Visually Guided Sound Source Separation and Localization using Self-Supervised Motion Representations

Lingyu Zhu
Tampere University, Finland
lingyu.zhu@tuni.fi

Esa Rahtu
Tampere University, Finland
esa.rahtu@tuni.fi

Abstract

In this paper, we perform audio-visual sound source separation, i.e. to separate component audios from a mixture based on the videos of sound sources. Moreover, we aim to pinpoint the source location in the input video sequence. Recent works have shown impressive audio-visual separation results when using prior knowledge of the source type (e.g. human playing instrument) and pre-trained motion detectors (e.g. keypoints or optical flows). However, at the same time, the models are limited to a certain application domain. In this paper, we address these limitations and make the following contributions: i) we propose a two-stage architecture, called Appearance and Motion network (AMnet), where the stages specialise to appearance and motion cues, respectively. The entire system is trained in a self-supervised manner; ii) we introduce an Audio-Motion Embedding (AME) framework to explicitly represent the motions that related to sound; iii) we propose an audio-motion transformer architecture for audio and motion feature fusion; iv) we demonstrate state-of-the-art performance on two challenging datasets (MUSIC-21 and AVE) despite the fact that we do not use any pre-trained keypoint detectors or optical flow estimators. Project page: https://ly-zhu.github.io/self-supervised-motion-representations

1. Introduction

Sound source separation is a classical task of extracting a target sound source from a given audio mixture [21, 36, 47, 10]. A well-known example is so called cocktail party problem, where one attempts to listen to a person while multiple people are speaking in the same space. Similarly, one might be interested in extracting the sound of a single instrument from a concert recording or other signals from the background noise. Despite being extensively studied, the audio-based source separation remains a challenging problem.

Recent works [33, 22, 13, 18, 34, 54, 53, 51, 19, 59, 60] have shown that visual observations of the sound source (e.g. speaking face) may substantially simplify the separation task. For instance, the lip movements can be applied to extract the desired speech signal in the cocktail party problem [33, 22, 13]. Similarly, other visual cues like object categories or motions can be exploited to guide the separation process [18, 34, 54, 53, 51, 19, 59, 60, 61]. This kind of problem setup is often referred as visual sound source separation or visually guided sound source separation.

Figure 1. We introduce a new sound-related motion representation for localizing and separating sound sources. When combined with the appearance cues, we obtain a highly effective model for the visual sound separation.

A popular approach in visual sound source separation is to encode both visual input and mixture audio into feature representations, and then fuse them to decode the component signal corresponding to the visual content. The visual feature encoding is one of the key elements in the approach, and previous works have studied different options for this part. For instance, several works utilise frame-based appearance features [54, 18, 19, 51, 60], which contain information about the object categories (e.g. instrument types) but not from the object motion. While the appearance alone can be a strong cue, the motion may often be the only reliable cue (e.g. lip motion while speaking).

Several recent works have proposed approaches to include motion information in the separation process. These models utilise optical flow [53], dynamic image [59], and keypoint-based human body dynamics [15]. While optical flow and dynamic images are good in representing the over-

all motion (e.g. for action recognition), they do not provide significant gain over the pure appearance-based methods in source separation. The recent work by Gan *et al.* [15] proposed to use human keypoints to encode the motion cues. Their main motivation was to explicitly model body and finger movements of musicians when they perform music. The results demonstrated impressive performance over the prior state-of-the-art, but unfortunately this approach is limited to sounds resulting from human motion. Moreover, the work relies on pre-trained human keypoint detectors. Owens *et al.* [34] presented an alternative approach for learning the visual representation by classifying artificially misaligned audio and visual streams. Their idea is interesting as the misalignment naturally encourages to focus on the correlation between motion and audio.

In this paper, we introduce a two-stage visual sound source separation architecture, called Appearance and Motion network (AMnet), where the stages specialise to appearance and motion cues, respectively. We propose an Audio-Motion Embedding (AME) framework to learn the motions of sounds in a self-supervised manner. Furthermore, we design a new Audio-Motion Transformer (AMT) module to facilitate the fusion of audio and motion cues.

We demonstrate the performance of the proposed AMnet with two challenging datasets, *MUSIC-21* and *AVE*, and obtain the state-of-the-art results. Interestingly, AMnet outperforms the keypoint-based approach [15] without any prior knowledge or limitations to human induced motion or pre-extracted keypoints. Moreover, we apply the learned motion cues from the AME to pinpoint the sound source location in the input video stream (see Figure 1).

In summary, our contributions are: i) AMnet for self-supervised visual sound separation. The approach makes no pre-assumptions on the sound source type (e.g. human induced motion); ii) self-supervised motion representation learned by mapping the audio and motions into a common embedding space; iii) audio-motion transformer module for audio and motion feature fusion; iv) state-of-the-art results in two challenging visual sound separation datasets.

2. Related Work

Audio-Visual Learning In recent years, leveraging constraints between different modalities, such as audio and vision, has been applied in various cross-model learning tasks. Aytar *et al.* [4] learned joint audio-visual embeddings to minimize the KL-divergence of their representations. Arandjelovic *et al.* [2, 3] associated the audio with vision by modeling their correspondence. More recently, researchers have demonstrated the works of audio-visual scene understanding [1, 38, 58, 45], audio-visual synchronization [34, 30, 9], talking face generation [56], audio-driven 3D facial animation [11], audio-visual navigation [16], and visual-to-auditory [26, 14, 57].

Visually Guided Sound Separation Early works of sound separation were mainly based on probabilistic methods [21, 36, 47, 10], while recent approaches utilise deep learning architectures [24, 7]. Despite of the substantial improvements, general form of the problem is challenging and highly underdetermined. Visual sound separation is gaining increasing attentions recently. Ephrat *et al.* [13] extracted face embeddings for speech separation. Similarly, Gao *et al.* [19, 18] applied object detection to facilitate source separation. Zhao *et al.* [54] proposed to separate sounds by a linear combination of semantic cues and sound spectrogram features. A subsequent work [53] introduced trajectory optical flows to the sound separation. Xu *et al.* [51] separated sounds by recursively removing large energy components from sound mixture. Zhu *et al.* [61] utilized slow and fast spectrograms to separate sounds. Gan *et al.* [15] associated body and finger movements with audio signals by learning a keypoint-based structured representation. While impressive, these methods either have limited capabilities to capture the motion cues or rely on prior knowledge (e.g. object detection, optical flows, or keypoints).

The works by Owens *et al.* [34] and Zhu *et al.* [59] are most related to ours. [34] presented a classification-based audio-visual misalignment model to analyse multisensory features for sound separation. In [59], the authors utilized visual features of all the sources to look for incorrectly assigned sound components between sources in a multi-stage manner. Our work learns new motion cues by mapping the audio and motions into a common embedding space, and separates sounds with the proposed AMnet, which specialises to appearance and the motion cues.

Motion Representations of Video Sequence Early works of video representations were largely based on hand-crafted spatio-temporal features [31, 29, 48, 49]. These have been recently shifted to deep neural networks, which can be roughly grouped into following categories: i) 2D CNN with summarized motions from dynamic images [6]; ii) 3D CNN on spatio-temporal video volume [44]; iii) two-stream CNNs [41], where motions are modeled from prior computed optical flows; iv) LSTM [12], Graph CNN [50] and attention clusters [32] based techniques. These methods are proposed mainly for action recognition problem. In contrast, our goal is to model the visual motions of sound and further to facilitate the sound separation task.

Sound Source Localization Visually identifying sound source location is another challenging task. Hershey *et al.* [25] utilized non-stationary Gaussian process to model audio-visual synchrony for locating sound sources. The subsequent work brought the ideas of canonical correlations [28] and temporal coincidences [5]. More recent works, including semantics [3, 40, 54, 51, 60, 8], tra-

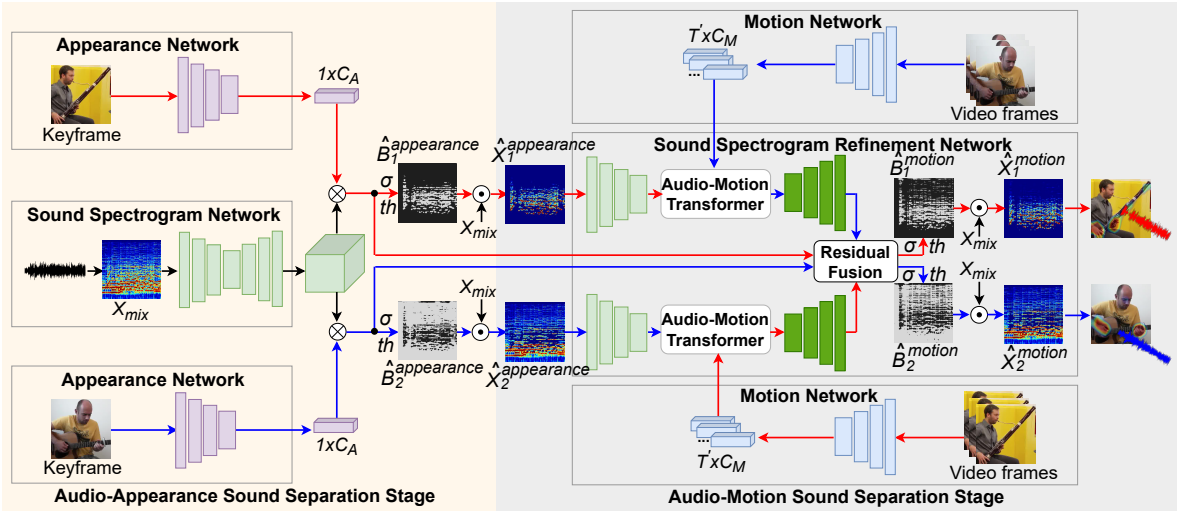

Figure 2. The overall architecture of the proposed Appearance and Motion network (AMnet). The Audio-Appearance stage encodes a video keyframe and a mixture audio spectrogram into an appearance feature vector and a spectrum feature volume, respectively. These are subsequently fused by calculating a weighted sum of spectrum features using appearance features as weights. The result is further converted to a binary mask and multiplied with the input spectrogram to produce the separated output. The Audio-Motion stage again encodes the spectrograms and the video sequence into feature representations. These are fused with the proposed Audio-Motion Transformer module, decoded, and passed to produce a refined mask, which is multiplied with the input mixture spectrogram to produce the final output.

jectory optical flows [53], misalignment [34, 30], location masking [59], spatial audios [17, 52, 16], and attention [3, 35, 9, 1, 60] based methods.

3. Approach

3.1. Overview

The input to the proposed system consists of a mixture audio and video sequences depicting the sound sources. The objective is to extract the component audio that corresponds to the sound source in the given video. Figure 2 illustrates the overall architecture of the proposed Appearance and Motion network (AMnet). The first part, called Audio-Appearance stage, performs source separation using pure appearance-based features (e.g. object types). To this end, we first randomly extract a frame from the video sequence and encode it into a visual feature vector of dimension C_A. The input mixture audio is converted to a $H_S \times W_S$ spectrogram image and then encoded into a feature volume of size $C_S \times H_S \times W_S$. The obtained feature volume is multiplied channel-wise with the visual feature vector and converted to a $H_S \times W_S$ binary mask. The output of the Audio-Appearance stage is formed by multiplying the mixture spectrogram with the binary mask.

The obtained result is further passed to the Audio-Motion stage (see Figure 2), which refines the source separation using motion cues. The corresponding features are extracted by the Motion Network, which encodes the video sequence into a motion representation of size $T^{'} \times C_M$, where $T^{'}$ corresponds to the time dimension of the sequence. Here we do not make any pre-assumptions of the motion type (e.g. human body motions). The subsequent Sound Spectrogram Refinement network (SSR) combines the motion features with the output spectrogram from the previous stage. The core part of the refinement network is the Audio-Motion Transformer (AMT) module that associates motion representations with the spectrogram features using a multimodal transformer architecture. The final output is formed by multiplying the original mixture spectrogram with a binary mask obtained from the refinement network. In addition, the Motion Network provides an estimate of the sound source location in the video (see Figure 5).

The Motion Network is trained using the new Audio-Motion Embedding (AME) framework. In AME, we learn mappings from both audio and video streams into a common embedding space (see Figure 3), where distance would be correlated with temporal alignment of the input sequences. The following sections provide further details of the system parts and the training procedure. We start by describing the AME framework and related learning objectives. The detailed network architectures are provided in the supplementary material.

3.2. Audio-Motion Embedding Framework

The proposed Audio-Motion Embedding (AME) framework exploits the natural correlation between the audio and motion of a natural video. The AME framework (see Figure 3) consists of Motion Network and Sound Network, which map the motion and audio sequences into a common embedding space, respectively. We formulate the learning

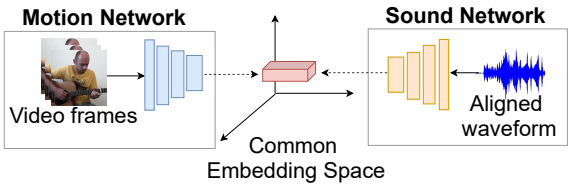

Figure 3. An illustration of the proposed Audio-Motion Embedding (AME) framework.

objectives to enforce small embedding distances between well synchronised streams and large distances for out-of-sync streams. We hypothesize that this learning objective encourages the Motion Network to focus mainly on the sound related motion features in the input video.

The AME maps the motions and audio into the common embedding space. The mappings are learned in the following manner: i) given a video clip v with an aligned audio $x_{aligned}$, we generate a misaligned audio $x_{misaligned}$ by randomly shifting the waveform in time domain; ii) we encode the video stream, aligned audio, and the misaligned audio into embedding representations using the motion and sound networks, respectively; iii) we calculate the distances between the video embedding and the both audio embeddings, and formulate a cost function using a triplet loss approach [39]; and iv) we optimise the embedding networks by minimizing the loss function over a large set of videos.

Previous works using audio-visual synchronization for self-supervised representation learning formulate the problem as a classification task where the system decides if the given audio and video are synchronized (1) or not (0) [34, 30, 9]. Moreover, the information along the temporal dimension is mostly neglected by marginalizing the corresponding dimension with pooling operation. In contrast, we formulate the problem as a mapping from motion and audio domains to a common embedding space, where the distances correlate with the temporal alignment. Moreover, the representation retains the temporal dimension at the output features. Following paragraphs outline details of the embedding networks and the learning objective.

Motion Network The Motion Network E_M maps the input video frames v into a vector space. We forward the input video sequence v of size $3 \times T \times H \times W$ to a 3D version of Res18 [23] and produce a representation f_{M1} of size $C_M \times T^{'} \times H^{'} \times W^{'}$, where $T^{'}$=$T/4$, $H^{'}$=$H/16$, and $W^{'}$=$W/16$. With an additional 3D convolution, we obtain a single channel feature map f_{M2} of size $1 \times T^{'} \times H^{'} \times W^{'}$. We obtain the final embedding vector f_{M3} of size $1 \times T^{'}$ by applying a spatial average pooling.

Sound Network The Sound Network E_S maps the audio waveform into a common embedding space with the Motion Network E_M. Here we use Res18-1D architecture, which consists of a series of strided 1D convolutions, applied until the size of the output representation matches to the Motion Network output f_{M3}, i.e. $1 \times T^{'}$.

Learning Objective We utilize the natural audio and motion temporal alignment to train the AME. Given a video clip, aligned audio, and misaligned audio $\{v,\ x_{aligned},\ x_{misaligned}\}$, and their corresponding embeddings $E_M(v)$, $E_S(x_{aligned})$, and $E_S(x_{misaligned})$, we define a triplet loss function as follows:

$$\begin{aligned}\mathcal{L}_{AME} = \max\Big(&d\big(E_M(v), E_S(x_{aligned})\big) \\ -\ &d\big(E_M(v), E_S(x_{misaligned})\big) + margin, 0\Big)\end{aligned} \tag{1}$$

where d measures the similarity between the motion and audio embeddings. We use $margin = 2.0$ in all experiments. The embedding networks are optimized with respect to the loss function over a large set of input triplets.

3.3. Audio-Appearance Sound Source Separation

The Audio-Appearance stage aims to perform source separation using object appearances. The stage consists of appearance network, sound spectrogram network, and sound source separation module.

Appearance Network The Appearance Network receives a random single frame from the input video and applies a dilated Res18-2D [23] to obtain a compact semantic representation. More specifically, given an input RGB image, the Appearance Network produces a representation f_A of size $1 \times C_A$ as the output of the last spatial average pooling layer.

Sound Spectrogram Network The input audio waveform is first converted to a spectrogram presentation X_{mix} using Short-time Fourier Transform (STFT). The Sound Spectrogram Network encodes X_{mix} into a set of feature maps. The network is implemented using MobileNetV2 (MV2) [37] architecture and it converts the input spectrogram of size $1 \times H_S \times W_S$ to a feature map f_{mix} of size $C_S \times H_S \times W_S$. Note that the number of produced feature maps C_S is equal to the appearance feature vector dimension C_A in the previous section.

Sound Source Separation The sound source separation module utilises the feature maps f_A and f_{mix} to produce an estimate of the component audio corresponding to the input video. More specifically,

$$\begin{aligned} f_{S,n}^{appearance} &= f_{A,n} \otimes f_{mix}, \\ \hat{B}_n^{appearance} &= th(\sigma(f_{S,n}^{appearance})), \\ \hat{X}_{S,n}^{appearance} &= \hat{B}_n^{appearance} \odot X_{mix}, \end{aligned} \tag{2}$$

where $\otimes$ and $\odot$ denote the channel-wise and element-wise product, respectively. σ represents the sigmoid operation. $th(x) = 1$ if $x > 0.5$ and 0 otherwise. $f_{A,n}$ is the appearance network output for the n-th source (n-th input video). The output spectrogram $\hat{X}^{appearance}_{S,n}$ is formulated by element-wise multiplying the binary mask $\hat{B}^{appearance}_{n}$ with the original mixture spectrogram X_{mix}.

3.4. Audio-Motion Sound Source Separation

The Audio-Motion stage utilises the motion cues for the sound source separation. The stage contains four components: motion network (see Section 3.2), sound spectrogram refinement (SSR) network, audio-motion transformer (AMT), and residual fusion module.

Sound Spectrogram Refinement Network The Sound Spectrogram Refinement (SSR) network is an encoder-decoder architecture, which consists of 7 down- and 7 up- convolutional layers followed by a BatchNorm layer and Leaky ReLU. The encoder (SSR_E) takes the spectrogram $\hat{X}^{appearance}_{S,n}$ from the Audio-Appearance stage and produces a feature representation $f^{motion,encoder}_{S,n}$ (the superscript *motion* refers to the Audio-Motion stage). The encoder is followed by the Audio-Motion Transformer (AMT), which fuses $f^{motion,encoder}_{S,n}$ with the motion features f_M (f_{M1} following with a spatial pooling, which is produced by the Motion Network of AME in Section 3.2). The transformed output $f^{AMT}_{S,n->m}$ from the AMT is passed to the up-convolutional decoder (SSR_D) to produce residual spectrum representation of size $H_S \times W_S$. We adopt similar opponent filter approach as in [59] to relocate the identified residual sound components from Audio-Appearance outputs to final corresponding outputs. The procedure of the SSR network is defined as Eq. 3 and 4.

$$\begin{aligned} f^{motion,encoder}_{S,n} &= SSR_E(\hat{X}^{appearance}_{S,n}), \\ f^{AMT}_{S,n->m} &= AMT(f_{M,m}, f^{motion,encoder}_{S,n}), \\ f^{motion,decoder}_{S,n->m} &= SSR_D(f^{AMT}_{S,n->m}), \end{aligned} \tag{3}$$

Audio-Motion Transformer The Audio-Motion Transformer (AMT) module is used to leverage the obtained motion cues (the motions of sound learned with AME) for sound source separation. The overall architecture follows [46, 27] and is visualised in Figure 4. First the input embeddings are positional encoded [46] to preserve ordering of the sequence. The following encoder and decoder modules are composed of stacked multi-head attention, point-wise, and fully connected layers.

The encoder applies $f_{M,m}$ of m-th source as query Q_M, key K_M, and value V_M inputs. This type of multi-head attention which has identical Q, K, and V is often referred as

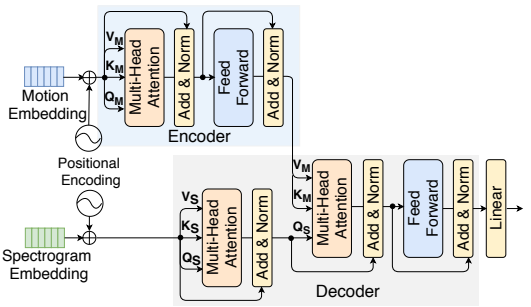

Figure 4. The architecture of the Audio-Motion Transformer.

self-multi-head attention. Similarly, at the very beginning of the decoder, the AMT has a self-multi-head attention (Q_S, K_S, and V_S) of n-th sound features $f^{motion,encoder}_{S,n}$. After self-attention, the decoder infuses the motions into the sound components by an audio-motion multi-head attention, which considers the sound embedding $f^{AMT}_{S,n}$ of n-th source as the query Q_S, motions embeddings $f^{AMT}_{M,m}$ of m-th source as the key K_M and value V_M, where the $m, n \in (1, \dots N), m \neq n$, and N is the number of sources.

Residual Fusion Module The SSR_D output $f^{motion,decoder}_{S,n->m}$ (Eq. 3) is interpreted as a residual spectrum. That is, the spectrum indicates the parts of the Audio-Appearance outputs (Eq. 2), which need to be reallocated. For instance $f^{motion,decoder}_{S,n->m}$ defines the component that should belong to source m but is currently assigned to source n. Similar to [59], we relocate these components between the spectrograms, as follows

$$\begin{aligned} f_{S,n} &= f^{appearance}_{S,n} \ominus f^{motion,decoder}_{S,n->m}, \\ f_{S,m} &= f^{appearance}_{S,m} \oplus f^{motion,decoder}_{S,n->m}, \\ \hat{B}^{motion}_{n} &= th(\sigma(f_{S,n})), \\ \hat{X}^{motion}_{S,n} &= \hat{B}^{motion}_{n} \odot X_{mix} \end{aligned} \tag{4}$$

where the $f^{appearance}_{S,n}$ is the output spectrum of Audio-Appearance stage for n-th source, $f^{motion,decoder}_{S,n->m}$ is the residual spectrum from sound n to sound m. $\oplus$ and $\ominus$ denote the element-wise sum and subtraction, respectively. The obtained feature maps $f_{S,n}$ are passed through sigmoid and thresholding operations, and the final spectrogram output $\hat{X}^{motion}_{S,n}$ is formed by multiplying the original mixture spectrogram with the obtained result. An inverse STFT is applied to produce the final separated audio waveforms.

3.5. Overall learning Objective

We formulate the overall learning objective in terms of the binary masks $\hat{B}_n$, which are used to obtain the final output spectrograms (Eq. 2 and 4). The ground truth masks B_n are formed as follows,

$$B_n(t, f) = [X_n(t, f) \geq X_m(t, f)] \tag{5}$$

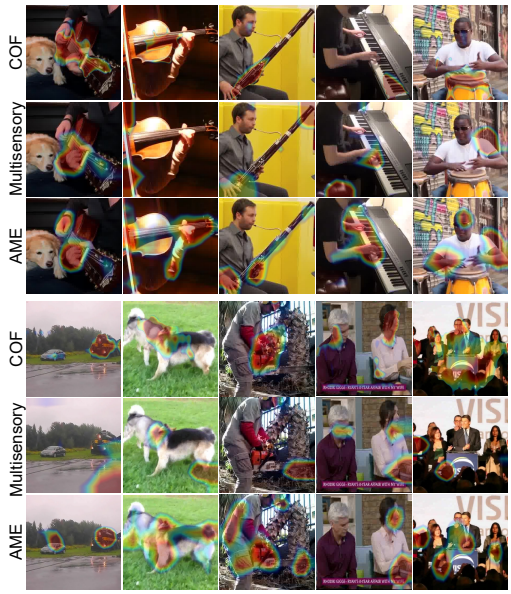

Figure 5. Visualization of the CAM responses for MUSIC-21 (top) and AVE (bottom) with our AME, COF [59], and Multisensory [34].

where $\forall m = (1, \ldots, N)$, (t, f) represents the time-frequency coordinates in the sound spectrogram X. The AMnet is trained by minimizing the binary cross entropy (BCE) loss between the estimated binary masks $\hat{B}_n$ and the ground-truth binary masks B_n,

$$\mathcal{L} = \sum_{n=1}^{N} BCE(\hat{B}_n^{appearance}, B_n) + BCE(\hat{B}_n^{motion}, B_n) \tag{6}$$

where $\hat{B}_n^{appearance}$ and $\hat{B}_n^{motion}$ represent the predicted binary masks at Audio-Appearance and Audio-Motion stages.

4. Experiments

In this section, we start by assessing the Audio-Motion Embedding framework and continue with source separation experiments. The results are compared with the current state-of-the-art baseline methods.

4.1. Datasets and evaluation metrics

We use MUSIC-21 [53] and Audio-Visual Event (AVE) [20, 43] datasets in our experiments. MUSIC-21 [53] contains 1365 videos from 21 instrumental categories. We extract video frames at 8 fps and sub-sample audio streams at 11kHz. The AVE [43] dataset, a subset of AudioSet [20], contains 4143 10-second videos covering 28 event categories. The dataset covers a wide range of audio-visual events from different domains, e.g. human activities and vehicle sounds. For AVE, we use the full frame-rate (29.97 fps) and sub-sample audio signal at 22kHz. The sound separation performance is measured in terms of: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR). For measures, higher value indicates better performance.

Models \ Metrics	MUSIC-21		AVE		UCF-101
	cIoU	AUC	cIoU	AUC	Acc
Multisensory [34]	44.62	45.99	23.88	28.64	64.29%
AME (ours)	67.18	54.58	25.62	29.81	71.69%

Table 1. Quantitative localization results on MUSIC-21 and AVE datasets, and action recognition results on UCF-101 (split 1).

Models	SDR	SIR	SAR
Multisensory [34]	3.18	11.42	6.68
Sound of Pixels [54]⋆	7.52	13.01	11.53
Co-Separation [19]⋆	7.64	13.8	11.3
Sound of Motions [53]⋆	8.31	14.82	13.11
Minus-Plus [51]	9.15	15.38	12.11
Cascaded Opponent Filter [59]	9.80	17.16	12.33
Music Gesture [15]⋆	10.12	15.81	-
AMnet (ours)	**11.08**	**18.00**	**13.22**

Table 2. The source separation performance using mixtures of two sources from the MUSIC-21 dataset. The results indicated with ⋆ are obtained from [15].

4.2. Implementation details

The Audio-Motion Embedding (AME) framework is trained with MUSIC-21 dataset. We adopt random scaling, horizontal flipping, and cropping (224×224) as the frame augmentation. A stream of $T = 48$ frames is forwarded to the Motion Network E_M. We randomly crop 6-second audio clip and randomly shift the audio forward or backward by 1 to 7 seconds. In the source separation experiments, we follow the same setup as in [15]. Both datasets are split into disjoint train, val (AVE), and test sets. The audio mixture is obtained by adding the audio tracks of N videos. For MUSIC-21, we use the same split as in [15]. The input audio is converted to spectrogram using STFT with a hanning window of size 1022 and a hop lengths of 256 and 184 for MUSIC-21 and AVE datasets, respectively.

4.3. Audio-Motion Embedding Framework

Although our main objective is to perform audio-visual source separation, we assess the AME based motion cues in three different motion related tasks: i) sound source localization; ii) action recognition; and iii) audio-visual sound source separation (in sec. 4.4). The motivation behind the additional comparisons is to provide wider picture of our model with respect to similar frameworks such as [34]. The evaluation details are provided in supplementary material.

Sound Source Localization For examining the ability of pinpointing the source location in the input stream, we mea-

Models	SDR	SIR	SAR
Multisensory [34]	0.84	3.44	6.69
Sound of Pixels [54]	1.21	7.08	6.84
Sound of Motions [53]	1.48	7.41	7.39
Minus-Plus [51]	1.96	7.95	8.08
Cascaded Opponent Filter [59]	2.68	8.18	8.48
AMnet (ours)	**3.71**	**9.15**	**11.00**

Table 3. The source separation performance for mixtures of two sources from the AVE dataset.

Models	SDR	SIR	SAR
Multisensory [34]	-1.92	2.13	5.71
Sound of Pixels [54]	2.31	9.34	5.77
Sound of Motions [53]	2.77	10.20	5.81
Minus-Plus [51]	3.36	9.22	7.15
Cascaded Opponent Filter [59]	4.08	9.95	7.68
AMnet (ours)	**4.82**	**11.75**	**7.77**

Table 4. The source separation performance with mixtures of three sources from the MUSIC-21 dataset.

sure the consensus Intersection over Union (cIoU) and Area Under Curve (AUC) [40] in Table 1 and visualize the source locations in Figure 5. Figure 5 contains example results obtained using the Class Activation Map (CAM) [55]. More examples are provided in the supplementary material.

The results in Table 1 and Figure 5 indicate that the sound related motion is nicely captured by the AME. For instance, the AME accurately captures both hands of the guitar player (first column in top Figure 5). When playing a congas (fifth column in top Figure 5), AME highlights hand and arm motions which are highly correlated with the output sound. The AVE dataset is more challenging due to the large scale of natural sounds it contains. Compared to the baseline models, our method can localize motions in various categories of videos. For example, our method detects the moving vehicles (first column in bottom Figure 5), both of the human body and the chainsaw motions (third column in bottom Figure 5) as these motions occur always with the sounds. Moreover, as shown in the fourth column (bottom Figure 5), the woman on the right side is speaking while the man on the left is not. Even though they have similar appearance, the AME module localizes the motions precisely, whereas both COF [59] and Multisensory [34] localize the head region of both the two persons. These results indicate that the AME mechanism facilitates the network to capture discriminative motions that correlated to the sound sources in the scenes instead of semantics.

Action Recognition We further evaluate the performance of the AME motion cues in comparison with Multisensory [34] for recognition tasks in Table 1 (last column). To study this, we fine-tuned the methods (motion only) for action recognition using the UCF-101 dataset [42].

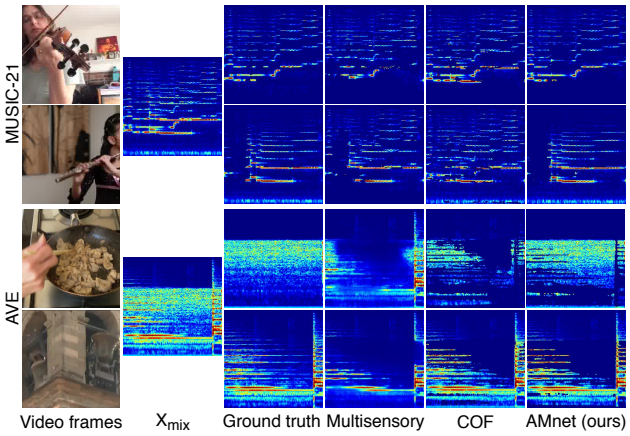

Figure 6. Visualization of the source separation results with the MUSIC-21 and AVE datasets.

4.4. Audio-Visual Sound Source Separation

Separating Two Sound Sources Tables 2 and 3 summarise the results for separating two sound sources using the proposed AMnet and recent baseline works [34, 54, 53, 19, 51, 15, 59]. The AMnet outperforms the baseline methods consistently and for most cases with a large margin. Compared to the closely related Multisensory [34] and COF [59] works, we obtain 7.90dB and 1.28dB improvement on MUSIC-21 and 2.87dB and 1.03dB improvement on AVE, respectively. Figure 6 contains qualitative examples, which clearly illustrate the differences. Additional examples are provided in the supplementary material.

Interestingly, the proposed AMnet also outperforms the keypoint-assisted Music Gesture [15] model, which is particularly designed to exploit human body and finger dynamics. This result further indicates that the introduced AME model is able to capture fine-grained motion information from the sequences. Moreover, the transformer-based fusion with the motion features seems to provide a strong combination for the sound separation.

Separating Three Sound Sources The separation task turns more difficult when the mixture contains more sources. To this end, we assess the methods by separating mixtures of three sources created from the MUSIC-21[1]. Table 4 contains the results for AMnet and the baselines [34, 54, 53, 51, 59]. We can observe a clear drop in all performances compared to the two sources case (Table 2). However, AMnet still outperforms the baselines with a clear margin.

Separating Sources of the Same Type In addition to the number of sources, the source type affects the performance.

[1]Due to limited information, we were not able to exactly reproduce the results reported in [53, 15] for three sources and same instrument mixtures.

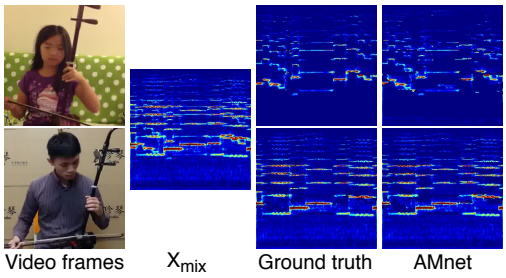

Figure 7. Illustration of the result for separating sound sources of the same type from MUSIC-21 dataset.

Models	cello	erhu	guzheng	pipa	xylophone
Copy-Paste	0.55	0.12	0.29	0.24	0.30
SoP	-0.44	-0.37	-0.02	-0.95	0.34
SoM	-4.04	-0.92	0.62	-2.34	0.25
Multisensory	-1.45	1.34	-0.65	-2.86	-2.12
COF	-3.47	2.33	-0.75	1.40	1.07
AMnet (ours)	**2.47**	**3.91**	**1.30**	**2.28**	**1.57**

Table 5. The source separation performance using mixtures of same instrument types from MUSIC-21 in terms of SDR.

Particularly difficult case occurs when the sources are of the same type (e.g. two cellos) and the appearance cues do not provide useful information. To this end, we pick video pairs with the same instrument type from MUSIC-21 and mix them for testing. We compare AMnet with Sound of Pixels (SoP) [54], Sound of Motions (SoM) [53], Multisensory [34], and COF [59]. The SoP is a pure appearance-based method, while the others utilise motion information in different ways. To gain further insight, we include a trivial "Copy-Paste" baseline that simply copies the input mixture spectrogram to the output.

Table 5 shows the results in terms of SDR for cello, erhu, guzheng, pipa, and xylophone categories. Interestingly, the Copy-Paste baseline outperforms comparison methods in most cases, whereas the proposed AMnet provides a clear improvement for all categories. The results indicate the challenges related to optical flow or dynamic image based motion modelling in self-supervised source separation. The classification-based multisensory features in [34] suffer similar challenges. The motions of sound that learned through the audio-motion common embedding space seem to be the best option for capturing the motion cues in a self-supervised manner.

Ablation Study We conducted ablation study to investigate the performance of the Audio-Motion Transformer (AMT) and the role of the appearance and motions in our model. For this purpose, we repeat the source separation experiments with MUSIC-21 using different configurations.

The AMT fuses the audio and motion for sound separation. To verify its efficacy, with a 1-stage model of using only motions to separate sounds, we replace the AMT module with the Early Fusion (EF) used in Multisensory [34]. The comparison results are shown in Table 6. Our proposed AMT module brings 1.11dB improvement in SDR.

stages	Models	SDR	SIR	SAR
1	Motion (EF)	3.71	10.61	8.22
1	Motion (AMT)	4.82	12.04	8.46

Table 6. The ablation results of using EF [34] and AMT.

stages	Models	SDR	SIR	SAR
1	Motion	4.82	12.04	8.46
1	Appearance	7.40	12.88	11.00
1	Appearance-Motion	8.99	16.38	11.32
2	Motion+Appearance	9.63	16.75	11.98
2	Appearance+Motion (AMnet)	11.08	18.00	13.22

Table 7. The ablation results comparing the appearance and motion stages, and their combination. "Appearance-Motion" represents a 1-stage model which fuses (by concatenation) the appearance and motions and continues to separate sound sources. "Motion+Appearance" and "Appearance+Motion" indicate 2-stage models, which includes appearance and motions at Audio-Appearance and Audio-Motion stage.

Table 7 shows the results using only Motion, only Appearance, Appearance-Motion, Motion+Appearance, and Appearance+Motion. The appearance results in better performance compared to using motion alone. Their combination in 1-stage improves the performance by 1.59dB in SDR. The 2-stage models of "Motion+Appearance" and "Appearance+Motion" achieve better results in comparison with their 1-stage counterparts by a large margin. However, the full model AMnet ("Appearance+Motion") provides clearly the best results suggesting that appearance and motion cues contain highly complementary information. Appearance and motion cues are different in nature and the AMnet lets each stage specialize in one type.

5. Conclusions

In this paper, we show that the motion representations that learned through the Audio-Motion Embedding (AME) framework, together with the Appearance and Motion network (AMnet), obtain the new state-of-the-art results on visually guided sound source separation. The proposed AME approach results in better sound source localization and action recognition in comparison to baselines. The AMnet contains an Audio-Appearance and Audio-Motion stage, which specialise to appearance and motion cues, respectively. Our method, trained in a self-supervised manner, has no limitation on source types and outperforms the methods specifically designed to utilise human body motions.

Acknowledgement This work is supported by the Academy of Finland (projects 327910 & 324346).

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020.

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

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

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016.

[5] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[6] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3042, 2016.

[7] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International conference on latent variable analysis and signal separation*, pages 258–266. Springer, 2017.

[8] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[9] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. *arXiv preprint arXiv:2008.05789*, 2020.

[10] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shunichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[11] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019.

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[13] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[14] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. *arXiv preprint arXiv:2007.10984*, 2020.

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

[16] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020.

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

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

[19] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3879–3888, 2019.

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[21] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. In *Advances in Neural Information Processing Systems*, pages 472–478, 1996.

[22] Elana Zion Golumbic, Gregory B Cogan, Charles E Schroeder, and David Poeppel. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, 33(4):1417–1426, 2013.

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

[24] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[25] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*, pages 813–819, 2000.

[26] Di Hu, Dong Wang, Xuelong Li, Feiping Nie, and Qi Wang. Listen to the image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7972–7981, 2019.

[27] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition Workshops, pages 958–959, 2020.

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

[29] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008.

[30] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018.

[31] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.

[32] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7834–7843, 2018.

[33] Wei Ji Ma, Xiang Zhou, Lars A Ross, John J Foxe, and Lucas C Parra. Lip-reading aids word recognition most in moderate noise: a bayesian explanation using high-dimensional feature space. *PloS one*, 4(3):e4638, 2009.

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

[35] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2970–2979, 2020.

[36] Sam T Roweis. One microphone source separation. In *Advances in neural information processing systems*, pages 793–799, 2001.

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[38] Valentina Sanguineti, Pietro Morerio, Niccolo Pozzetti, Danilo Greco, Marco Cristani, and Vittorio Murino. Leveraging acoustic images for effective self-supervised audio representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 119–135. Springer, 2020.

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

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

[41] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012.

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

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[45] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11612–11621, 2021.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[47] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[48] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR 2011*, pages 3169–3176. IEEE, 2011.

[49] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[50] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[51] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–891, 2019.

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

[53] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744, 2019.

[54] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018.

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[56] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019.

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

[58] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021.

[59] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation using cascaded opponent filter network. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[60] Lingyu Zhu and Esa Rahtu. Leveraging category information for single-frame visual sound source separation. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2021.

[61] Lingyu Zhu and Esa Rahtu. V-slowfast network for efficient visual sound separation. *arXiv preprint arXiv:2109.08867*, 2021.