FAIR-Play dataset (Section 5.1.4).

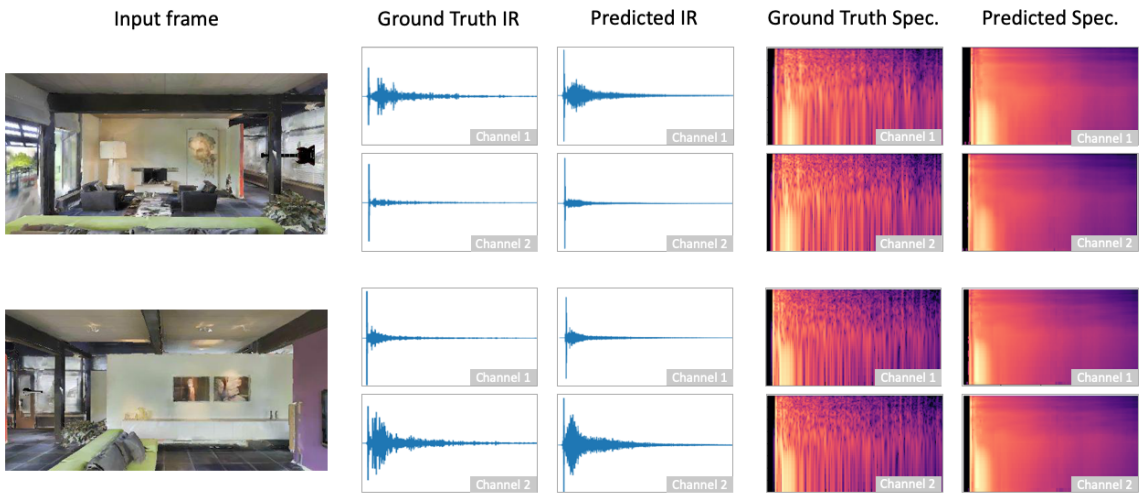

Figure 5.1: IR Prediction: The first column is the input frame to the encoder. The second column depicts the ground truth IR for the frame and the fourth column is the corresponding spectrogram of this IR. The third and fifth columns show the predicted IR waveform and spectrogram, respectively. This predicted IR waveform is estimated from the spectrogram generated by our network.

5.1.1 IR Prediction Case Study

First we perform a case study on the task of predicting the binaural IR directly from a single visual frame. We train separately on this task to see the feasibility of learning the IR directly from the visual frame alone. The visual frames have some information about the acoustic environment which we predict by looking at one snapshot of the scene. We predict the magnitude spectrogram of the IR for the two channels as we work on the binaural task. We also obtain the actual waveform of the IR. Since the predicted spectrogram $\mathcal{X}^t_{(pred)}$ is a magnitude spectrogram instead of a complex one, we cannot directly do ISTFT. Therefore, we use the Griffin-Lim algorithm [Griffin and Lim, 1984] to generated the predicted IR. We present qualitative examples of predictions from the test set in Figure 5.1. We can see that we can get a fairly accurate general idea of the IR, and the difference between the response in each channel is also captured well.

To evaluate if we capture the materials and geometry effectively, we also train another task to predict the reverberation time RT_{60} of the IR from the visual frame. A more accurate prediction of RT_{60} means that our network understands how the

wave will interact with the room and materials and whether it takes more or less time to decay. We formulate this as a classification task. We discretize the range of the RT_{60} number into 10 bins or classes such that they each have roughly the same number of samples based on the training set. We then use a classifier to predict this range class of RT_{60} using only the visual frame as input. The classifier consists of a ResNet-18 with 10 classes as the last layer and takes the video frame as input. The classifier has a test accuracy of **61.5%** which demonstrates the networks ability to estimate the RT_{60} range quite well (a random classifier has 10% accuracy).

5.1.2 Baselines

We compare our method to the following baselines to demonstrate that our model learns desirable information.

Flipped-Visual: In this baseline, we flip the visual frame's pixels from left to right while testing to evaluate if the features actually learn the spatial geometric information. Since this baseline uses incorrect visual information, it ought to be at a disadvantage if the visual frame is significant for our results.

Audio Only: In this setting, we do not provide the visual frames to the network in order to to verify if the visual information is essential to learning. Therefore we only evaluate the performance of the backbone network with mono audio as input while other configurations are the same.

Mono-Mono: In this baseline, both channels have the same input mono audio repeated as the two-channel output. It helps verify if we are actually distinguishing between the channels.

Mono2Binaural [Gao and Grauman, 2019a]: A state-of-the-art model for this task.

APNet [Zhou et al., 2020]: A state-of-the-art model that handles binauralization and audio source separation. We use the APNet network from their method and train only on binaural data for stereo audio.

For the existing methods, we reproduce the results carefully using the code provided by the authors. For our model, we perform multi-task training using the auxiliary losses simultaneously with the baseline network.

	Scene-Split		Position-Split	
	STFT	ENV	STFT	ENV
Mono-Mono	1.334	0.159	1.315	0.161
Audio-Only	0.872	0.127	0.857	0.127
Flipped-Visual	1.082	0.142	1.075	0.141
Mono2Binaural[Gao and Grauman, 2019a]	0.824	0.123	0.803	0.123
APNet [Zhou et al., 2020]	0.816	0.122	0.777	0.122
Backbone+IR Pred	0.803	0.122	0.745	0.120
Backbone+Spatial	0.807	0.121	0.750	0.119
Backbone+Geom	0.812	0.122	0.744	0.118
Full Model	**0.777**	**0.118**	**0.719**	**0.117**

Table 5.1: Quantitative results of binaural audio prediction on SimBinaural. For both the metrics, lower is better.

5.1.3 Experiments on SimBinaural

We evaluate our model on two different splits of our data: ***Scene-Split*** and ***Position-Split***. In the ***Scene-Split***, the train/test/val splits do not have any overlapping scenes from the Matterport3D [Chang et al., 2017] dataset and hence the model has never seen the room of the video. In this respect, it is therefore a harder task compared to FAIR-Play, which is recorded in one room. The ***Position-Split*** may have the same Matterport3D scene in the split, with each scene consisting of several rooms, but the exact configuration of the source object and receiver position is not seen before. Therefore, during training, the model might have seen the room from a different perspective with the sound-making object in some other position, but it still is quite different from the test data.

Table 5.1 shows the results of our experiments on SimBinaural. The backbone model is also trained only using Equation 3.1 without any of our additional losses. We provide an ablation and demonstrate the effect of each loss on the model individually and then present our full model that uses all the proposed losses. For both splits, our model outperforms all the baselines, including the two state-of-the-art prior methods.

First, we can see from Table 5.1 that the ***Scene-Split*** is a fundamentally harder task to solve. This is because we must predict the sound, as well as other characteristics like the IR, from visuals that we have not observed before. This forces the model to try to generalize its encoding to generic visual properties (wall orienta-

	STFT	ENV
Mono-Mono	1.215	0.157
Audio-Only	1.102	0.145
Flipped-Visual	1.134	0.152
Mono2Binaural[Gao and Grauman, 2019a]	0.947	0.142
APNet [Zhou et al., 2020]	0.904	0.138
Backbone+Spatial	0.873	**0.134**
Backbone+Geom	0.874	0.135
Full Model	**0.869**	**0.134**

Table 5.2: Quantitative results of binaural audio prediction on the FAIR-Play dataset. For both the metrics, lower is better.

tions, major furniture, etc.) that have intra-class variations and geometry changes compared to the training scenes.

Second, the ablations shed light on the impact of each of the proposed losses in our multi-task framework. The Backbone+IR Pred model, trained by adding Equation 3.4 which uses the IR, provides a significant improvement over the baseline across both splits. This indicates that with the availability of the IR during training, the visual features can learn to extract information relevant for more accurate prediction. Note that the true IR is never provided at test time, consistent with real-world applications where this would not be measurable directly by the system (but, as our results show, could be predicted). The Backbone+Spatial denotes adding only the loss from Equation 3.2, while Backbone+Geom denotes adding only the loss from Equation 3.3 to the backbone network. Both the losses individually perform better than the state-of-the-art indicating that there is some geometry inferred as desired. The full model uses all the losses as in Equation 3.5. This outperforms other methods significantly on both splits. It also outperforms using each of the losses individually, which demonstrates the losses can combine to jointly learn better visual features for generating spatial audio.

5.1.4 Experiments on FAIR-Play

Table 5.2 shows the results of our experiments on the FAIR-Play real video dataset.

For our method, since we do not have the ground truth impulse responses for

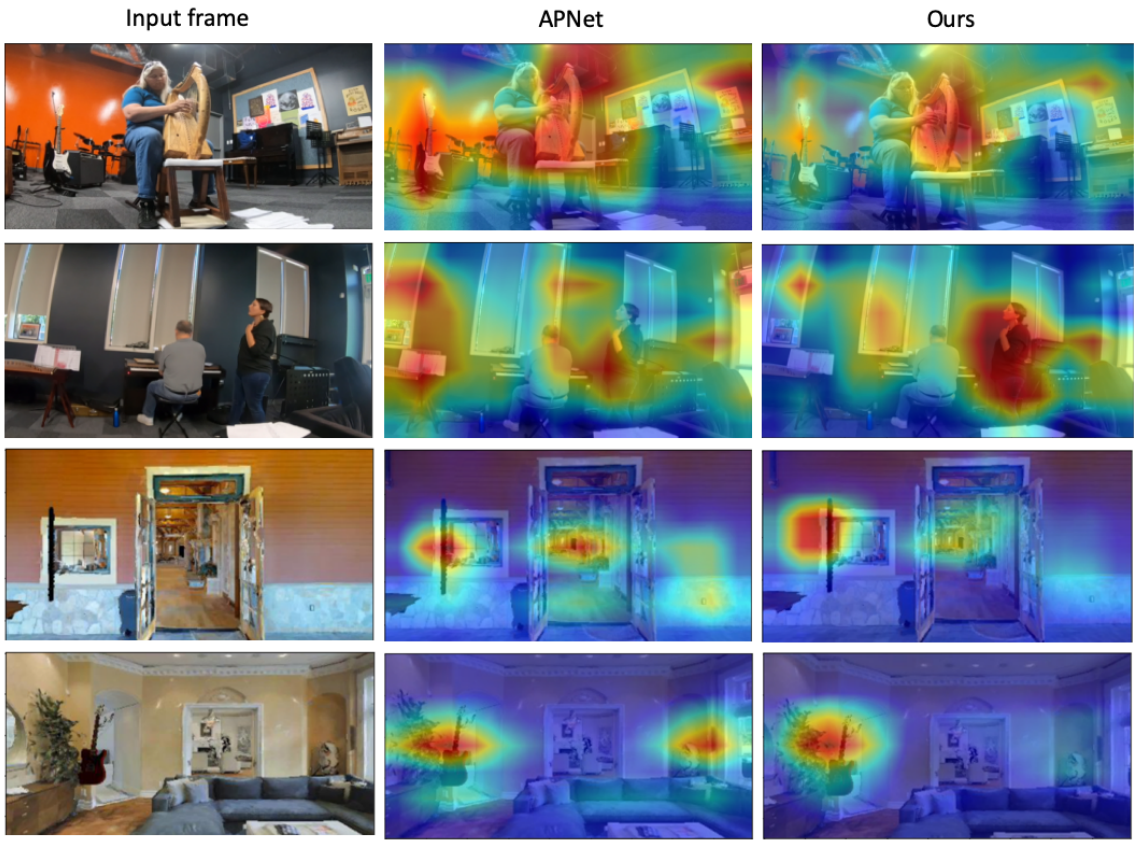

Figure 5.2: Qualitative visualization of the activation maps of the visual features from the APNet [Zhou et al., 2020] baseline and our model. While their method produces more diffuse maps, our method localizes the object better within the image. This indicates that the visual features in our method is better at identifying the regions which might be emitting sound to generate more accurate binaural audio.

the FAIR-Play dataset, we omit the IR prediction network. The Backbone+Spatial denotes adding only the loss from Equation 3.2, while Backbone+Geom denotes adding only the loss from Equation 3.3 to the backbone network. We can observe that both variants of our method outperforms the state-of-the-art. Therefore, enforcing the geometric and spatial constraints is beneficial to the binaural audio generation task. We get the best results when we combine both the losses highlighting that both contribute together to improve the predictions.

Figure 5.2 visualizes the activation maps of the visual features when the visual frame is passed through the ResNet. We can see that while the activation maps

for APNet [Zhou et al., 2020] are diffused and focusing on non essential parts like objects in the background, our method focuses more on the object/region producing the sound and its location. This shows that we can learn better visual features which can more accurately capture the important aspects required by the backbone networks for binaural audio generation.

Chapter 6

Conclusion

We presented a multi-task setting to learn geometry-aware visual features for mono to binaural audio conversion in videos. Our method exploits the inherent room and object geometry and spatial information encoded in the visual frames to generate rich binaural audio. We also generated a large-scale video dataset with binaural audio in photo-realistic environments to better understand and learn the relation between visuals and binaural audio. This dataset will be made publicly available to support further research in this direction. Our state-of-the-art results on two datasets demonstrate the efficacy of our proposed formulation.

The simulated data allows for a numerous possibilities for learning different tasks and then applying them for other real world challenges. In the future, we plan to better harness the vast amounts of data for learning and generalizing to other videos. We can also explore improving the binaural sound generation by better using the IR and transferring the IR knowledge to infer characteristics in real videos.

Bibliography

[fre, 2021] (2021). Freesound. https://freesound.org/. [Online; accessed 1-April-2021].

[Afouras et al., 2018] Afouras, T., Chung, J. S., and Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. In *Interspeech*.

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

[Arandjelovic and Zisserman, 2017] Arandjelovic, R. and Zisserman, A. (2017). Look, listen and learn. In *ICCV*.

[Arandjelović and Zisserman, 2018] Arandjelović, R. and Zisserman, A. (2018). Objects that sound. In *ECCV*.

[Aytar et al., 2016] Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*.

[Chang et al., 2017] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017). Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

[Chen et al., 2020a] Chen, C., Al-Halah, Z., and Grauman, K. (2020a). Semantic audio-visual navigation. *arXiv preprint arXiv:2012.11583*.

[Chen et al., 2020b] Chen, C., Jain, U., Schissler, C., Gari, S. V. A., Al-Halah, Z., Ithapu, V. K., Robinson, P., and Grauman, K. (2020b). Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer.

[Chen et al., 2020c] Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S. K., and Grauman, K. (2020c). Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*, 1(2):6.

[Chen et al., 2020d] Chen, P., Zhang, Y., Tan, M., Xiao, H., Huang, D., and Gan, C. (2020d). Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29:8292–8302.

[Christensen et al., 2020] Christensen, J. H., Hornauer, S., and Stella, X. Y. (2020). Batvision: Learning to see 3d spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1581–1587. IEEE.

[Chung et al., 2017] Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE.

[Chung et al., 2020] Chung, S.-W., Choe, S., Chung, J. S., and Kang, H.-G. (2020). Facefilter: Audio-visual speech separation using still images. In *INTERSPEECH*.

[Engel et al., 2019] Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*.

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

[Gabbay et al., 2018] Gabbay, A., Shamir, A., and Peleg, S. (2018). Visual speech enhancement. In *INTERSPEECH*.

[Gan et al., 2020a] Gan, C., Huang, D., Chen, P., Tenenbaum, J. B., and Torralba, A. (2020a). Foley music: Learning to generate music from videos. *arXiv preprint arXiv:2007.10984*, 4(6):7.

[Gan et al., 2020b] Gan, C., Huang, D., Zhao, H., Tenenbaum, J. B., and Torralba, A. (2020b). Music gesture for visual sound separation. In *CVPR*.

[Gan et al., 2020c] Gan, C., Zhang, Y., Wu, J., Gong, B., and Tenenbaum, J. B. (2020c). Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE.

[Gao et al., 2020a] Gao, R., Chen, C., Al-Halab, Z., Schissler, C., and Grauman, K. (2020a). Visualechoes: Spatial image representation learning through echolocation. In *ECCV*.

[Gao et al., 2018] Gao, R., Feris, R., and Grauman, K. (2018). Learning to separate object sounds by watching unlabeled video. In *ECCV*.

[Gao and Grauman, 2019a] Gao, R. and Grauman, K. (2019a). 2.5d visual sound. In *CVPR*.

[Gao and Grauman, 2019b] Gao, R. and Grauman, K. (2019b). Co-separating sounds of visual objects. In *ICCV*.

[Gao et al., 2020b] Gao, R., Oh, T.-H., Grauman, K., and Torresani, L. (2020b). Listen to look: Action recognition by previewing audio. In *CVPR*.

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

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

[Hershey and Movellan, 2000] Hershey, J. R. and Movellan, J. R. (2000). Audio vision: Using audio-visual synchrony to locate sounds. In *NeurIPS*.

[Hu et al., 2016] Hu, D., Li, X., et al. (2016). Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582.

[Hu et al., 2020] Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., and Dou, D. (2020). Discriminative sounding objects localization via self-supervised audiovisual matching. In *NeurIPS*.

[Kazakos et al., 2019] Kazakos, E., Nagrani, A., Zisserman, A., and Damen, D. (2019). Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*.

[Kidron et al., 2005] Kidron, E., Schechner, Y. Y., and Elad, M. (2005). Pixels that sound. In *CVPR*.

[Kingma and Ba, 2015] Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

[Korbar et al., 2018] Korbar, B., Tran, D., and Torresani, L. (2018). Co-training of audio and video representations from self-supervised temporal synchronization. In *NIPS*.

[Lee et al., 2021] Lee, J., Jain, M., Park, H., and Yun, S. (2021). Cross-attentional audio-visual fusion for weakly-supervised action localization.

[Long et al., 2018] Long, X., Gan, C., Melo, G., Liu, X., Li, Y., Li, F., and Wen, S. (2018). Multimodal keyless attention fusion for video classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[Lu et al., 2019] Lu, Y.-D., Lee, H.-Y., Tseng, H.-Y., and Yang, M.-H. (2019). Self-supervised audio spatialization with correspondence classifier. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3347–3351. IEEE.

[Morgado et al., 2020] Morgado, P., Li, Y., and Vasconcelos, N. (2020). Learning representations from audio-visual spatial alignment. *arXiv preprint arXiv:2011.01819*.

[Morgado et al., 2018] Morgado, P., Vasconcelos, N., Langlois, T., and Wang, O. (2018). Self-supervised generation of spatial audio for 360° video. *arXiv preprint arXiv:1809.02587*.

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

[Nagrani et al., 2020] Nagrani, A., Sun, C., Ross, D., Sukthankar, R., Schmid, C., and Zisserman, A. (2020). Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10326.

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

[Owens et al., 2016a] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. (2016a). Visually indicated sounds. In *CVPR*.

[Owens et al., 2016b] Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., and Torralba, A. (2016b). Ambient sound provides supervision for visual learning. In *ECCV*.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[Perraudin et al., 2013] Perraudin, N., Balazs, P., and Søndergaard, P. L. (2013). A fast griffin-lim algorithm. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE.

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

[Richard et al., 2021] Richard, A., Markovic, D., Gebru, I. D., Krenn, S., Butler, G., de la Torre, F., and Sheikh, Y. (2021). Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*.

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

[Rouditchenko et al., 2019] Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., and Torralba, A. (2019). Self-supervised audio-visual co-segmentation. In *ICASSP*.

[Savva et al., 2019] Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al. (2019). Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347.

[Schissler et al., 2017] Schissler, C., Loftin, C., and Manocha, D. (2017). Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE transactions on visualization and computer graphics*, 24(3):1246–1259.

[Schroeder, 1965] Schroeder, M. R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(6):1187–1188.

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

[Straub et al., 2019] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., et al. (2019). The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

[Tang et al., 2020] Tang, Z., Bryan, N. J., Li, D., Langlois, T. R., and Manocha, D. (2020). Scene-aware audio rendering via deep acoustic analysis. *IEEE transactions on visualization and computer graphics*, 26(5):1991–2001.

[Tian et al., 2020] Tian, Y., Li, D., and Xu, C. (2020). Unified multisensory perception: weakly-supervised audio-visual video parsing. *arXiv preprint arXiv:2007.10558*.

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

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

[Wu et al., 2019] Wu, Y., Zhu, L., Yan, Y., and Yang, Y. (2019). Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6292–6300.

[Xu et al., 2019] Xu, X., Dai, B., and Lin, D. (2019). Recursive visual sound separation using minus-plus net. In *ICCV*.

[Xu et al., 2021] Xu, X., Zhou, H., Liu, Z., Dai, B., Wang, X., and Lin, D. (2021). Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

[Yang et al., 2020] Yang, K., Russell, B., and Salamon, J. (2020). 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.

[Yu et al., 2020] Yu, J., Zhang, S.-X., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., and Yu, D. (2020). Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE.

[Zhao et al., 2019] Zhao, H., Gan, C., Ma, W.-C., and Torralba, A. (2019). The sound of motions. In *ICCV*.

[Zhao et al., 2018] Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. (2018). The sound of pixels. In *ECCV*.

[Zhou et al., 2019a] Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. (2019a). Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*.

[Zhou et al., 2019b] Zhou, H., Liu, Z., Xu, X., Luo, P., and Wang, X. (2019b). Vision-infused deep audio inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 283–292.

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

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