Visually Guided Sound Source Separation using Cascaded Opponent Filter Network

Lingyu Zhu[0000−0001−6707−6665] and Esa Rahtu[0000−0001−8767−0864]

Tampere University, Tampere, Finland
lingyu.zhu@tuni.fi, esa.rahtu@tuni.fi

Abstract. The objective of this paper is to recover the original component signals from a mixture audio with the aid of visual cues of the sound sources. Such task is usually referred as visually guided sound source separation. The proposed *Cascaded Opponent Filter* (COF) framework consists of multiple stages, which recursively refine the source separation. A key element in COF is a novel opponent filter module that identifies and relocates residual components between sources. The system is guided by the appearance and motion of the source, and, for this purpose, we study different representations based on video frames, optical flows, dynamic images, and their combinations. Finally, we propose a *Sound Source Location Masking* (SSLM) technique, which, together with COF, produces a pixel level mask of the source location. The entire system is trained in an end-to-end manner using a large set of unlabelled videos. We compare COF with recent baselines and obtain the state-of-the-art performance in three challenging datasets (*MUSIC*, *A-MUSIC*, and *A-NATURAL*).

1 Introduction

Sound source separation [1–4] is a classical audio processing problem, where the objective is to recover original component signals from a given mixture audio. Well known example of such task is the cocktail party problem, where multiple people are talking simultaneously (e.g. at a cocktail party) and the observer is attempting to follow one of the discussions. The general form of the problem is challenging and highly underdetermined. Fortunately, one is often able to leverage additional constraints from external cues, such as vision. For instance, the cocktail party problem turns more tractable by observing the lip movements of people [5]. Similar visual cues have also been applied in other sound separation tasks [6–12]. This type of problem setup is often referred as visually guided sound source separation (see e.g. Fig. 1).

Besides separating the component signals from the mixture, one is often interested in identifying the source location. Such task would be intractable from a single audio channel, but could be approached using e.g. microphone arrays [13]. Alternatively, the sound source location could be determined from the visual data [14, 15], which are more often available.

This paper proposes a new approach for visually guided sound source separation and localisation. Our system (Fig. 2), referred as Cascaded Opponent

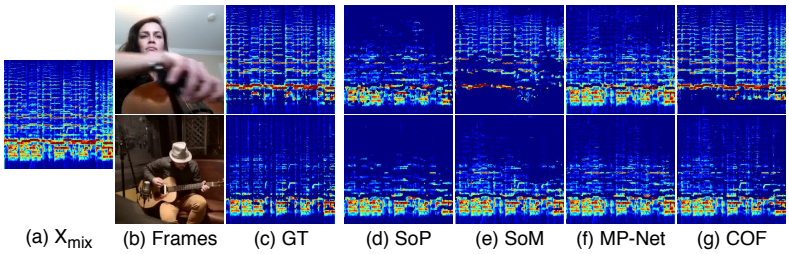

Fig. 1. Visually guided sound source separation aims at splitting the input mixture (column (a)) into component signals corresponding to the given visual cues (column (b)). The proposed COF approach results in better separation performance over the baseline methods SoP [8], SoM [9], and MP-Net [10] on MUSIC dataset [8].

Filter (COF), consists of an initial separation stage and one or more subsequent cascaded Opponent Filter (OF) modules (Fig. 4). The OF module utilises visual cues from *all* videos to reconstruct each component audio. This is in contrast to most previous works (e.g. [8, 9]), where the separation is done only based on the corresponding video. The OF module is very light containing only 17 parameters (in our case) and we show that it can greatly improve the sound separation performance over the recent single stage systems [8, 9] and recursive method [10].

Moreover, since motion is strongly correlated to sound formation [9], we build our system on both appearance and motion representations. To this end, we examine multiple options based on video frames, optical flows, dynamic images [16], and their combinations. Finally, we introduce a Sound Source Location Masking (SSLM) network that, in conjunction with COF, is able to pin point pixel level segmentation of the sound source location. Qualitative results indicate sharper and more accurate results compared to the baselines [8–10]. The entire system is trained using a self-supervised setup with a large set of unlabelled videos.

2 Related Work

Cross-modal Learning from Audio and Vision Aytar *et al.* [17] presented a method for learning joint audio-visual embeddings by minimizing the KL-divergence of their representations. Owens *et al.* [18] proposed a synchronization based cross-modal approach for visual representation learning. Arandjelovic *et al.* [19, 20] associated the learnt audio and visual embeddings by asking whether they originate from the same video. Nagrani *et al.* [21] learned to identify face and voice correspondences. More recent works, include transferring mono- to binaural audio using visual features [11], audio-video deep clustering [22], talking face generation [23], audio-driven 3D facial animation prediction [24], vehicle tracking with stereo sound [25], visual-to-auditory [26, 27], audio-visual navigation [28, 29], and speech embedding disentanglements [30]. Unlike these works, (visually guided) sound source separation aims at splitting the input audio into original components signals.

Video Sequence Representations Most early works in video representations were largely based on direct extensions of the image based models [31–33]. More recently, these have been replaced by deep learning alternatives operating on stack of consecutive video frames. These works can be roughly divided into following categories: 1) 3D CNN applied on spatio-temporal video volume [34]; 2) two-stream CNNs [35–37] applied on video frames and separately computed optical flow frames; 3) LSTM [38], Graph CNN [39] and attention clusters [40] based techniques; and 4) 2D CNN with the concept of dynamic image [16]. Since most of these methods are proposed for action recognition problem, it is unclear which representation would be best suited for self-supervised sound source separation. Therefore, this paper evaluates multiple options and discusses their pros and cons.

(Visually Guided) Sound Source Separation The sound source separation task is extensively studied in the audio processing community. Early works were mainly based on probabilistic models [1–4], while recent methods utilise deep learning architectures [41–44]. Despite of the substantial improvements, the pure audio based source separation remains a challenging task. At the same time, visually guided sound source separation has gained increasing attention. Ephrat *et al.* [5] extracted face embeddings to facilitate speech separation. Similarly, Gao *et al.* [12, 6] utilised object detection and category information to guide source separation. Gan *et al.* [45] associated body and finger movements with audio signals by learning a keypoint-based structured representation. While impressive, these methods rely on the external knowledge of the video content (e.g. speaking faces, object types, or keypoints).

The works by Zhao *et al.* [8, 9] and Xu *et al.* [10] are most related to ours. In [8] the input spectrogram is split into components using U-Net [46] architecture and the separated outputs are constructed as a linear combinations of these. The mixing coefficients are estimated by applying Dilated ResNet to the keyframes representing the sources. The subsequent work [9] introduced motion features and improvements to the output spectrogram prediction. Both of these methods operate in a single stage manner directly predicting the final output. Alternatively, Xu *et al.* [10] proposed to separate sounds by recursively removing large energy components from the sound mixture. Our work explores multiple approaches to utilize the appearance and motion information to refine the sound source separation in multi-stages. The proposed Opponent Filter uses visual features of a sound source to look for incorrectly assigned sound components from opponent sources, resulting in accurate sound separation.

Sound Source localization Early work by Hershey *et al.* [47] localised sound sources by modeling the audio-visual synchrony as a non-stationary Gaussian process. Barzelay *et al.* [48] applied cross-modal association and visual localization by temporal coincidences. Based on canonical correlations, Kidron *et al.* [49] localized visual events associated with sound sources. Recently, Senocak *et al.* [50] learned to localize sound sources in visual scenes by transferring the sound-guided visual concepts to sound context vector. Arandjelovic *et al.* [20] ob-

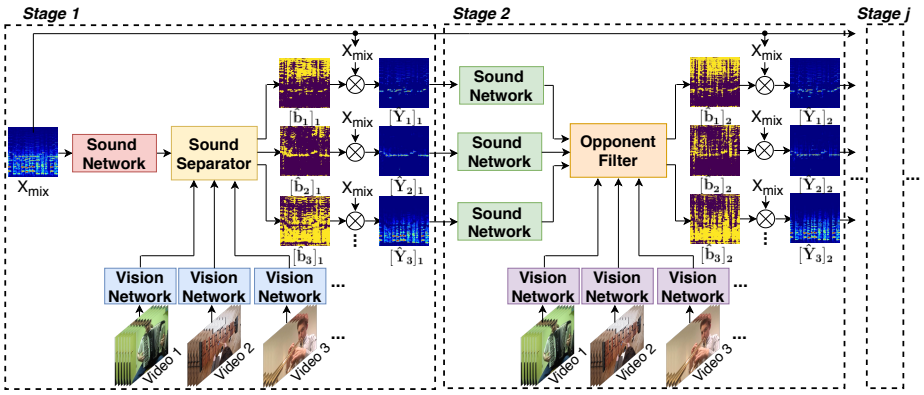

Fig. 2. Architecture of the proposed Cascaded Opponent Filter (COF) network. COF operates in multiple stages: In the first stage, visual representations (vision network) and sound features (sound network) are passed to the sound separator and further produce a binary mask $\hat{b}$ (Eq. (1), (2)) for each output source. Stage two refines the separation result $\hat{Y}$ using the opponent filter (OF) module guided by the visual cues. Later stages are identical to second stage with OF module. The sound networks share parameters only if they are in the same stage. The vision networks share parameters (within and across stages) if they have same architecture.

tained locations by comparing visual and audio embeddings using a coarse grid. Class activation maps were used by [51, 7]. Gao *et al.* [12] localised potential sound sources via a separate object detector. Rouditchenko *et al.* [52] segmented visual objects by leveraging a task of sound separation. Zhao *et al.* [8, 9] and Xu *et al.* [10] visualise the sound sources by calculating the sound volume at each spatial location. In contrast to these methods, which either produce coarse sound location or rely on the external knowledge, we propose a self-supervised SSLM network to localise sound sources on a pixel level.

3 Method

This section describes the proposed visually guided sound source separation method. We start with a short overview and then continue to detailed describe each component.

3.1 Overview

The inputs to our system consist of a mixture audio (e.g. band playing) and a set of videos, each representing one component of the mixture (e.g. person playing a guitar). The objective of the system is to recover the component sound signals corresponding to each video sequence. Fig. 2 illustrates an overview of the approach. Note that the audio signals are represented as spectrograms, which are obtained from the audio stream using Short-term Fourier transform (STFT).

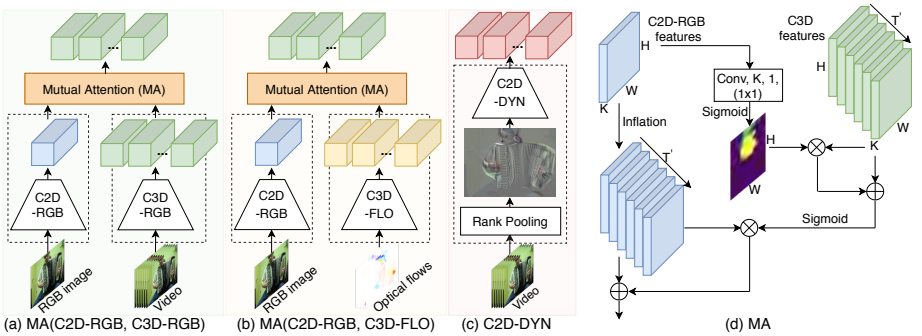

Fig. 3. Architecture of (a) MA(C2D-RGB, C3D-RGB), (b) MA(C2D-RGB, C3D-FLO), (c) C2D-DYN, and (d) MA: Mutual Attention module.

The proposed system consists of multiple cascaded stages. The first stage contains three components: 1) a sound network that splits the input spectrogram into a set of feature maps; 2) a vision network that converts the input video sequences into compact representations; and 3) a sound separator that produces spectrum masks (not shown in Fig. 2) of the component audios (one per video) based on the outputs of the sound and vision networks.

The second stage contains similar sound and vision networks as the first one (internal details may differ). However, instead of the sound separator, the second stage contains a special Opponent Filter (OF) module, which enhances the separation result by transferring sound components between the sources. The output of the filter is passed to the next stage or used as the final output. The following stages are identical to the second one and, for this reason, we refer our method as cascaded opponent filter (COF) network. The final component audios are produced by applying the inverse STFT to the predicted component spectrograms.

In addition, we propose a new Sound Source Location Masking (SSLM) network (not shown in Fig. 2) that indicates the pixels with highest impact on the sound source separation (i.e. source location). The entire network is trained in an end-to-end fashion using artificially generated examples. That is, we take two or more videos and create an artificial mixture by summing the corresponding audio tracks. The created mixture and video frames are provided to the system, which then has to reproduce the original component audios. In the following sections, we will present each component with more details and provide the learning objective used in the training phase.

3.2 Vision Network

The vision network aims at converting the input video sequence (or keyframe) into a compact representation that contains the necessary information of the sound source. Sometimes already a pure appearance of the source (e.g. instrument type) might be sufficient, but, in most cases, the motions are vital cues

to facilitate the source separation (e.g. hand motion, mouth motion, etc.). The appropriate representation may have high model/computation complexity and, to seek for a balance between computational complexity and performance, we study several visual representation options. The models are introduced in the following and the detailed network architectures are provided in the supplementary material. In all cases, we assume that the input video sequence is of size $3 \times 16H \times 16W$ and has T frames.

The first option, referred as **C2D-RGB**, is a pure appearance-based representation. This is obtained by applying a dilated ResNet18 [53] to a single keyframe extracted from the sequence. More specifically, given an input RGB image of size $3 \times 16H \times 16W$, the C2D-RGB produces a representation of size $K \times H \times W$. Dynamic image [16] is a compact representation, which summarises the appearance and motion of the entire video sequence into a single RGB image by rank pooling the original pixel data. The second option, referred as **C2D-DYN**, first converts the input video into a dynamic image (size $3 \times 16H \times 16W$) and then applies a dilated ResNet18 [53] to produce a representation of size $K \times H \times W$. Fig. 3c illustrates C2D-DYN option.

The third option, referred as **C3D-RGB**, applies 3D CNN to extract the appearance and motion information from the sequence simultaneously. C3D-RGB uses 3D version of ResNet18 and produces a representation of size $T^{'} \times K \times H \times W$. The optical flow [35, 54, 55] explicitly describes the motion between the video frames. The fourth option, referred as **C3D-FLO**, first estimates the optical flow between the consecutive video frames using LiteFlowNet [55], and then applies 3D ResNet18 to the obtained flow sequence. C3D-FLO produces a representation of size $T^{'} \times K \times H \times W$.

In addition, following the recent work [36] in action recognition, we propose a set of two stream options by combining pairs of C2D-RGB, C3D-RGB, and C3D-FLO representations using Mutual Attention (MA) module. The module is depicted in Fig. 3d. It enhances the sound source relevant motions and eliminates motion irrelevant appearance by giving the mutual attention mechanism. Finally, we receive the mutual attentive features of dimension $T^{'} \times K \times H \times W$ from the two-stream structures, which are referred to as **MA(C2D-RGB, C3D-RGB)** and **MA(C2D-RGB, C3D-FLO)**. Fig. 3a and 3b illustrate these options. We omit the model of two 3D streams MA(C3D-RGB, C3D-FLO) due to large size of the resulting model.

3.3 Sound Network

The sound network splits the input audio spectrogram into a set of feature maps. The network is implemented using U-Net [46] architecture and it converts the input spectrogram of size $HS \times WS$ into an output of size $HS \times WS \times K$. Note that the number of created feature maps K is equal to the visual feature dimension K in the previous section. At the first stage, the input to the sound network is the original mixture spectrogram X_{mix}, while in later stages, the sound network operates on the current estimates of the component spectrograms. This

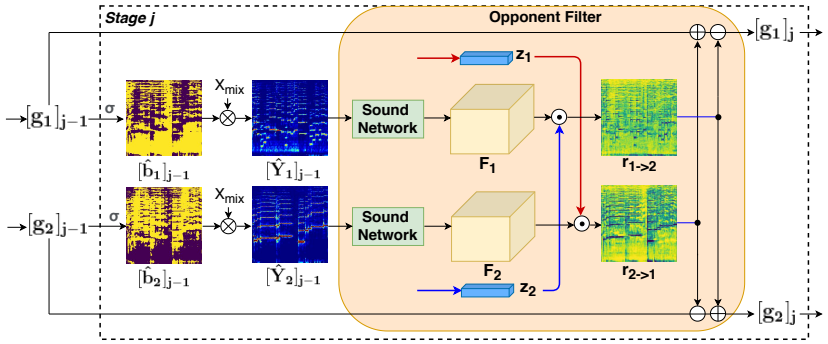

Fig. 4. An illustration of the Opponent Filter (OF) module at stage j in the case of two sound sources. The input consists of the visual representation $\mathbf{z}$ and the previous spectrum mask $[g]_{j-1}$, for both sources. First, we obtain the spectrograms $\hat{Y}$ for both sources from the spectrum masks (Eq. (2)). Second, the spectrograms are turned into feature maps F with the sound network (Sec. 3.3). Third, the visual representation $\mathbf{z}_2$ and the feature map F_1 are used to identify components from the source 1 that should belong to the source 2 ($r_{1->2}$ in figure). The spectrum masks are updated accordingly by subtracting from $[g_1]_{j-1}$ and adding to $[g_2]_{j-1}$. Similar operation is done for the source 2. Finally, the updated spectrum masks $[g_1]_j$ and $[g_2]_j$ are passed to the next stage.

allows stages to focus on different details of the spectrogram. In the following, we will denote the kth feature map, produced by the sound network for an input spectrogram X, as $S(X)_k$.

3.4 Sound Separator

The sound separator combines the visual representations with the sound network output and produces an estimate of the component signals. First, we apply global max pooling operation over the spatial dimensions ($H \times W$) of the visual representation. For 3D CNN-based options, we further apply max pooling layer along the temporal dimension (T'). As a result, we obtain a feature vector $\mathbf{z}$ with K elements. We combine $\mathbf{z}$ with sound network output using a linear combination to predict the spectrum masks g as Eq. (1).

$$g(\mathbf{z}, X) = \sum_{k=1}^{K} \alpha_k \, \mathbf{z_k} * S(X)_k + \beta, \tag{1}$$

where α_k and β are learnable weight parameters, $\mathbf{z}_k$ is the kth element of visual vector $\mathbf{z}$, and $S(X)_k$ is the kth sound network feature map for a spectrogram X.

3.5 Opponent Filter Module

The structure of the Opponent Filter (OF) module is illustrated in Fig. 4 in the case of two sound sources. The main idea in the OF is to use visual representation

of the source n to identify spectrum components from the source m that should belong to source n but are currently assigned to m. These are then transferred from source m to n. The motivation behind the construction is to utilise all visual representations $z_1, \ldots, z_N$ to determine each component audio, instead of using only the corresponding one. This is in contrast to the previous works SoP [8], SoM [9] (and approximately for MP-Net [10]), where the output for each source is determined solely by the same visual input. Our approach leads to more efficient use of the visual cues, which is reflected by the performance improvements shown in the experiments (see Sec. 4.2). Moreover, in our case ($K = 16$), the selected architecture requires only 17 parameters (consist of 16 α_k values and one β as shown in Eq. (3)), which makes it very light and efficient to learn. The OF module is used in all but the first stage of the COF.

More specifically, the OF module takes the visual representation $\mathbf{z}$ and the previous spectrum mask $[g]_{j-1}$ for each sound source as an input. Firstly, the spectrum masks are converted to the spectrograms $\hat{Y}$ as

$$\hat{b} = th(\sigma(g)), \quad \hat{Y} = \hat{b} \otimes X_{mix} \tag{2}$$

where σ denotes the sigmoid function, th represents the thresholding operation with value 0.5, and $\otimes$ is the element-wise product. In other words, we first map g into a binary mask $\hat{b}$, and then produce the estimate of the output component spectrogram as an element-wise multiplication between the binary mask $\hat{b}$ and the original mixture spectrogram X_{mix}. g and $\hat{Y}$ are provided for the upcoming stage as inputs (or used as the final output). We denote the outputs corresponding to nth video at stage j as $[g_n]_j$, $[\hat{b}_n]_j$, and $[\hat{Y}_n]_j$. The obtained spectograms are passed to the sound network (see Sec. 3.3), which converts them to feature maps of size $HS \times WS \times K$ denoted by F_n for the source n.

Secondly, the OF module takes one source at a time, referred using index $n \in [1, N]$, and iterates over the remaining sources $m \in \{[1, N] | m \neq n\}$. N is the number of sources in sound mixture. For each pair (n, m) the filter determines a component of source m that should be reassigned to the source n as

$$r_{m->n} = \sum_{k=1}^{K} \alpha_k \mathbf{z}_{\mathbf{n},\mathbf{k}} * F_{m,k} + \beta \tag{3}$$

where $\mathbf{z}_{n,k}$ is the kth element of visual representation of sound n. $F_{m,k}$ is the kth sound network feature maps of sound m. The $r_{m->n}$ denotes the residual spectrum components identified from source m that should belong to source n but are currently assigned to m. The obtained component will be subtracted from the spectrum mask $[g_m]_{j-1}$ and added to $[g_n]_{j-1}$ as follows

$$[g_m]_j = [g_m]_{j-1} \ominus r_{m->n} \tag{4}$$

$$[g_n]_j = [g_n]_{j-1} \oplus r_{m->n} \tag{5}$$

where the $[g_n]_j$ is the spectrum mask (Eq. (1)) of nth video in stage j, $r_{m->n}$ is the residual spectrum components from sound m to sound n. $\oplus$ and $\ominus$ denote the element-wise sum and subtraction, respectively.

The overall process can be summarized in the following Algorithm 1,

Algorithm 1 Algorithm of Opponent Filter (OF) module

for n = 1 ... N **do**
  for m = 1 ... N **do**
    if $n \neq m$ **then**
      $r_{m->n} = \sum_{k=1}^{K} \alpha_k \mathbf{z_{n,k}} * F_{m,k} + \beta$ ▷ obtain $r_{m->n}$
      $[g_m]_j \leftarrow [g_m]_{j-1} \ominus r_{m->n}$ ▷ subtract $r_{m->n}$ from $[g_m]_{j-1}$
      $[g_n]_j \leftarrow [g_n]_{j-1} \oplus r_{m->n}$ ▷ add $r_{m->n}$ to $[g_n]_{j-1}$
    end if
  end for
end for
return $[g]_j$ ▷ return $[g]_j$ of all the sound sources

3.6 Learning Objective

The model parameters are optimised with respect to the binary cross entropy (BCE) loss that is evaluated between the predicted and ground truth masks over all stages. More specifically,

$$\mathcal{L}_{sep} = \sum_{j=1}^{J} r_j \, BCE([\hat{b}]_j, b_{gt}) \tag{6}$$

where r_j is a weight parameter, $[\hat{b}]_j$ is the predicted binary mask, b_{gt} is the ground truth mask (determined by whether the target sound is the dominant component in the mixture), and J is the total number of stages.

3.7 Sound Source Location Masking Network

The objective of the Sound Source Location Masking (SSLM) network is to identify a minimum set of input pixels, for which the COF network would produce almost identical output as for the entire image. In practice, we follow the ideas presented in [56], and build an auxiliary network to estimate a sound source location mask that is applied to the input RGB frames. The SSLM is trained together with the overall model in a self-supervised manner (please see supplementary material). The input video frames are first passed through the SSLM component which outputs a weighted location mask [0,1] having same spatial size as the input frame. The input video frames are multiplied element-wise with the mask, and the result is passed to the COF model. We illustrate the overall structure of the SSLM in Fig. 5a. The final optimisation is done by minimising the following loss function,

$$\mathcal{L} = \sum_{j=1}^{J} r_j \, l_{diff}([\hat{b}_{SSLM}]_j, [\hat{b}]_j) + \lambda \frac{1}{q} \parallel SSLM(I) \parallel_1, \tag{7}$$

where l_{diff} denotes the difference between the $[\hat{b}_{SSLM}]_j$ and $[\hat{b}]_j$ by L_1 norm, $[\hat{b}_{SSLM}]_j$ is the output sound separation mask obtained using only selected pixels,

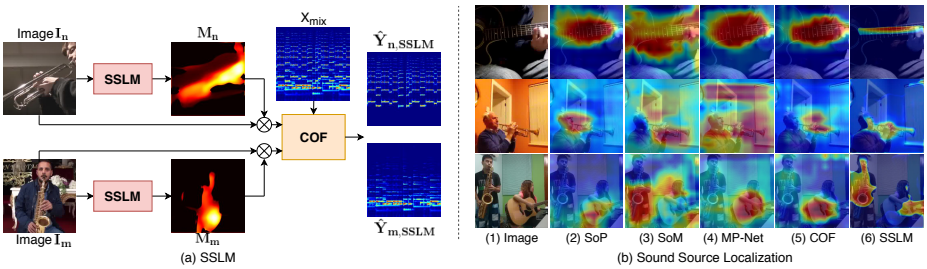

Fig. 5. (a) The diagram of the Sound Source Location Masking (SSLM) network and (b) Visualizing sound source location of our methods in comparison with baseline models SoP [8], SoM [9], and MP-Net [10] on MUSIC dataset.

$[\hat{b}]_j$ is the output separation mask for the original image. r_j and λ are hyperparameters which control the contribution of each loss factor. The $\lambda\frac{1}{q} \parallel SSLM(I) \parallel_1$ norm produces a location mask with only small number of non-zero values. q is the total number of pixels of the $SSLM(I)$.

4 Experiments

We evaluate the proposed approach using Multimodel Sources of Instrument Combinations (MUSIC) [8] dataset, and two sub-sets of AudioSet [57]: A-MUSIC and A-NATURAL. The proposed model is trained using artificial examples, generated by adding audio signals from two or more training videos. The performance of the final sound source separation is measured in terms of standard metrics: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR). Higher is better for all metrics[1].

4.1 Datasets and Implementation Details

MUSIC The Multimodel Sources of Instrument Combinations (MUSIC) [8] dataset is a relatively small, but has high quality. Most of the video frames are well aligned with the audio signals and have little off-screen noise. Part of the original MUSIC dataset is no longer available in YouTube (10% missing at the time of writing). In order to keep dataset size, we replaced the missing entries with similar YouTube videos. Baseline methods (e.g., SoP [8]) in original paper split the dataset into 500 training and 130 validation videos, and report the performances on the validation set (train/test split are not published). Instead, we follow the standard practice of reporting the performance on a separate hold-out test set. For this purpose, we randomly split the dataset into 400 training, 100 validation, and 130 test videos. This leads to 20% less training videos compared to [8]. All tested methods are trained and evaluated with the same data and pre-processing steps (see implementation details).

[1] Note that SDR and SIR scores measure the separation accuracy, SAR captures only the absence of artifacts (and hence can be high even if separation is poor)

Table 1. The sound separation results of the proposed COF network, conditioning on appearance cues, on MUSIC test dataset

Models	SDR	SIR	SAR
COF - 1 stage	5.38	11.00	9.77
$\text{COF}_{addition}$ - 2 stages	6.29	11.83	10.21
$\text{COF}_{subtraction}$ - 2 stages	6.30	12.61	10.13
COF - 2 stages	**8.25**	**14.24**	**12.02**

A-MUSIC and A-NATURAL AudioSet consists of an expanding ontology of 632 audio event classes and is a collection of over 2 million 10-second sound clips drawn from YouTube videos. Many of the AudioSet videos have limited quality and sometimes the visual content might be uncorrelated to the audio track. A-MUSIC dataset is a trimmed musical instrument dataset from AudioSet. It has around 25k videos spanning ten instrumental categories. A-NATURAL dataset is a trimmed natural sound dataset from AudioSet. It contains around 10k videos which cover 10 categories of natural sounds. We split both the A-MUSIC and A-NATURAL dataset samples to 80%, 10%, and 10% as train, validation and test set. More details of datasets are discussed in the supplementary material.

Implementation Details We sub-sample each audio signals at 11kHz and randomly crop an audio clip of 6 seconds for training. A Time-Frequency (T-F) spectrogram of size 512×256 is obtained by applying STFT, with a Hanning window size of 1022 and a hop length of 256, to the input sound clip. We further re-sample this spectrogram to a T-F representation of size 256×256 on a log-frequency scale. We extract video frames at 8fps and give a single RGB image to the C2D-RGB model, $T = 48$ frames to C2D-DYN and all the discussed C3D models. Further implementation details are provided in the supplementary material.

4.2 Opponent Filter

In this section, we assess the performance of the OF module. For simplicity we use only the appearance based features (C2D-RGB) in all stages. The baseline is provided by the basic single stage version denoted as COF - 1 stage, which does not contain the opponent filter module. The results provided in Table 1 indicate a clear improvement from the OF stages.

In addition, we evaluate the impact of the "addition" and "subtraction" branches in the OF module. To this end, we implement two versions $\text{COF}_{addition}$ and $\text{COF}_{subtraction}$, which include only the "addition" (Eq. (5)) or "subtraction" (Eq. (4)) operation in the OF, respectively. The corresponding results in Table 1 indicate that both versions obtain similar performance which is between the baseline and the full model. We conclude that both operations are essential part of the OF module and contribute equally to the sound separation result.

Table 2. The sound separation results with COF, conditioning on different visual cues, on the MUSIC test dataset. Table contains three blocks: 1) single-stage COF associated with visual cues predicted from MA-RGB, MA-FLO, and C2D-DYN; 2) two-stage extension of the models in the first block; 3) two-stage COF with C2D-RGB at stage 1 and C3D-RGB, C3D-FLO, or C2D-DYN at stage 2

	Models	SDR	SIR	SAR
1	COF(MA-RGB)	6.68	12.24	10.63
	COF(MA-FLO)	5.84	11.39	10.27
	COF(C2D-DYN)	6.37	11.75	10.79
2	COF(MA-RGB, MA-RGB)	8.78	15.07	12.10
	COF(MA-FLO, MA-FLO)	8.71	15.07	11.83
	COF(C2D-DYN, C2D-DYN)	8.95	15.03	12.07
3	COF(C2D-RGB, C3D-RGB)	8.97	15.06	**12.53**
	COF(C2D-RGB, C3D-FLO)	9.04	15.28	12.24
	COF(C2D-RGB, C2D-DYN)	**9.17**	**15.32**	12.37

4.3 Visual Representations

We firstly separate sounds by implementing a single stage network, associating with the appearance and motion cues discussed in Sec. 3.2. We denote the MA(C2D-RGB, C3D-RGB) and MA(C2D-RGB, C3D-FLO) as MA-RGB and MA-FLO in Table 2. As is shown in the block 1 of Table 2, the results with appearance and motion cues clearly surpass the network with only appearance cues from C2D-RGB in Table 1, which proposes that the motion representation is important for the sound separation quality. Block 2 shows the performance of how the visual information separates sounds in a two-stage manner. Explicitly, we replace the vision network at each stage in Fig. 2 with MA-RGB, MA-FLO, and C2D-DYN. Block 1 and 2 report that the three two-stage networks obtain similar performance and outperform their single-stage counterparts from block 1 with a large margin.

Finally, we evaluate an option where the first stage utilises only appearance based option and the second stage applies motion cues. In practice, we combine C2D-RGB with C3D-RGB, C3D-FLO, or C2D-DYN. The results in block 3 of Table 2, indicate that this combination obtains similar or even better performance than the options where motion information was provided for both stages. We conclude that the appearance information is enough to facilitate coarse separation at first stage. The motion information is only needed at the later stages to provide higher separation quality. It is worth noting that the COF(C2D-RGB, C2D-DYN) combination has less computation complexity and better performance compared to the 3D CNN alternatives. **Therefore, we apply C2D-RGB for the 1st stage and C2D-DYN for the later stages for all the remaining experiments**.

Table 3. Sound separation performance with 2 and 3 stages COF models compared with three recent baselines SoP [8], SoM [9], and MP-Net [10], on MUSIC, A-MUSIC, and A-NATURAL datasets. The top 2 results are bolded.

Models \ Datasets	MUSIC			A-MUSIC			A-NATURAL		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
SoP	5.38	11.00	9.77	2.05	5.36	10.69	2.83	7.24	8.51
SoM	4.83	11.04	8.67	2.56	5.98	8.80	2.56	7.69	8.02
MP-Net	5.71	11.36	10.45	2.34	5.27	**11.27**	3.20	8.17	8.68
COF - 2 stages	**9.17**	**15.32**	**12.37**	**3.31**	**7.08**	10.74	**4.00**	**8.85**	**8.70**
COF - 3 stages	**10.07**	**16.69**	**13.02**	**5.42**	**9.47**	**10.94**	**4.10**	**8.60**	**10.58**

4.4 Comparison with the State-of-the-Art

We compare the 2-stage and 3-stage of the proposed COF model with three recent baseline methods SoP [8], SoM [9], and MP-Net [10]. For SoP we use the publicly available implementation from the original authors. For SoM and MP-Net we use our own implementations since there were no publicly available versions. The corresponding results for MUSIC[2], A-MUSIC, and A-NATURAL datasets are provided in Table 3, Fig. 1, and Fig. 6. The quantitative results indicate that our model outperforms the baselines with a large margin across all three datasets.

Increasing the number of stages: We observe that the computational cost increases approximately linearly with respect to the number of stages. The performance generally improves until it reaches a plateau. COF with 2, 3, 4, and 5 stages obtain SDRs of 9.17, 10.07, 10.12, and 10.32 on MUSIC dataset, respectively. The corresponding FLOPs (GMACS) are 8.05, 12.06, 16.06, and 20.07. The performance plateaus at 3 stages, which led to a compromise at this point.

Mixture of three sources: We assess the COF model using a mixture of three sound sources from the MUSIC dataset. In this case, the two-stage model obtains SDR: 3.33, SIR: 10.32, and SAR: 6.70 which are clearly higher than SDR: 1.30, SIR: 8.66, and SAR: 5.73 obtained with MP-Net [10] that is particularly designed for the multi-source case. As discussed in Sec. 3.5, the computational cost of COF scales approximately linearly with the number of sources. For instance, the FLOPs (GMACS) for 2, 3, 4, 5, 10, and 15 sources are 8.05, 11.09, 14.12, 17.16, 32.36, and 47.62 respectively.

4.5 Visualizing Sound Source Locations

We compare the sound source localizing capability of our best two-stage model with state-of-the-art methods in Fig. 5b. Columns (2)-(5) display the sound

[2] We note that due to the differences in the dataset and evaluation protocol (see Sec. 4.1.) the absolute results differ from those reported in [8] and [9] for MUSIC.

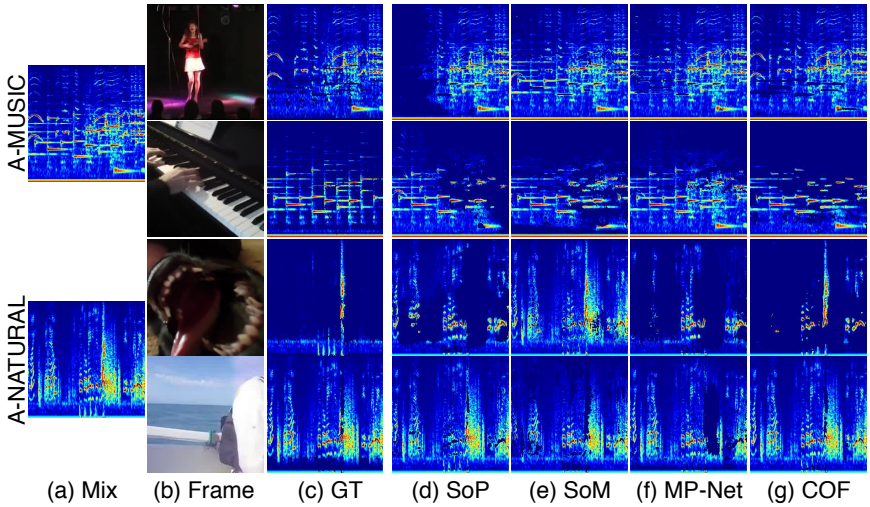

Fig. 6. Visualizing sound source separation of our 2-stage COF model on A-MUSIC and A-NATURAL datasets, in comparison with baseline methods SoP [8], SoM [9], and MP-Net [10].

energy distributions of spatial location in heatmaps on input frame during inference. COF produces precise associations between visual representation and separated sounds, though columns (5) is just the visualization from the first stage of COF. As we know, the spatial features from ConvNet usually have small resolution (14×14 pixels in this work). Thus, the final visualized location is generally coarse after up-sampling the heatmap to the resolution of the input image. Differently, our proposed SSLM learns to predict a pixel-level sound source location mask, as shown in column (6), which precisely localizes sound sources and preserves high quality of sound separation. Further examples are provided in the supplementary material.

5 Conclusions

We proposed an innovative framework of visually guided Cascaded Opponent Filter (COF) network to recursively refine sound separation with visual cues of sound sources. The proposed Opponent Filter (OF) module uses visual features of all sound sources to look for incorrectly assigned sound components from opponent sounds, resulting in accurate sound separation. For this purpose, we studied different visual representations based on video frames, optical flows, dynamic images, and their combinations. Moreover, we introduced a Sound Source Location Making (SSLM) network, together with COF, to precisely localize sound sources.

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

References

1. Ghahramani, Z., Jordan, M.I.: Factorial hidden markov models. In: Advances in Neural Information Processing Systems. (1996) 472–478
2. Roweis, S.T.: One microphone source separation. In: Advances in neural information processing systems. (2001) 793–799
3. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.i.: Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons (2009)
4. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE transactions on audio, speech, and language processing **15** (2007) 1066–1074
5. 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. arXiv preprint arXiv:1804.03619 (2018)
6. 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) 35–53
7. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 631–648
8. 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) 570–586
9. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 1735–1744
10. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 882–891
11. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 324–333
12. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3879–3888
13. Pertilä, P., Mieskolainen, M., Hämäläinen, M.S.: Closed-form self-localization of asynchronous microphone arrays. In: 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, IEEE (2011) 139–144
14. 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) 247–263
15. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9248–9257
16. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S.: Dynamic image networks for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3034–3042
17. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in neural information processing systems. (2016) 892–900

18. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: European conference on computer vision, Springer (2016) 801–816
19. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 609–617
20. Arandjelovic, R., Zisserman, A.: Objects that sound. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 435–451
21. Nagrani, A., Albanie, S., Zisserman, A.: Seeing voices and hearing faces: Cross-modal biometric matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8427–8436
22. Alwassel, H., Mahajan, D., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. arXiv preprint arXiv:1911.12667 (2019)
23. 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. Volume 33. (2019) 9299–9306
24. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 10101–10111
25. 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) 7053–7062
26. Hu, D., Wang, D., Li, X., Nie, F., Wang, Q.: Listen to the image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7972–7981
27. Gan, C., Huang, D., Chen, P., Tenenbaum, J.B., Torralba, A.: Foley music: Learning to generate music from videos. arXiv preprint arXiv:2007.10984 (2020)
28. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Audio-visual embodied navigation. arXiv preprint arXiv:1912.11474 (2019)
29. Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B.: Look, listen, and act: Towards audio-visual embodied navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE (2020) 9701–9707
30. Nagrani, A., Chung, J.S., Albanie, S., Zisserman, A.: Disentangled speech embeddings using cross-modal self-supervision. arXiv preprint arXiv:2002.08742 (2020)
31. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2008) 1–8
32. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. (2009)
33. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. (2008)
34. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. (2015) 4489–4497
35. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. (2014) 568–576
36. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 6299–6308

37. Zhan, X., Pan, X., Liu, Z., Lin, D., Loy, C.C.: Self-supervised learning via conditional motion propagation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1881–1889
38. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 2625–2634
39. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Proceedings of the European conference on computer vision (ECCV). (2018) 399–417
40. Long, X., Gan, C., De Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: Purely attention based local feature integration for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7834–7843
41. Simpson, A.J., Roma, G., Plumbley, M.D.: Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In: International Conference on Latent Variable Analysis and Signal Separation, Springer (2015) 429–436
42. Chandna, P., Miron, M., Janer, J., Gómez, E.: Monoaural audio source separation using deep convolutional neural networks. In: International conference on latent variable analysis and signal separation, Springer (2017) 258–266
43. Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2016) 31–35
44. Grais, E.M., Plumbley, M.D.: Combining fully convolutional and recurrent neural networks for single channel audio source separation. In: Audio Engineering Society Convention 144, Audio Engineering Society (2018)
45. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10478–10487
46. 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, Springer (2015) 234–241
47. Hershey, J.R., Movellan, J.R.: Audio vision: Using audio-visual synchrony to locate sounds. In: Advances in neural information processing systems. (2000) 813–819
48. Barzelay, Z., Schechner, Y.Y.: Harmony in motion. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2007) 1–8
49. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Volume 1., IEEE (2005) 88–95
50. 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. (2018) 4358–4366
51. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Learning sight from sound: Ambient sound provides supervision for visual learning. International Journal of Computer Vision **126** (2018) 1120–1137
52. Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., Torralba, A.: Self-supervised audio-visual co-segmentation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2019) 2357–2361

53. 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. (2016) 770–778
54. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8934–8943
55. Hui, T.W., Tang, X., Change Loy, C.: Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8981–8989
56. Hu, J., Zhang, Y., Okatani, T.: Visualization of convolutional neural networks for monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3869–3878
57. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2017) 776–780