Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation

Hang Zhou*, Xudong Xu*, Dahua Lin, Xiaogang Wang, and Ziwei Liu

CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
{zhouhang@link,xx018@ie,dhlin@ie,xgwang@ee}.cuhk.edu.hk
zwliu.hust@gmail.com

Abstract. Stereophonic audio is an indispensable ingredient to enhance human auditory experience. Recent research has explored the usage of visual information as guidance to generate binaural or ambisonic audio from mono ones with stereo supervision. However, this fully supervised paradigm suffers from an inherent drawback: the recording of stereophonic audio usually requires delicate devices that are expensive for wide accessibility. To overcome this challenge, we propose to leverage the vastly available mono data to facilitate the generation of stereophonic audio. Our key observation is that the task of visually indicated audio separation also maps independent audios to their corresponding visual positions, which shares a similar objective with stereophonic audio generation. We integrate both stereo generation and source separation into a unified framework, **Sep-Stereo**, by considering source separation as a particular type of audio spatialization. Specifically, a novel associative pyramid network architecture is carefully designed for audio-visual feature fusion. Extensive experiments demonstrate that our framework can improve the stereophonic audio generation results while performing accurate sound separation with a shared backbone[1].

1 Introduction

Sight and sound are both crucial components of human perceptions. Sensory information around us is inherently multi-modal, mixed with both pixels and vocals. More importantly, the stereophonic or spatial effect of sound received by two ears gives us the superiority to roughly reconstruct the layout of the environment, which complements the spatial perception in the vision system. This spatial perception of sound makes it appealing for content creators to create audio information with more than one channel. For example, the user's experience will be greatly promoted if stereo music instead of mono is provided when watching a recording of a concert.

However, it is still inconvenient for portable devices to record stereophonic audio. Normally, cell phones and cameras have only mono or line array microphones that can not record real binaural audio. To achieve such goals, dummy

[*] Equal Contribution.

[1] Code, models and demo video are available at https://hangz-nju-cuhk.github.io/projects/Sep-Stereo.

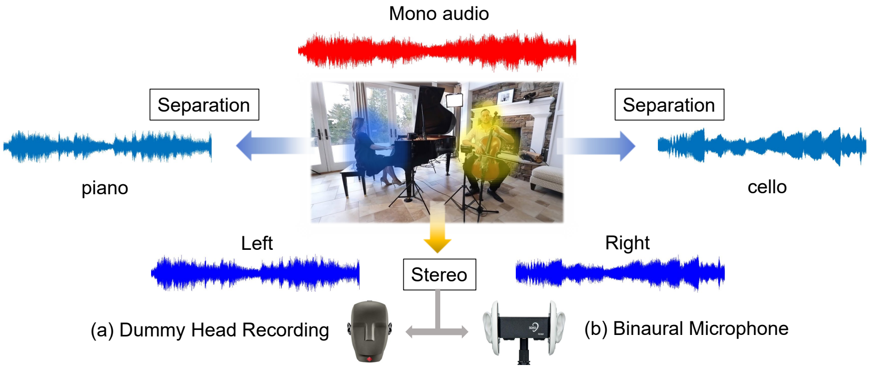


Fig. 1. We propose to integrate audio source separation and stereophonic generation into one framework. Since both tasks build associations between audio sources and their corresponding visual objects in videos. Below is the equipment for recording stereo: (a) dummy head recording system, (b) Free Space XLR Binaural Microphone of 3Dio. The equipment is not only expensive but also not portable. This urges the need for leveraging mono data in stereophonic audio generation

head recording systems or fake auricles need to be employed for creating realistic 3D audio sensations that humans truly perceive. This kind of system requires two microphones attached to artificial ears (head) for mimicking a human listener in the scenery, as shown in Fig. 1. Due to the cost and weight of the devices, such recorded binaural audio data is limited, particularly the ones associated with visual information. Therefore, developing a system to generate stereophonic audio automatically from visual guidance is highly desirable.

To this end, Gao and Grauman [17] contribute the FAIR-Play dataset collected with binaural recording equipment. Along with this dataset, they propose a framework for recovering binaural audio with the mono input. Nevertheless, their method is built in a data-driven manner, fusing visual and audio information with a simple feature concatenation, which is hard to interpret. Moreover, the data is valuable yet insufficient to train a network that can generalize well on all scenarios. Similarly, Morgado *et al.* [30] explore to use 360° videos uploaded on youtube for generating ambisonic audio.

On the other hand, videos recorded with a single channel are much easier to acquire. Ideally, it would be a great advantage if networks can be benefited from training with additional mono-only audios without stereophonic recordings. This paradigm of leveraging unlabeled data has rarely been explored by previous research because of its inherent challenge. We observe that **an essential way to generate stereo audio requires disentangling multiple audio sources and localizing them, which is similar to the task of visually indicated sound source separation** [47,44,46,12]. While previous methods have also explored performing separation along with generating spatial audio, they either implement it as an intermediate representation by using stereo data as supervision [30], or train another individual network for separation [17]. Thus it is valuable to investigate a new way of leveraging both mono and stereo together for solving both problems. Moreover, it is particularly desirable if the two tasks can be solved within one unified framework with a shared backbone as illustrated in Fig. 1. However, there are considerable differences between these two tasks and great challenges to overcome. The details are explained in Section 3.1.

Our key insight is to **regard the problem of separating two audios as an extreme case of creating binaural audio**. More specifically, we perform duet audio separation with the hypothesis that the two sources are only visible at the edges of human sight. And no visual information can be provided about the whole scene. Based on this assumption and our observation above, we propose to explicitly associate different local visual features to different responses in spectrograms of audios. In this manner, the intensity of sound can be represented on both the image domain visually and the audio domain auditorily. A novel **associative pyramid network** architecture is proposed for better fusing the information within the two modalities. The whole learning process can be divided into two parts, namely **separative learning** and **stereophonic learning**. . We perform multi-task training with two different sets of data: mono ones (MUSIC [47]) for separation, and binaural ones (FAIR-Play [17]) for stereophonic audio generation. Following traditional multi-task learning settings, the two learning stages share a same backbone network but learn with different heads with the same architecture. This framework is uniformly called **Sep-Stereo**.

Extensive experiments regarding stereophonic audio generation have validated the effectiveness of our proposed architecture and learning strategy. At the same time, the trained model can preserve competitive results on the task of audio separation simultaneously. Moreover, we show that with the aid of mono data, our framework is able to achieve generalization under a low data regime using only a small amount of stereo audio supervision.

Our **contributions** are summarized as follows: **1)** We unify audio source separation and stereophonic audio generation into a principled framework, **Sep-Stereo**, which performs joint **separative** and **stereophonic** learning. **2)** We propose a novel **associative pyramid network** architecture for coupling audio-visual responses, which enables effective training of both tasks simultaneously with a shared backbone. **3)** Our **Sep-Stereo** framework has a unique advantage of leveraging mono audio data into stereophonic learning. Extensive experiments demonstrate that our approach is capable of producing more realistic binaural audio while preserving satisfying source separation quality.

2 Related Works

2.1 Joint Audio-Visual Learning

The joint learning of both audio and visual information has received growing attention in recent years [53,19,15,35,23]. By leveraging data within the two modalities, researchers have shown success in learning audio-visual self-supervision [4,2,3,25,31,22], audio-visual speech recognition [21,39,48,45], localization [47,38,37,34], event localization (parsing) [41,43,40], audio-visual navigation [13,5], cross-modality generation between the two modalities [9,51,8,6,48,7,52,49,42,50] and so on. General representation learning across the two modalities is normally conducted in a self-supervised manner. Relja *et al.* [2,3] propose to learn the association between visual objects and sound, which supports localizing the objects that sound in an image. Owens *et al.* [31] and Korbar *et al.* [25] train neural networks to predict whether video frames and audios are temporally aligned. Researchers have also explored the possibility of directly generating sound according to videos [32,8,51], which is more related to our task. Different from their aims, our Sep-Stereo framework exploits visual-audio correspondence to improve the generation of stereophonic audios.

2.2 Audio Source Separation and Spatialization

Source Separation. Source separation with visual guidance has been an interest of research for decades [11,29,33,16,47,46,1,44,10,12]. Compared with audio-only source separation, visual information could provide rich clues about the types and movements of audio sources. Thus the performance of audio separation is expected to improve with the guidance of vision. Recently, deep learning has been widely applied into this filed of research. For separating speech segments, Afouras *et al.* [1] propose to leverage mouth movements as guidance, and Ephrat *et al.* [10] use cropped human faces. Owens *et al.* [31] do not crop faces and modify their pipeline from learning synchronization. On the other hand, instrumental music [47,46,16,18,44] is the field that we care more about. Gao *et al.* [16] propose to combine non-negative matrix factorization (NMF) with audio features. Zhao *et al.* [47] use a learnable U-Net instead, and match feature activations with different audio channels. Based on this work, motion information is merged into the main framework to achieve better performance in [46]. In [18], object detection and instrument labels are leveraged to co-separating sound. In our work, we will not model motion explicitly, thus adopt a similar setting as [47].

Spatialization. Visually guided audio spatialization has received relatively less attention [26,28,30,17,14] compared with separation. Recently, Li *et al.* [26] leverage synthesised early reverberation and measured late reverberation tail to generate stereo sound in a specific room, which cannot generalize well to other scenarios. With the assistance of deep learning, Morgado *et al.* [30] propose to generate ambisonic for 360° videos using recorded data as self-supervision. The work mostly related to ours is [17]. They contribute a self-collected binaural audio dataset, and propose a U-Net based framework for mono-to-binaural generation on normal field of view (NFOV) videos. These works all only leverage the limited stereophonic audio. In this paper, we propose to boost the spatialization performance with additional mono data.

3 Our Approach

Our proposed framework, **Sep-Stereo**, is illustrated in Fig. 2. This whole pipeline consists of two parts: (a) stereophonic learning and (b) separative learning. We will first introduce the overall framework of visually guided stereophonic audio generation and source separation (section 3.1). Then we demonstrate how our proposed network architectures can effectively associate and integrate audio and visual features into a unified network with a shared backbone.

3.1 Framework Overview

Stereophonic Learning. The whole process of stereophonic learning is depicted in the lower part in Fig. 2. In the setting of stereophonic learning, we care for the scenario that human perceives and has access to binaural data. The visual information V_s corresponds to its audio recording of the left ear $a_l(t)$ and the right-ear one $a_r(t)$. Notably, all spatial information is lost when they are averaged to be a mono clip $a_{mono} = (a_l + a_r)/2$, and our goal is to recover left and right given the mono and video. We operate in the Time-Frequency (TF) domain by transferring audio to spectrum using Short-Time Fourier Transformation (STFT) as a common practice. Here we use S_l^t and S_r^t to denote the STFT of the ground truth left and right channels, with t here represents "target". The input of our network is the mono audio by averaging the STFT of the two audio channels:

$$S_{mono} = (S_l^t + S_r^t)/2 = \mathrm{STFT}(a_{mono}). \tag{1}$$

This can be verified by the property of the Fourier Transformation. Please note that due to the complex operation of STFT, each spectrum $S = S_R + j * S_I$ is a complex matrix that consists of the real S_R and imagery part S_I. So the input size of our audio network is $[T, F, 2]$ by stacking the real and imagery channels.

Separative Learning. The task of separation is integrated for its ability to leverage mono data. Our separative learning follows the Mix-and-Separate training procedure [47], where we elaborately mix two independent audios as input and manage to separate them using ground truth as supervision. It is illustrated at the top of Fig. 2. Given two videos V_A and V_B with only mono audios accompanied, the input of the separative phase is the mixture of two mono audios $a_{mix} = (a_A + a_B)/2$. They can be represented in the STFT spectrum domain as S_A^t, S_B^t and S_{mix}. Aiming at disentangling the mixed audio, separative learning targets at recovering two mono audios with the guidance of corresponding videos.

Connections and Challenges. Apart from our observation that both tasks connect salient image positions with specific audio sources, they all take mono audio clips as input and attempt to split them into two channels. One can easily find a mapping from stereo to separation as: $\{S_{mono} \Rightarrow S_{mix}, S_l^t \Rightarrow S_A^t, S_r^t \Rightarrow S_B^t\}$. From this point of view, the two tasks are substantially similar. However, the goals of the two tasks are inherently different. While each separated channel should contain the sound of one specific instrument, both sources should be audible, *e.g.* in the task of stereo for a scene shown in Fig. 1. Also, the spatial effect would exist if there is only one source, but separation is not needed in such a case. As for the usage of visual information, the separation task aims at finding the most salient area correctly while the stereo one is affected by not only the sources' positions but also the environment's layout. Thus neither an existing stereo framework [17] nor a separation one [47] is capable of handling both tasks.

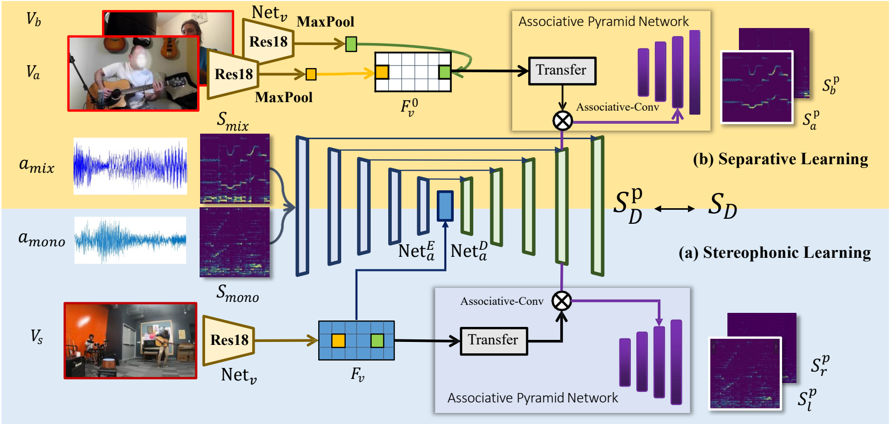


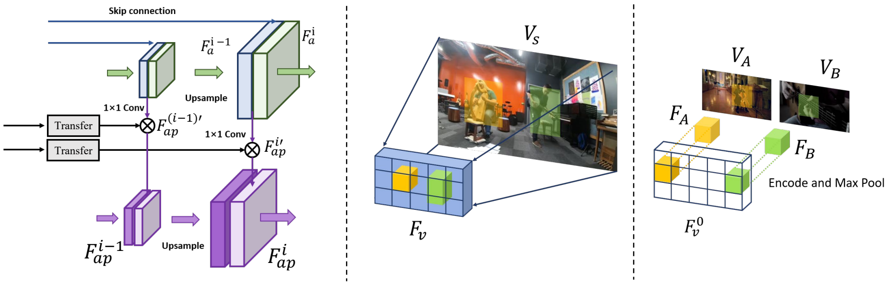
Fig. 2. The whole pipeline of our Sep-Stereo framework. It aims at learning the associations between visual activation and the audio responses. The framework consists of a U-Net backbone whose weights are shared, and our proposed Associative Pyramid Network (APNet). The visual features are firstly extracted and fused with audio features through a multi-scale feature mapping network (APNet). To perform multitask training, (a) stereophonic learning and (b) separative learning stages are leveraged to tackle this problem. One set of APNet parameters are trained for each task

3.2 Associative Neural Architecture

Backbone Network. Our audio model is built upon Mono2Binaural [17], which is a conditional U-Net [36]. This audio backbone is denoted as Net_a. It consists of skip-connected encoder Net_a^E and decoder Net_a^D. The visual features are extracted using a visual encoder with ResNet18 [20] architecture called Net_v. The input video clip with input dimension $[T, W_v, H_v, 3]$ is encoded into a feature map F_v of size $[w_v, h_v, C_v]$, by conducting a temporal max pooling operation. We assume that the feature map would correspond to each spatial part in the original video with high responses on salient positions.

Associative Pyramid Network. Based on the backbone network, we propose a novel Associative Pyramid Network (APNet) for both learning stages. It is inspired by PixelPlayer [47] that maps one vision activation with one source feature map, but with a different formulation and underlying motivation. Our key idea is to associate different intensities of audio sources with different vision activations in the whole scene with feature map re-scheduling. As illustrated in Fig. 2 and 3, it works as a side-way network along-side the backbone in a coarse-to-fine manner.

We operate on each layer of the decoder Net_a^D in the U-Net after the upsample deconvolutions. Suppose the ith deconv layer's feature map F_a^i is of shape $[W_a^i, H_a^i, C_a^i]$, we first reshape F_v to $[(w_v \times h_v), C_v]$ and multiply it by a learned weight with size $[C_v, C_a^i]$ to be K_v^i with dimension $[1, 1, C_a^i, (h_v \times w_v)]$. This is called the kernel transfer operation. Then K_v^i operates as a 1×1 2D-convolution



(a) APNet Structure **(b) Stereophonic Learning** **(c) Separative Learning**

Fig. 3. Figure (a) is the architecture of the Associative Pyramid Network. After kernel transfer of the visual feature, it associates audio channels with spatial visual information through 1 × 1 convolution and coarse-to-fine tuning. Figure (b) and (c) shows the difference between stereo and separative learning with respect to visual embedding. For (b) stereophonic learning, a visual feature map F_v is directly extracted, with each bin corresponds to a region in the image. While in (c), the visual feature is max-pooled to a 1 × 1 bin and manually sent to an all-zero feature map F_v^0. This process is the *rearrangement* of visual feature

kernel on the audio feature map F_a^i, and renders an entangle audio-visual feature $F_{ap}^{i'}$ of size $[W_a^i, H_a^i, (h_v \times w_v)]$. This process can be formulated as:

$$F_{ap}^{i'} = \underset{K_v^i}{\text{Conv2d}}(F_a^i). \tag{2}$$

Note that each $[1, 1, C_a^i]$ sub-kernel of K_v^i corresponds to one area in the whole image space. So basically, audio channels and positional vision information are associated through this learned convolution. This operation is named Associative-Conv. The output feature $F_{ap}^{i'}$ can be regarded as the stack of audio features associated with different visual positions.

The $F_{ap}^{1'}$ is the first layer of APNet F_{ap}^1. When $i > 1$, the $(i-1)$th feature map F_{ap}^{i-1} will be upsampled through a deconvolution operation to be of the same size as $F_{ap}^{i'}$. Then a new feature map F_{ap}^i can be generated by a concatenation:

$$F_{ap}^i = \text{Cat}([\text{DeConv}(F_{ap}^{i-1}), F_{ap}^{i'}]). \tag{3}$$

In this way, the side-way APNet can take advantage of the pyramid structure by coarse-to-fine tuning with both low-level and high-level information.

The goal of APNet is to predict the target left and right channels' spectrum in both learning stages. Thus, two parallel final convolutions are applied to the last layer of F_{ap}^i, and map it to two outputs with real and imagery channels. As discussed before, each channel in the APNet is specifically associated with one visual position, the final convolution acts also as a reconfiguration to different source intensities.

3.3 Learning Sep-Stereo

Directly predicting spectrum is difficult due to the large dynamic range of STFTs, so we predict the complex masks $M = \{M_R, M_I\}$ following [17] as our objectives. Suppose the input spectrum is $S_{mono} = S_{R(mono)} + j * S_{I(mono)}$, then a prediction can be written as:

$$S^p = (S_{R(mono)} + j * S_{I(mono)})(M_R + j * M_I). \tag{4}$$

The outputs of our networks are all in the form of complex masks, and the predictions are made by the complex multiplication stated above.
Stereophonic Learning. The base training objective of the backbone network for stereo is to predict the subtraction of the two spectrums $S^t_D = (S^t_l - S^t_r)/2$ as proposed in [17]. The left and right ground truth can be written as:

$$S^t_r = S_{mono} + S^t_D, \quad S^t_l = S_{mono} - S^t_D. \tag{5}$$

The backbone network is to predict the difference spectrum S^p_D, the training objective is:

$$L_D = ||S^t_D - S^p_D||^2_2. \tag{6}$$

The APNet's outputs are the left and right spectrums, the loss fuction is:

$$L_{rl} = ||S^t_l - S^p_l||^2_2 + ||S^t_r - S^p_r||^2_2. \tag{7}$$

Separative Learning. While the backbone network and APNet seem to be suitable for learning binaural audio, it has the advantage of handling separation by manually modifying features. Our key insight is to regard separation as an extreme case of stereo, that the visual information is only available at the left and right edges. Normally, the visual feature map F_v is a global response which contains salient and non-salient regions. During the separation stage, we manually create the feature map.

Specifically, we adopt max-pooling to the visual feature map F_v to be of size $[1, 1, C_v]$. The feature vectors for video A and B are denoted as F_A and F_B respectively. Then we create an all-zero feature map F^0_v of the same size as $[w_v, h_v, C_v]$ to serve as a dummy visual map. Then the max-pooled vectors are sent to the left and right most positions as illustrated in Fig. 3. It can be written as:

$$F^0_v(\lceil H/2 \rceil,\ 1) = F_A, \quad F^0_v(\lceil H/2 \rceil, W) = F_B. \tag{8}$$

Then we replace the F_v with F^0_v. This process is called the **rearrangement** for visual feature.

The intuition for the separative learning to work is based on our design that each channel of the APNet layers corresponds to one visual position. While with separative learning, take the left-ear for instance. Most information correlates to the left-ear is zero, thus the left-ear spectrum should correspond to only the left-most visual information. Training the separation task provides especially the backbone network with more audio spectrum and vision information instead of only overfitting to the limited binaural data. Besides, it is also assumed that the non-salient visual features also help our APNet understand the sound field. Thus, without the environment information, we can expect the network to implicitly ignore the distribution of sound around the space but focus on the two sides.

At the separative learning stage, only the APNet predicts the masks for audios A and B, as the left and right channels in the same way as Eq. 4. The predicted spectrums can be represented as S_a^p and S_b^p. So the training objective is:

$$L_{ab} = ||S_a^t - S_a^p||_2^2 + ||S_b^t - S_b^p||_2^2. \tag{9}$$

Final Objective. The final objective of our network is the combination of all the losses for training stereo and separation.

$$L_{all} = L_D + \lambda_1 L_{rl} + \lambda_2 L_{ab} \tag{10}$$

where λ_1 and λ_2 are loss weights that are empirically set to 1 in the experiments through cross-validation.

4 Experiments

4.1 Implementation Details

Preprocessing. We fix all of our audio sampling rate to 16kHz and clip the raw audios to ensure their values are between -1 and 1. For performing STFT, our window size is 512, and the hop length is 160. During stereophonic training, we sample a 0.63s clip randomly from the whole 10s video. Thus can lead to an STFT map with the size of $[257, 64]$. Separative learning samples a 0.63s clip from each individual video as well, and mixes them up as inputs. Other configurations are the same as [17]. The length of the sliding window for testing is 0.1s. The videos are extracted to frames at 10 fps. At each training time step, the center frame is used as the input of the visual embedding network.

Model Configurations. The backbone audio U-Net Net_a is borrowed from [17], which consists of 5 downsample convolution and 5 de-convolution layers with 4 skip connections between feature maps of the same scale. The Associative Pyramid Network consists of 4 Associative-Conv which couples visual features with audio features. Additionally, there are 3 upsampling operations in APNet. The visual embedding network Net_v is adopted from [47], which is a modified ResNet18 network [20]. The final pooling and fully-connected layers are removed from this network, and the dilation of the network's kernels is 2. Thus F_v is of size $[14, 7, 512]$, where 512 is its channel size.

Training Details. The networks are trained using Adam [24] optimizer with learning rate at 5e-4 and batch size 144. For stereophonic learning, we use the same data augmentation diagram as Mono2Binaural [17]. For separative learning, the amplitude of selected audio is augmented with a random scale disturb of 0.5

to 1.5. The separative learning part is firstly trained, then both data of stereo and separation are sent into the network at the same time. Our original design is to share the backbone and APNet parameters through both learning stages. However, it requires careful tuning for both tasks to converge simultaneously. In our final version, the parameters of Net_a and Net_v are **shared** across two learning stages while **different** sets of APNet parameters are trained for different stages. As the backbone takes up most of the parameters, sharing it with separative learning is the key for improving stereophonic learning. Moreover, visual information is also fused into the backbone, thus our insights all stand even without sharing the APNet parameters.

4.2 Datasets and Evaluation Metrics

In the sense of improving audiences' experiences, videos with instrumental music are the most desired scenario for stereophonic audio. Thus in this paper, we choose music-related videos and audios as a touchstone. Our approach is trained and evaluated on the following datasets:

FAIR-Play. The FAIR-Play dataset is proposed by Gao and Grauman [17]. It consists of 1,871 video clips. The train/val/test has already been split by the authors. We follow the same split and evaluation protocol of [17].

YT-MUSIC. This is also a stereophonic audio dataset that contains video recordings of music performances in 360° view. It is the most challenging dataset collected in paper [30]. As the audios are recorded in first-order ambisonics, we convert them into binaural audios using an existing converter and also follow the protocol of [17].

MUSIC. We train and evaluate the visually indicated audio source separation on the solo part of MUSIC dataset [47]. Note that in our comparing paper [46], this dataset is enriched to a version with 256 videos for testing named MUSIC21. We follow this setting and use an enriched version with 245 videos for testing, so the comparisons are basically fair. Please be noted that our whole Sep-Stereo model with separative learning is trained on this dataset.

Stereo Evaluation Metrics. We evaluate the performance of audio spatialization using similar metrics used in Mono2Binaural [17] and Ambisonics [30].

- *STFT Distance* ($STFT_D$). As all existing methods are trained in the form of STFT spectrum, it is natural to evaluate directly using the training objective on the test set.
- *Envelope Distance* (ENV_D). As for evaluations on raw audios, we use the envelope distance. It is well-known that direct comparison on raw audios is not informative enough due to the high-frequency nature of audio signals. So we follow [30] to use differences between audio envelopes as a measurement.

Separation Evaluation Metrics. We use these source separation metrics following [47]: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). The units are dB.

Table 1. Comparisons between different approaches on FAIR-Play and YT-Music dataset with the evaluation metric of STFT distance and envelop distance. The lower the score the better the results. The training data types for each method are also listed. It can be seen from the results that each component contributes to the network

	Training Data		FAIR-Play		YT-Music	
Method	Stereo	Mono	STFT_D	ENV_D	STFT_D	ENV_D
Mono2Binaural [17]	✓	✗	0.959	0.141	1.346	0.179
Baseline (MUSIC)	✓	✓	0.930	0.139	1.308	0.175
Assoicative-Conv	✓	✗	0.893	0.137	1.147	0.150
APNet	✓	✗	0.889	0.136	1.070	0.148
Sep-Stereo (Ours)	✓	✓	**0.879**	**0.135**	**1.051**	**0.145**

4.3 Evaluation Results on Stereo Generation

As the model of Mono2Binaural is also the baseline of our model, we re-produce Mono2Binaural with our preprocessing and train it carefully using the authors' released code. Besides, we perform extensive ablation studies on the effect of each component in our framework brings on the FAIR-Play and MUSIC dataset. Our modification to the original Mono2Binaural are basically the following modules:
1) Associative-Conv. While our APNet associates the visual and audio features at multiple scales, we conduct an additional experiment with Associative-Conv operating at only the outmost layer. This module aims to validate the effectiveness of the associative operation.
2) APNet. Then we perform the whole process of stereo learning with the complete version of APNet. Four layers of associative mappings are utilized to perform the coarse-to-fine tuning of the stereo learning.
3) Baseline (MUSIC). It is not possible for Mono2Binaural to use mono data in its original setting. Nevertheless, we manage to integrate our separative learning into the baseline by using our **rearrangement** module for the visual feature illustrated in Fig. 3 (c). The other parts of the network remain the same. This model can also validate the advantage of our proposed separative learning over Mono2Binaural.
4) Sep-Stereo (Ours). Finally, we add the data from MUSIC dataset for training separation. In this model, the separative learning and stereo learning are working together towards more accurate stereophonic audio generation.

The results of the experiments tested on FAIR-Play and YT-MUSIC dataset are listed in Table 1. The "Training Data" column shows whether these models are trained on MUSIC. Due to different preprocessing and sliding window sizes for testing, the results reported in paper [17] in not directly comparable. So we use our re-produced results for comparison. It can be seen that step by step adding our module can lead to better stereo results. Particularly, adding Associative-Conv can shorten the STFT distance by a large margin, which proves that this procedure can efficiently merge audio and visual data. Then improvements can be seen when expanding it to be APNet. Finally, integrating separative learning

Table 2. Source separation results on MUSIC dataset. The units are dB. †Note that DDT results are directly borrowed from the original paper [46]. It uses additional motion information, while other methods only use static input

Metric	Baseline(MUSIC)	Associative-Conv	PixelPlayer	Sep-Stereo(Ours)	DDT†
SDR	5.21	5.79	7.67	8.07	8.29
SIR	6.44	6.87	14.81	10.14	14.82
SAR	14.44	14.49	11.24	15.51	14.47

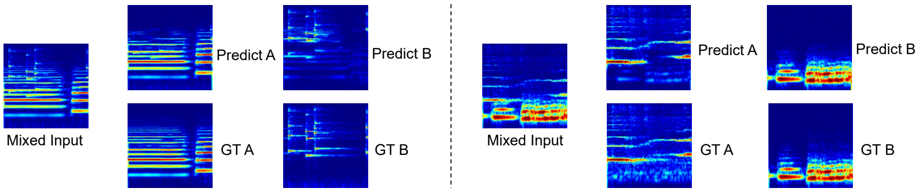


Fig. 4. The visualization of separation results from spectrums. Our results are very similar to the ground truth (GT)

into our framework gives the network more diverse data for training which leads to be the best outcome.

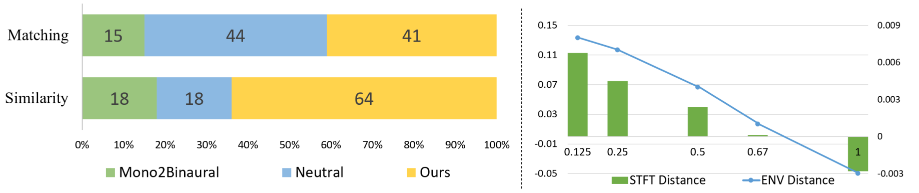
4.4 Evaluation Results on Source Separation

Our competing methods are the baseline (MUSIC) trained directly for separation, ablation of Associative-Conv, the results of self-implemented PixelPlayer [47] and DDT [46] which are originally designed for the separation task. With or without training on FAIR-Play for predicting stereo has little influence on our separation results, so we report the duet trained ones.

As shown in Table 2 that the baseline (MUSIC) and Associative-Conv model cannot achieve satisfying results. However, our Sep-Stereo can outperform PixelPlayer by two metrics and can keep competitive results with DDT. Note that the results of DDT are the reported ones from the original paper [46], thus the results are not directly comparable. The state-of-the-art DDT uses motion information which we do not leverage. Reaching such a result shows the effectiveness of APNet and the value of our model. There is no doubt that we have the potential for further improvements. We visualize two cases of separation results with duet music in Fig. 4. It can be observed that our method can mostly disentangle the two individual spectrums from the mixed one.

4.5 Further Analysis

User Study. We conduct user studies to verify the stereophonic quality of our results. We show the users a total of 10 cases from the FAIR-Play [17] dataset.



(a) User Study

(b) Our performance curve *w.r.t* training partition

Fig. 5. (a) User study for their preferences between ours and Mono2Binaural [17]. The **matching** stands for audio-visual matching quality and **similarity** represents their similarity with ground truth. The **Neutral** selection means the users cannot tell which one is better. The results are shown in percentage (%). It can be seen that ours are more preferred by users. (b) The curve of our relative performances $w.r.t$ the percentage of training data we use on FAIR-Play. The X-axis is the fraction of training data used. Axis on the left represents STFT distance and right is ENV. The lower the better. The curve is drawn in a relative setting, where the performance of Mono2Binaural serves as the reference (zero in both metrics). It can be observed that our framework can reach their full performance using only 67% of the training data

Four are results selected from Mono2Binaural's video and six are results generated by our own implementation. A total number of 15 users with normal hearing are asked to participate in the study, and a monitoring-level earphone Shure SE846 is used for conducting the study in a quiet place.

The users are asked to listen to the audio and watch the video frames at the same time. They will listen to the generated audios first and listen to the ground truth. They are responsible for telling their preferences over (1) the audio-visual **matching** quality; which of the two audios better matches the video. And (2) **similarity** to the ground truth; which of the two audios are closer to the ground truth. The users can listen to the clips multiple times. One *Neutral* option is provided if it is really difficult to tell the difference. The results show the users' preferences in Fig. 5 (a). The final results are averaged per video and per user. The table shows the ratio between selections. It can be seen that it is hard to tell the differences for certain untrained users without the ground truth. However, more people prefer our results than Mono2Binaural under both the two evaluations. The confusion is less when the ground truth is given. It can be inferred that our results are more similar to the ground truth.

Audio-Based Visual Localization. We illustrate the visually salient areas learned from our model in Fig. 6. The way is to filter intense responses in feature F_v back to the image space. It can be seen that our network focuses mostly on instruments and humans, which are undoubtedly the potential sound sources.

Generalization under Low Data Regime. We show the curve of our relative performance gains $w.r.t$ the percentage of training data used on FAIR-Play in Fig. 5 (b). The curve is drawn in a relative setting, where the performance of Mono2Binaural serves as the reference (zero in both metrics). It can be observed that our framework can reach their full performance by using only 67% of the

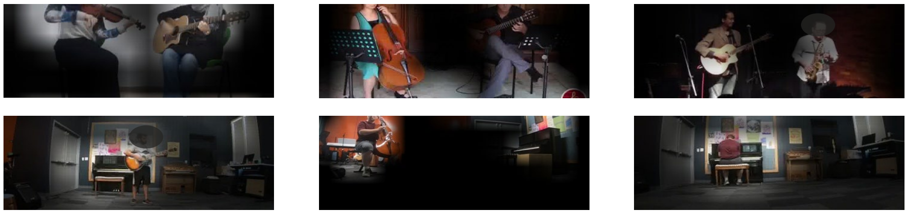


Fig. 6. The visualization of the visual responses according to audio-visual associative learning. The bright places are the salient regions in the feature map, which correspond to intense audio information. The images are selected from MUSIC and FAIR-Play

training data, which is an inherent advantage that our separative and stereophonic learning with mono data brings under low data regime [27].

We also highlight our model's ability of generalization to unseen scenarios by leveraging separative learning. Previous methods such as Mono2Binaural can only be trained with stereo data. It is difficult for them to handle out-of-distribution data if no supervision can be provided. While our method is naturally trained on mono ones, by additional training on only a small portion of stereophonic data with supervision, our method can generalize to in-the-wild mono scenarios. The video results and comparisons can be found at https://hangz-nju-cuhk.github.io/projects/Sep-Stereo.

5 Conclusion

In this work, we propose to integrate the task of stereophonic audio generation and audio source separation into a unified framework namely **Sep-Stereo**. We introduce a novel perspective of regarding separation as a particular type of stereo audio generation problem through manual manipulation on visual feature maps. We further design Associative Pyramid Network (APNet) which associates the visual features and the audio features with a learned Associative-Conv operation. Our proposed Sep-Stereo has the following appealing properties that are rarely achieved before: **1)** Rich mono audio clips can be leveraged to assist the learning of binaural audios. **2)** The task of audio separation and spatialization can be solved with a shared backbone with different heads, thus additional parameters for an entire extra network can be removed. **3)** Stereophonic generation can be generalized to low data regime with the aid of mono data. Extensive evaluation, analysis and visualization demonstrate the effectiveness of our proposed framework.

Acknowledgements. This work is supported by SenseTime Group Limited, the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14208619, CUHK14213616, CUHK14203518, and Research Impact Fund R5001-18.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. Proc. Interspeech 2018 (2018) 4
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017) 3, 4
3. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 3, 4
4. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems (NeurIPS) (2016) 3
5. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Audio-visual embodied navigation. Proceedings of the European Conference on Computer Vision (ECCV) (2020) 4
6. Chen, L., Li, Z., K Maddox, R., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 4
7. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4
8. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia (2017) 4
9. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? BMVC (2017) 4
10. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG) (2018) 4
11. Fisher III, J.W., Darrell, T., Freeman, W.T., Viola, P.A.: Learning joint statistical models for audio-visual fusion and segregation. In: Advances in neural information processing systems (NeurIPS) (2001) 4
12. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 2, 4
13. Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B.: Look, listen, and act: Towards audio-visual embodied navigation. ICRA (2020) 4
14. Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: Proceedings of the IEEE International Conference on Computer Vision (2019) 4
15. Gao, R., Chen, C., Al-Halah, Z., Schissler, C., Grauman, K.: Visualechoes: Spatial image representation learning through echolocation. Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3
16. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 4
17. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13
18. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) 4

19. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2016) 6, 9
21. Hu, D., Li, X., lu, X.: Temporal multimodal learning in audiovisual speech recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3
22. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
23. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
25. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Advances in Neural Information Processing Systems (NeurIPS) (2018) 3, 4
26. Li, D., Langlois, T.R., Zheng, C.: Scene-aware audio for 360 videos. ACM Transactions on Graphics (TOG) **37**(4), 1–12 (2018) 4
27. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 14
28. Lu, Y.D., Lee, H.Y., Tseng, H.Y., Yang, M.H.: Self-supervised audio spatialization with correspondence classifier. In: 2019 IEEE International Conference on Image Processing (ICIP) (2019) 4
29. Maganti, H.K., Gatica-Perez, D., McCowan, I.: Speech enhancement and recognition in meetings with an audio–visual sensor array. IEEE Transactions on Audio, Speech, and Language Processing **15**(8), 2257–2269 (2007) 4
30. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: Advances in Neural Information Processing Systems (NeurIPS) (2018) 2, 4, 10
31. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. European Conference on Computer Vision (ECCV) (2018) 3, 4
32. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 4
33. Parekh, S., Essid, S., Ozerov, A., Duong, N.Q., Pérez, P., Richard, G.: Motion informed audio source separation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017) 4
34. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Learning to visually localize multiple sound sources via a two-stage manner code. Proceedings of the European Conference on Computer Vision (ECCV) (2020) 4
35. Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multi-modal movie scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 3
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (MICCAI). Springer (2015) 6

37. Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., Torralba, A.: Self-supervised audio-visual co-segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019) 4
38. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
39. Son Chung, J., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3
40. Tian, Y., Li, D., Xu, C.: Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 4
41. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 4
42. Wen, Y., Raj, B., Singh, R.: Face reconstruction from voice using generative adversarial networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2019) 4
43. Wu, Y., Zhu, L., Yan, Y., Yang, Y.: Dual attention matching for audio-visual event localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) 4
44. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) 2, 4
45. Yu, J., Zhang, S., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., Yu, D.: Audio-visual recognition of overlapped speech for the lrs2 dataset. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020) 3
46. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) 2, 4, 10, 12
47. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 2, 3, 4, 5, 6, 9, 10, 12
48. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2019) 3, 4
49. Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019) 4
50. Zhou, Y., Li, D., Han, X., Kalogerakis, E., Shechtman, E., Echevarria, J.: Makeittalk: Speaker-aware talking head animation. arXiv preprint arXiv:2004.12992 (2020) 4
51. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
52. Zhu, H., Huang, H., Li, Y., Zheng, A., He, R.: Arbitrary talking face generation via attentional audio-visual coherence learning. In: International Joint Conference on Artificial Intelligence (IJCAI) (2020) 4
53. Zhu, H., Luo, M., Wang, R., Zheng, A., He, R.: Deep audio-visual learning: A survey. arXiv preprint arXiv:2001.04758 (2020) 3