Binaural Audio-Visual Localization

Xinyi Wu,[1] Zhenyao Wu,[1] Lili Ju,[2,*] Song Wang[1,*]
[1]Department of Computer Science and Engineering, University of South Carolina, USA
[2]Department of Mathematics, University of South Carolina, USA
{xinyiw, zhenyao}@email.sc.edu, ju@math.sc.edu, songwang@cec.sc.edu

Abstract

Localizing sound sources in a visual scene has many important applications and quite a few traditional or learning-based methods have been proposed for this task. Humans have the ability to roughly localize sound sources within or beyond the range of the vision using their binaural system. However most existing methods use monaural audio, instead of binaural audio, as a modality to help the localization. In addition, prior works usually localize sound sources in the form of object-level bounding boxes in images or videos and evaluate the localization accuracy by examining the overlap between the ground-truth and predicted bounding boxes. This is too rough since a real sound source is often only a part of an object. In this paper, we propose a deep learning method for pixel-level sound source localization by leveraging both binaural recordings and the corresponding videos. Specifically, we design a novel Binaural Audio-Visual Network (BAVNet), which concurrently extracts and integrates features from binaural recordings and videos. We also propose a point-annotation strategy to construct pixel-level ground truth for network training and performance evaluation. Experimental results on Fair-Play and YT-Music datasets demonstrate the effectiveness of the proposed method and show that binaural audio can greatly improve the performance of localizing the sound sources, especially when the quality of the visual information is limited.

Introduction

Humans are able to extract a wealth of useful information through eyes and ears and then integrate and interpret them to further understand the surroundings. Imagine the scene that two persons in front of you are singing a song together but actually one of them is a lip-syncher. Can you locate the sound source to identify the real singer with only your eyes? What about with only your ears? It is generally quite difficult to reach such a goal under both situations. On the one hand, it is not hard for a human to pretend to do something to trick others' eyes. On the other hand, although ears seem more difficult to be fooled than eyes, the sound can't give us accurate location except rough direction information. However, by using eyes and ears together we could localize accurately sound sources in many situations.

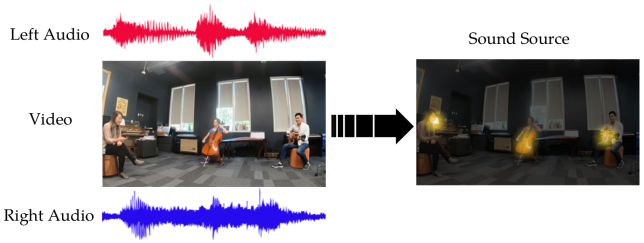

Figure 1: Binaural audio-visual localization task. Given the binaural recordings and the corresponding visual scene, our method localizes the sound sources in pixel level and outputs a localization map. Note that the pixel-level localization only identifies the part of the object that is making the sound.

Audio-visual localization is a well-established task, which aims at localizing sound sources in visual scenes by integrating both visual and audio information. Most of the existing methods (Senocak et al. 2018; Arandjelovic and Zisserman 2018; Owens and Efros 2018; Zhao et al. 2018; Hu, Nie, and Li 2019) integrate only monaural audio, which is a mixture of sounds in a scene, to help localize sound sources. The localization results are promising due to the assistance of sound information. However, our perception of monaural audio often cannot provide location information or even a rough direction without the help of visual information. In contrast, using binaural recordings to localize sound sources is more reasonable and appeals more to common sense. Binaural audio-visual localization would also benefit a few other applications, such as robot navigation and action recognition. Another practically useful application is that it can help police officers to accurately locate the gunman in the shooting, where the binaural audio can be achieved by simply adding two microphones on the two sides of the surveillance cameras, respectively. It is worth mentioning that the importance of binaural recordings should be especially emphasized when the quality of visual information is low or the videos are recorded at night or in foggy days.

In addition, to get a quantitative analysis for the localization results, many prior methods such as Senocak et al. (2018), Hu, Nie, and Li (2019) and Zhao et al. (2018) manually annotate the sound objects using bounding boxes and then calculate the accuracy by examining the overlap, e.g.,

*Co-corresponding authors.

the Intersection over Union (IoU), between the annotated and predicted bounding boxes. However, the sound sources often are much more similar to humans' fixation which is concentrated and small, e.g., a person's throat or the sounding objects they are holding. Other methods like (Gao and Grauman 2019a; Arandjelovic and Zisserman 2018) can provide pixel-level visualization results but no quantitative results due to the lack of ground truth, since the localization results are only by-products of their networks.

In this paper we consider the sound source localization problem in pixel level and the process of our binaural audio-visual localization is illustrated in Fig. 1. We propose a novel Binaural Audio-Visual Network (BAVNet) to address this problem and present concrete and accurate quantitative analysis. The input of our BAVNet consists of frame sequences and the corresponding binaural recordings. We first extract features from video frames, binaural audios, single left audio and single right audio, respectively. Then we fuse the frame features and the three types of audio features together. Considering that the left and right recordings are horizontally symmetrical in the visual space as illustrated in Fig. 2, we learn the mapping function between the visual and auditory modalities. In the way that the left recording is mapped onto the original frame while the right one onto the flipped image. We also use Convolutional LSTM (ConvLSTM) (Shi et al. 2015) to model the dynamic changes in both visual and auditory information. Finally, a decoder CNN is fed with the intermediate results to recover the resolution and produce localization result.

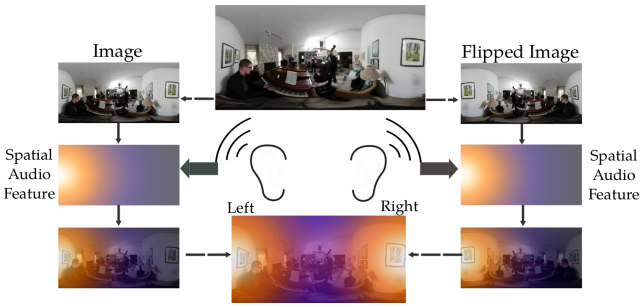

Figure 2: The horizontal symmetry of the left and right recordings in a visual scene. We map the left recording onto the original frame and map the right recording onto the horizontally flipped frame.

The main contributions of this paper are as follows: (1) We redefine the sound source localization problem in pixel level instead of object level, which better reflects the reality; (2) We propose a novel deep learning network of BAVNet to tackle this problem by jointly extracting and integrating the features from the binaural recordings and the corresponding video; (3) We annotate two audio-visual datasets, FairPlay and YT-Music, for providing pixel-level supervision and quantitative analysis; (4) Experimental results demonstrate the effectiveness of the proposed method and show that binaural audio can greatly improve the performance of localizing the sound sources.

Related Work

Audio-visual analysis. As audio has been shown to be an important modal for understanding visual scenes, a bunch of audio-visual tasks has been introduced and addressed in the community including sound classification (Aytar, Vondrick, and Torralba 2016; Arandjelovic and Zisserman 2017), sound source separation (Owens and Efros 2018; Zhao et al. 2018; Gao, Feris, and Grauman 2018; Gao and Grauman 2019a,b; Xu, Dai, and Lin 2019; Zhao et al. 2019), sound source localization (Senocak et al. 2018; Arandjelovic and Zisserman 2018; Owens and Efros 2018; Zhao et al. 2018; Hu, Nie, and Li 2019), audio-visual event localization (Tian et al. 2018; Wu et al. 2019), audio denoising (Gao, Feris, and Grauman 2018; Gao and Grauman 2019b), sound generation (Zhou et al. 2018) and audio inpainting (Zhou et al. 2019).

Monaural audio-visual localization. Prior to the widespread use of deep learning techniques, sound source localization has already been a long-lasting topic which received extensive attention in the literature. Traditional approaches rely on projecting audio-visual data to low-dimension subspace (Fisher III et al. 2001), synchrony (Hershey and Movellan 2000) and motion cues such as trajectory (Barzelay and Schechner 2007) and optical flow (Izadinia, Saleemi, and Shah 2012). With the development of deep learning techniques, Senocak et al. (2018) first proposed the learning-based sound source localization in visual scenes. Specifically, a two-stream structure is designed to extract the audio and visual features independently and then attention mechanisms are applied to the integrated features to localize the sound sources. In addition, in order to construct an unsupervised training setting, audio and corresponding visual features are also constrained to be close to each other in the feature space. A similar technique was developed by Arandjelovic and Zisserman (2018) for localizing objects that sound in the visual scenes. Concurrently, Owens and Efros (2018) and Zhao et al. (2018) proposed some different networks for sound source localization, which can additionally perform the separation of mixed speech messages and musical sound. More recently, Hu, Nie, and Li (2019) integrated K-means into a two-stream network to help distinguish objects or sounds captured by video for both sound localization and separation. Different from this line of researches, the goal of our work is to localize sound sources using binaural recordings, which is more similar to the normal auditory system of humans.

Binaural audio-visual tasks. Until now, few works have focused on localizing sound sources based on binaural audio-visual data in a supervised fashion in the computer vision community. The most relevant work was reported in Gao and Grauman (2019a), which proposed a network to convert the monaural audio to binaural audio using the visual information. The localization results are only by-products while performing binauralization during training, thus only some qualitative results are presented instead of quantita-

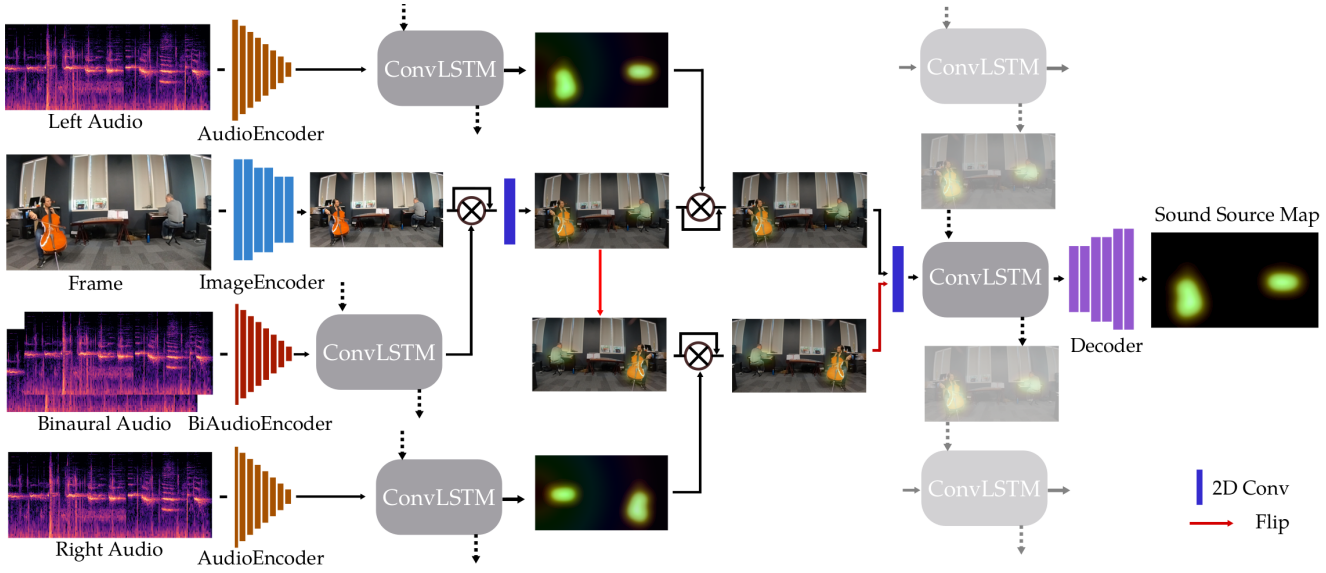

Figure 3: The architecture of the proposed BAVNet. The network is fed with an input consisting of video frames, binaural audio, single left and right audio channels, and generates the sound source localization map in an end-to-end fashion.

tive evaluations. Recently, Gao and Grauman (2019a) proposed a novel teacher-student based network which successfully transfers knowledge learned from visual modality to the stereo audio modality for a vehicle tracking problem. Although that method and our work in this paper both try to leverage the left-right audio for localization (Gao and Grauman (2019a) actually for tracking/detection), they differ in several critical aspects. First, our goal is to localize sound sources which often locates on parts of the whole object, whereas Gao and Grauman (2019a) is to localize the whole object as in Arandjelovic and Zisserman (2018). Next, and equally critically, Gan et al. (2019) takes the meta-data of the camera, including camera height, pitch angle, and orientation between the camera and a street, as input for both student-network training and the inference, thus it is only easy to be applied into a self-recorded dataset. Instead, our method can be applied to any in-the-wild dataset without the meta-data of the camera. BatVision (Christensen, Hornauer, and Yu 2019) is another recent work that takes advantage of "two ears" to generate a disparity-like map to show details about the depth and obstacles in the room.

Our Approach

In this section, we design a novel convolutional neural network to learn localization of sound sources based on integration of both visual and binaural auditory information. Different from most previous methods, which use two-stream networks to process video and audio separately and then conduct a fusion at the end to identify the final sound object, we jointly model both visual and binaural audio features and propose a new flipping operation for leveraging the horizontal symmetry of the binaural audio-visual data to further localize sound source in pixel level.

Network Architecture

The architecture of the proposed BAVNet is illustrated in Fig. 3, which takes video frames, binaural audio, single left audio and single right audio as the input. We first extract features from video sequence, binaural audio, single left recording and single right recording concurrently, which are followed by a ConvLSTM layer to model the temporal information for each of the audio branch. Then we fuse the image features and the binaural audio features together, and the output is going to be combined with the mapping results produced by single left and right recording features. The concatenation of these two feature maps are then fed into a ConvLSTM layer to model the dynamic changes. Finally, the decoder which contains convolution and upsampling layers is used to recover the resolution and generate the localization map.

Video Feature Extraction

Following the way to extract multi-level feature for static images with VGG16 in Cornia et al. (2016), we use the VGG19 (pretrained on ImageNet) as the encoder (ImageEncoder in Fig. 3) and take the features from three layers to form the multi-level feature I including the third and fourth max-pooling layers and the last convolution layer. Specifically, we remove the last max-pooling layer and modify the stride of the fourth max-pooling layer to 1. Finally, the size of I is $\frac{w}{8} \times \frac{h}{8}$, where w and h denote the width and height of the input frame, respectively.

Binaural Audio Feature Extraction

We first perform short-time Fourier transform (STFT) (Griffin and Lim 1984) on both left and right audio recordings to obtain two spectrograms. Fig. 4 shows a sample of binaural audio spectrogram representation along with the corresponding audio.

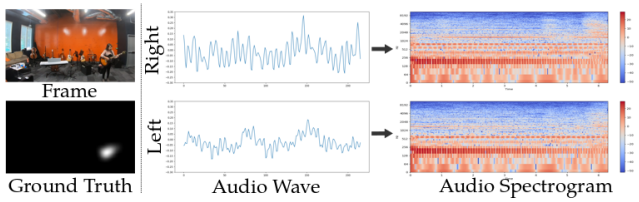

Figure 4: Binaural audio waves and the corresponding transformed spectrograms by applying STFT.

The binaural audio encoder (BiAudioEncoder in Fig. 3) takes the concatenation of left and right audio spectrograms as input. The BiAudioEncoder contains nine convolutional blocks and each block is composed of a 2D convolutional layer, a batch normalization layer and a LeakyReLU layer. After performing the encoding, the pair of audio spectrograms is converted to a one-dimensional audio feature. To map the audio information to the visual space, we reshape it to the resolution of $\frac{w}{8} \times \frac{h}{8}$ and obtain an initial spatial representation (A_b) for binaural audio. Then we feed A_b into a ConvLSTM (Shi et al. 2015) layer, which is to capture the temporal information of binaural audios:

$$A'_b, C_t, H_t = \text{ConvLSTM}(A_b, C_{t-1}, H_{t-1}), \quad (1)$$

where C_t and H_t are respectively the current cell state and hidden state, while C_{t-1} and H_{t-1} are respectively the preceding cell state and hidden state. Then, the output A'_b is passed into a softmax layer (Softmax) to obtain the final spatial representation for binaural audio:

$$\tilde{A}_b = \text{Softmax}(A'_b). \quad (2)$$

Jointly Modeling Left and Right Recordings and Video

Firstly, inspired by the widely-used soft-attention mechanism, we fuse the binaural audio feature ($\tilde{A}_b$) and frame feature (I) with a residual connection where the fusion result F_1 is defined by

$$F_1 = \tilde{A}_b \odot I + I, \quad (3)$$

where the operator $\odot$ denotes the elementwise multiplication. Then we perform a convolution operation on F_1 to produce F'_1. Following the similar way for binaural audio encoding, we convert each channel of the binaural audio (left and right) to its corresponding spatial representation ($\tilde{A}_l$ and $\tilde{A}_r$) respectively. Left and right audio encoders and the following ConvLSTM layer (AudioEncoder in Fig. 3) share weights. Such a design partly comes from that the same function needs to be applied to map both left and right audio recordings onto a frame. Due to the fact that the left and right audio recordings are horizontally symmetrical in visual scenes as shown in Fig. 2, we also make use of the horizontal symmetry by applying flipping operation to learn the mapping function (the common weight-sharing encoder) between the visual and auditory modalities.

To be specific, the single left audio feature $\tilde{A}_l$ is multiplied to the former fusion result F'_1, and the single right audio feature $\tilde{A}_r$ is multiplied to the flipped fusion result $flip(F'_1)$. The final fusion result F_2 is then defined by

$$F_2 = \langle \tilde{A}_l \cdot F'_1 + F'_1,\ \text{flip}(\tilde{A}_r \cdot \text{flip}(F'_1) + \text{flip}(F'_1)) \rangle, \quad (4)$$

where $\langle \cdot, \cdot \rangle$ denotes the concatenation operation. F_2 is then fed into a convolution layer and followed by ConvLSTM layer which is able to capture dynamic changes in temporal dimension for both videos and audios.

Finally, a decoder which consists of six convolution layers and three bilinear upsampling layers in between is used to recover the resolution back to the original input size and generate the sound source localization map in pixel level.

Loss Function

We first choose the Kullback-Leibler divergence and the Mean Squared Error to be part of the loss functions for the ground-truth sound source localization map (GT) and the predicted sound source localization map (P), where the KL divergence is a distribution-based loss function which was widely used for visual saliency estimation. We adopt it here in order to evaluate the localization prediction with a probabilistic interpretation, while the MSE loss can constrain the prediction to be pixel-wisely similar to the ground truth. In addition, based on the observation that human can very roughly localize the sound source by using their binaural system only, we also add two additional terms to the loss function, which minimize the distance between the intermediate result A and the rescaled ground-truth map (1/8 of the original resolution). Thus, our loss function is finally defined as follows:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{KL}(GT, P) + \beta \mathcal{L}_{MSE}(GT, P) + \\ \alpha(\mathcal{L}_{KL}(\mathcal{S}(GT), \tilde{A}_b) + \beta \mathcal{L}_{MSE}(\mathcal{S}(GT), \tilde{A}_b)), \end{aligned} \quad (5)$$

where α and β are weighting parameters (set to be respectively 0.2 and 100 in all experiments), $\mathcal{L}_{KL}$ and $\mathcal{L}_{MSE}$ are the Kullback-Leibler divergence and Mean Squared Error respectively, and the rescaling function $\mathcal{S}$ is used to rescale the ground-truth map GT to be the same scale of $\tilde{A}_b$.

We calculate the probabilistic representation T and F for the ground-truth localization map GT and the predicted localization map P, respectively:

$$\begin{aligned} T(i,j) &= \frac{GT(i,j)}{\sum_{(i,j)} GT(i,j) + \epsilon}, \\ F(i,j) &= \frac{P(i,j)}{\sum_{(i,j)} P(i,j) + \epsilon}, \end{aligned} \quad (6)$$

where ϵ is set to be 10^{-20}, then the K-L divergence L_{KL} is defined as:

$$\mathcal{L}_{KL} = \sum_{(i,j)} T(i,j) \log \left(\frac{T(i,j)}{F(i,j)} + \epsilon \right). \quad (7)$$

The Mean Squared Error L_{MSE} is defined as:

$$\mathcal{L}_{MSE} = \frac{1}{hw} \sum_{(i,j)} (GT(i,j) - P(i,j))^2. \quad (8)$$

Experimental Results

Datasets

To generate accurate sound source localization maps and fulfill the purpose of supervision for training, we manually label the datasets, FAIR-Play and YT-MUSIC, by annotating a set of points to construct pixel-level ground-truth locations of the sound sources, as illustrated in Fig. 5. Note that, we only annotate the middle frame for each video clip and the number of points we use on that frame depends on the content of the frame, e.g. the number of sound sources and the area of each sound source on that frame. Then we apply the Gaussian blur with the radius of 4 to the labeled points to produce a continuous representation of the sound source localization map as the ground truth.

(a) (b) (c)

Figure 5: An illustration of the ground-truth annotation of sound source localization. (a) A frame, (b) the labeled points, and (c) the sound-source localization map by Gaussian blur.

FAIR-Play (Gao and Grauman 2019a): FAIR-Play is the first audio-visual dataset recorded with both videos and professional binaural audios in a music room. Specifically, audios are recorded by 3Dio binaural microphones and a GoPro is mounted on the top to record the corresponding videos, i.e., the whole system is trying to simulate the auditory and vision of humans to collect data (See Gao and Grauman (2019a) for a vivid description). In total, the dataset consists of 1,871 pairs of videos with a resolution of 320×176 and binaural audios (10 seconds for each). For each video, we extract 102 frames with a sampling rate of 10 and manually label the sound sources in the middle frame (the fiftieth one). We use the 'split1' proposed in Gao and Grauman (2019a) to construct our train/test splits for training and evaluation.

YT-MUSIC (Morgado et al. 2018): The YT-MUSIC dataset is collected from Youtube for spatial audio generation by Morgado et al. (2018), which contains 397 videos that are all 360° videos with resolution of 448×224. Because a small number of videos have been removed by the creators and some videos have inconsistent audio and video, we finally use 317 videos (a subset of the original dataset) with 250 videos for training and 67 for testing. Following the guidance of Gao and Grauman (2019a), we use the head related transfer function (HRTF) to transfer the ambisonic audio into the corresponding binaural audio.

Both datasets include singing and instruments playing scenarios with indoor (FAIR-Play, YT-MUSIC) and outdoor (YT-MUSIC) cases.

Metrics

Given the format similarity between the sound-source localization map in our work and image-saliency map, we propose to use three metrics that are often used in saliency prediction (Jiang et al. 2018; Bak et al. 2017) for quantitatively evaluating the pixel-level sound-source localization accuracy: Pearson's correlation coefficient (CC), Similarity Metric (SIM) and Earth Mover's Distance (EMD). For this purpose, we normalize both the ground-truth localization map (after Gaussian blur) and the predicted localization map to probability maps (adding all the elements of a map is one), and then reshape them into vectors before applying the three metrics.

CC is a statistical method to measure the linear correlation between two normalized variables, which has been widely used for saliency detection. The value of CC is range from -1 to 1, where 1 is the perfect positive linear correlation, 0 means no correlation and -1 represents negative linear correlation.

SIM measures the similarity between two distributions, which was firstly introduced for evaluating image matching accuracy (Swain and Ballard 1991). SIM being equal to one means that the distributions are identical. The larger SIM, the better the similarity.

EMD is used to measure the spatial distance between two distributions by computing the minimum transformation cost that one distribution would take to match the other, which is first introduced for image matching (see (Rubner, Tomasi, and Guibas 2000) for details). The EMD value of two identical distributions is 0.

Model Specification

Training setting. BAVNet is implemented using Pytorch and trained with one Nvidia 2080Ti GPU. We take Adam as the optimizer by setting weight decay to be 0.0001. The starting learning rate is set to 0.0001, then it decayed by multiplying it with the decay factor 0.8 for every 10 epochs. We train the network for 200 epochs in total with the batch size being 1.

More details. For video data pre-processing, we randomly pick a video clip whose length ranges from 20 to 50 frames from each video. Note that the last frame of each clip is fixed at the one which has the label for later use of calculating the loss. We use the data augmentation strategy of horizontal flipping (the frame is horizontally flipped while the left-right channels of the binaural audio are swapped with each other). We also randomly shift the audio segmentation window and the shifting ranges from -1 ms to 1 ms, while the video is fixed. For the audio data pre-processing, we first resample the audio at 16 kHz, then STFT is calculated with the window length of 64 ms, the hop length of 8 ms and the FFT size of 512.

Baselines and Comparison Results

We compare our full-setting model (BAVNet) with the following baselines to evaluate the proposed method:

Methods	FAIR-Play			YT-MUSIC		
	CC	SIM	EMD	CC	SIM	EMD
Video-only	0.679	0.544	2.129	0.375	0.325	4.532
Single left audio w/ video	0.742	0.579	1.965	0.415	0.337	4.323
Single right audio w/ video	0.741	0.582	1.959	0.414	0.335	4.326
Monaural audio w/ video	0.743	0.583	1.698	0.414	0.332	4.319
Waveform w/ video	0.693	0.556	2.097	0.383	0.329	4.458
BAVNet	**0.776**	**0.625**	**1.618**	**0.434**	**0.364**	**4.312**

Table 1: Quantitative comparisons of baseline approaches and the proposed BAVNet on the FAIR-Play test set and the YT-MUSIC test set. For CC and SIM, the higher the better, while for EMD, the lower the better.

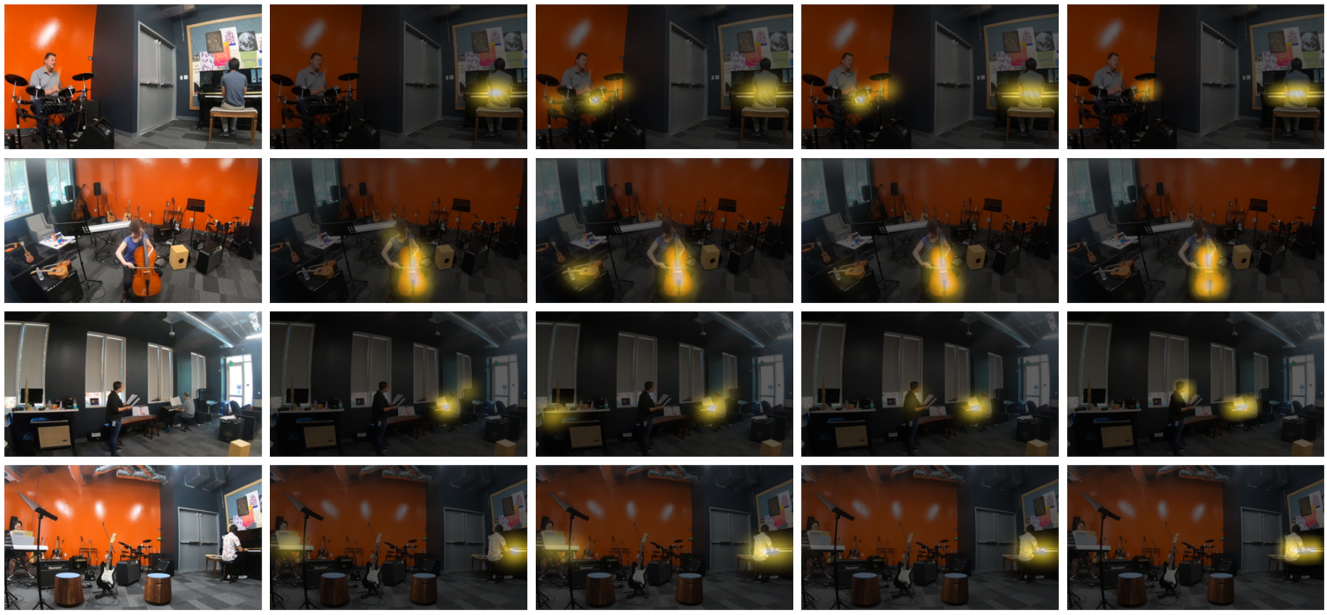

Figure 6: Visual comparisons of (a) the video-only, (b) the monaural audio with video, and (c) the proposed BAVNet on the FAIR-Play test set.

- **Video-only:** Only use the visual information to predict sound sources.
- **Single audio with video:** Use only either left channel or right channel as the audio input.
- **Monaural audio with video:** For each video clip, we add the left and right channels together to form the one channel monaural audio. The format of the input is the same as many previous methods (Senocak et al. 2018; Arandjelovic and Zisserman 2018; Owens and Efros 2018; Zhao et al. 2018; Hu, Nie, and Li 2019).
- **Waveform of binaural audio:** Uses the raw binaural audio wave as the input instead of the converted spectrogram. When the waveform is applied, we change all the 2D convolution operations into 1D convolution for the binaural audio encoder.

Comparison results with the baselines are reported in Table 1 and our BAVNet clearly outperforms all the others on both datasets. It is obvious that using both audio and video gets better performance than using video only. In the human auditory system, our brain can perceive subtle differences in intensity, spectral and timing to localize sound sources (Hearing 1983; Thompson 2018) whose cues are not kept by single audio or the monaural audios. Similarly, our method also shows that the binaural auditory modality does further significantly improve the performance of sound source localization than the monaural auditory one.

Some visual comparisons are shown in Fig. 6, where we again can see that the binaural audio really plays an important role in localizing the sound sources. For example, only the BAVNet captures that the woman is singing in the third picture, and without the binaural audio, the network localizes the left piano in the last picture by mistake.

We also observe that the spectrogram is a better representation compared to the raw audio wave when using BAVNet to solve the sound source localization problem. We hypothesize that the reason is that the spectrogram can offer a more intuitive representation for the differences of amplitude, frequency and timing cues than the waveform.

Model variants		FAIR-Play CC	FAIR-Play SIM	YT-MUSIC CC	YT-MUSIC SIM
BAVNet		**0.776**	**0.625**	**0.434**	**0.364**
Visual branch	w/o multi-level feature	0.658	0.583	0.407	0.323
Binaural audio branch	w/o the branch	0.761	0.615	0.431	0.357
Single audio branches	w/o the two branches	0.752	0.592	0.421	0.332
	w/o only flipping operation	0.749	0.591	0.422	0.339
Fusion	w/o residual connection	0.656	0.578	0.411	0.329
-	w/o ConvLSTM	0.611	0.429	0.357	0.314
Loss function	w/o KL loss	0.701	0.609	0.412	0.325
	w/o MSE loss	0.704	0.612	0.419	0.325
	w/o auxiliary loss function	0.769	0.621	0.432	0.357

Table 2: Comparisons on several model variants for justification of the main components of the proposed BAVNet on the FAIR-Play test set and the YT-MUSIC test set.

Ablation Study and Analysis

To evaluate the effectiveness of the main components of our BAVNet, we also conduct ablation studies on several model variants on the FAIR-Play and YT-MUSIC datasets. The following components are justified: 1) the visual branch; 2) the binaural audio branch; 3) the single audio branch; 4) the fusion operation; 5) the dynamic change modeling and 6) the loss function (auxiliary loss, KL loss and MSE loss).

Results are reported in Table 2, from which we can see that all of these components of BAVNet are useful. For the visual branch, the multi-level feature from the image encoder is helpful for the image feature extraction. Both of the binaural audio branch and the single audio branch can help encode the audio feature for the network. Specifically, the flipping operation for the single audio branch is crucial to map the audio feature to the visual space. The residual connection for fusing the audio feature and the image feature can improve the performance, while three ConvLSTMs in the audio branch and another one before decoder are essential to obtain the dynamic change from the previous state. In the end, the auxiliary loss terms on the intermediate result and the combination of the KL loss and the MSE loss are also well justified.

To compare our proposed BAVNet with other visual-audio localization networks, we implement several state-of-the-art methods and train them by adding a supervised loss against the ground-truth annotations as done in our method. Table 3 includes the auido localization performance of the AVOLNet (Arandjelovic and Zisserman 2018) and two visual-audio network based on (Owens and Efros 2018; Senocak et al. 2018).

Model	CC	SIM
AVOL-Net	0.453	0.398
Owens*et al.*	0.748	0.540
Senocak *et al.*	0.712	0.603
BAVNet	**0.776**	**0.625**

Table 3: Comparisons of the proposed BAVNet with some existing audio localization models on the FAIR-Play test set.

Figure 7: The sound source localization results of three examples from the YT-Music test set by using the proposed BAVNet. The top row is the input frame, the middle row is the prediction by the proposed BAVNet and the bottom row is the ground truth.

In Fig. 7, we provide some examples of the sound source localization results predicted by our BAVNet on the YT-MUSIC dataset, which are visually very consistent with the ground truth.

Conclusion

In this paper we studied the sound source localization problem by stressing that the sound sources should be more concentrated and accurately represented in pixel level instead of object level. Following the way that humans usually use to localize the sound sources, we proposed a novel Binaural Audio-Visual Network (BAVNet) which takes in both visual and binaural auditory information as the input, to generate the source localization map in an end-to-end fashion. We also manually label two existing datasets, FAIR-Play and YT-MUSIC, to provide pixel-level supervision for the training of BAVNet. From the experimental results, we found that binaural audio can greatly promote the localization accuracy than monaural audio, especially when the visual scene contains some degradation effects.

References

Arandjelovic, R.; and Zisserman, A. 2017. Look, Listen and Learn. In *IEEE International Conference on Computer Vision (ICCV)*.

Arandjelovic, R.; and Zisserman, A. 2018. Objects that sound. In *European Conference on Computer Vision (ECCV)*, 435–451.

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, 892–900.

Bak, C.; Kocak, A.; Erdem, E.; and Erdem, A. 2017. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia* 20(7): 1688–1698.

Barzelay, Z.; and Schechner, Y. Y. 2007. Harmony in motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8. IEEE.

Christensen, J. H.; Hornauer, S.; and Yu, S. 2019. BatVision: Learning to See 3D Spatial Layout with Two Ears. *arXiv preprint arXiv:1912.07011* .

Cornia, M.; Baraldi, L.; Serra, G.; and Cucchiara, R. 2016. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 3488–3493. IEEE.

Fisher III, J. W.; Darrell, T.; Freeman, W. T.; and Viola, P. A. 2001. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in neural information processing systems*, 772–778.

Gan, C.; Zhao, H.; Chen, P.; Cox, D.; and Torralba, A. 2019. Self-Supervised Moving Vehicle Tracking With Stereo Sound. In *IEEE International Conference on Computer Vision (ICCV)*.

Gao, R.; Feris, R.; and Grauman, K. 2018. Learning to separate object sounds by watching unlabeled video. In *European Conference on Computer Vision (ECCV)*, 35–53.

Gao, R.; and Grauman, K. 2019a. 2.5D Visual Sound. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gao, R.; and Grauman, K. 2019b. Co-Separating Sounds of Visual Objects. In *IEEE International Conference on Computer Vision (ICCV)*.

Griffin, D.; and Lim, J. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32(2): 236–243.

Hearing, S. 1983. The psychophysics of human sound localization. *J. Blauert* .

Hershey, J. R.; and Movellan, J. R. 2000. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*, 813–819.

Hu, D.; Nie, F.; and Li, X. 2019. Deep Multimodal Clustering for Unsupervised Audiovisual Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Izadinia, H.; Saleemi, I.; and Shah, M. 2012. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia* 15(2): 378–390.

Jiang, L.; Xu, M.; Liu, T.; Qiao, M.; and Wang, Z. 2018. Deepvs: A deep learning based video saliency prediction approach. In *European Conference on Computer Vision (ECCV)*, 602–617.

Morgado, P.; Nvasconcelos, N.; Langlois, T.; and Wang, O. 2018. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*, 362–372.

Owens, A.; and Efros, A. A. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, 631–648.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40(2): 99–121.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and So Kweon, I. 2018. Learning to Localize Sound Source in Visual Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, 802–810.

Swain, M. J.; and Ballard, D. H. 1991. Color indexing. *International journal of computer vision* 7(1): 11–32.

Thompson, D. M. 2018. *Understanding audio: getting the most out of your project or professional recording studio*. Hal Leonard Corporation.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision (ECCV)*, 247–263.

Wu, Y.; Zhu, L.; Yan, Y.; and Yang, Y. 2019. Dual Attention Matching for Audio-Visual Event Localization. In *IEEE International Conference on Computer Vision (ICCV)*.

Xu, X.; Dai, B.; and Lin, D. 2019. Recursive Visual Sound Separation Using Minus-Plus Net. In *IEEE International Conference on Computer Vision (ICCV)*.

Zhao, H.; Gan, C.; Ma, W.-C.; and Torralba, A. 2019. The Sound of Motions. In *IEEE International Conference on Computer Vision (ICCV)*.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The sound of pixels. In *European Conference on Computer Vision (ECCV)*, 570–586.

Zhou, H.; Liu, Z.; Xu, X.; Luo, P.; and Wang, X. 2019. Vision-Infused Deep Audio Inpainting. In *IEEE International Conference on Computer Vision (ICCV)*.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.