Co-Learn Sounding Object Visual Grounding and Visually Indicated Sound Separation in A Cycle

Yapeng Tian[1,*], Di Hu[2,*], and Chenliang Xu[1]
[1]University of Rochester [2]Baidu Research

1. Introduction

There are rich synchronized audio and visual events in our daily videos. Inside the events, audio scenes are associated with the corresponding visual objects, meanwhile, sounding objects can indicate and help to separate their individual sounds in the audio track. Based on this observation, in this paper, we propose a cyclic co-learning (CCoL) paradigm that can jointly learn sounding object visual grounding and visually indicated audio separation in a unified framework. Concretely, we locate sounding objects from all objects in videos with a visual grounding network and then learn an audio-visual sound separation network to separate sounds from individual sounding objects. Moreover, in the framework, sounding object visual grounding labels can be adaptively adjusted based on sound separation results to simultaneously improve grounding and separation models, which builds a co-learning cycle for the two tasks. Extensive experiments show that the proposed framework achieves state-of-the-art performance on both sounding object visual grounding and visually indicated sound separation tasks and they can benefit from each other with our cyclic co-learning.

2. Method

We first give an overview of our co-learning framework for sounding object visual grounding and visually indicated sound separation in Sec. 2.1. Upon the framework, we propose a grounding network that can recognize sounding objects for both single and mixed sounds in Sec. 2.2 and introduce an audio-visual sound separation network to separate sounds for grounded individual sounding objects in Sec. 2.3. Finally, we make co-learning in a cycle with an adjustment learning method to correct grounding labels in Sec. 2.4, which can simultaneously improve grounding and separation performance.

2.1. Co-Learning Framework

Given an unlabeled video clip V with the synchronized sound $s(t)$, $\mathcal{O} = \{O_1, ..., O_N\}$ are N detected objects in the video frames and the sound mixture $s(t) = \sum_{n=1}^{N} s_n(t)$. Here, $s_n(t)$ is the sound of the object O_n. When it is silent, $s_n(t) = 0$. Our co-learning aims to recognize each sounding object O_n and then separate its sound $s_n(t)$ for the object. The framework, as illustrated in Fig. 1, mainly contains two modules: sounding object visual grounding network and visually indicated sound separation network. To learn sound separation in the framework, we adopt a commonly used mix-and-separate strategy [2, 4, 11] during training. Given two training video and sound pairs $\{V^{(1)}, s^{(1)}(t)\}$ and $\{V^{(2)}, s^{(2)}(t)\}$, we obtain a mixed sound

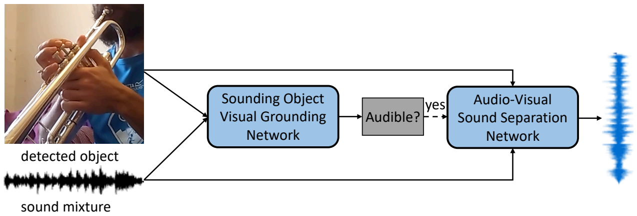
detected object
sound mixture
Sounding Object Visual Grounding Network
Audible?
yes
Audio-Visual Sound Separation Network

Figure 1. Given an detected object from a video and the sound mixture, our model will first recognize whether it is audible via a sounding object visual grounding network and then separate its sound with an audio-visual sound separation network if it is a sound source.

$$s_m(t) = s^{(1)}(t) + s^{(2)}(t) = \sum_{n=1}^{N_1} s_n^{(1)}(t) + \sum_{n=1}^{N_2} s_n^{(2)}(t), \quad (1)$$

and find objects $\mathcal{O}^{(1)} = \{O_1^{(1)}, ..., O_{N_1}^{(1)}\}$ and $\mathcal{O}^{(2)} = \{O_1^{(2)}, ..., O_{N_2}^{(2)}\}$ from the two videos. The sounding object visual grounding network will recognize audible objects from $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ and the visually indicated sound separation network will separate sounds for the grounded objects. Sounds are processed in a Time-Frequency space with the short-time Fourier transform (STFT).

2.2. Sounding Object Visual Grounding

Videos contain various and diverse sounds and visual objects, and not all objects are audible. To find sounding objects in videos $V^{(1)}$ and $V^{(2)}$ and further utilize grounding results for separation, we formulate sounding object visual grounding as a binary matching problem.

Sounding Object Candidates We first find potential audible visual objects from videos using an object detector. In our implementation, we use the Faster R-CNN [8] object detector trained on Open Images dataset [6] from [2] to detect objects from video frames in $V^{(1)}$ and $V^{(2)}$ and obtain $\mathcal{O}^{(1)} = \{O_1^{(1)}, ..., O_{N_1}^{(1)}\}$ and $\mathcal{O}^{(2)} = \{O_1^{(2)}, ..., O_{N_2}^{(2)}\}$.

[*] Equal contribution.

Next, we learn to recognize sounding objects in $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ associated with $s^{(1)}(t)$ and $s^{(2)}(t)$, respectively. For simplicity, we use an object O and a sound $s(t)$ as a example to illustrate our grounding network.

Audio Network Raw waveform $s(t)$ is transformed to an audio spectrogram S with the STFT. An VGG [10]-like 2D CNN architecture: VGGish followed by a global max pooling (GMP) is used to extract an audio embedding f_s from S.

Visual Network The visual network extracts features from detected visual object O. We use the pre-trained ResNet-18 [3] model before the last fully-connected layer to extract a visual feature map and perform a GMP to obtain a visual feature vector f_o for O.

Grounding Module The audio-visual grounding module takes audio feature f_s and visual object feature f_o as inputs to predict whether the visual object O is one of the sounding makers for $s(t)$. We solve it using a binary classification network. It first concatenates f_s and f_o and then uses a 3-layer Multi-Layer Perceptron (MLP) with a Softmax function to output a probability score $g(s(t), O) \in \mathcal{R}^2$. Here, if $g(s(t), O)[0] >= 0.5$, $\{s(t), O\}$ is a positive pair and $s(t)$ and O are matched; otherwise, O is not a sound source.

Training and Inference To train the sounding object visual grounding network, we need to sample positive/matched and negative/mismatched audio and visual object pairs. It is straightforward to obtain negative pairs with composing audio and objects from different videos. For example, $s^{(1)}(t)$ from $V^{(1)}$ and an randomly selected object $O_r^{(2)}$ from $V^{(2)}$ can serve as a negative pair. However, positive audio-visual pairs are hard to extract since not all objects are audible in videos. If an object from $V^{(1)}$ is not audio source, the object and $s^{(1)}(t)$ will be a negative pair, even though they are from the same video. To address the problem, we cast the positive sample mining as a multiple instance learning problem and sample the most confident pair as a positive sample with a grounding loss as the measurement:

$$\hat{n} = \underset{n}{\operatorname{argmin}}\, f(g(s^{(1)}(t), O_n^{(1)}), y_{pos}), \quad (2)$$

where $f(\cdot)$ is a cross-entropy function; $y_{pos} = [1, 0]$ is an one-hot encoding for positive pairs; $O_{\hat{n}}^{(1)}$ and $s^{(1)}$ will be the positive audio-visual pair for training. With the sampled negative and positive data, we can define the loss function to learn the sounding object visual grounding:

$$l_1 = \frac{1}{2}(f(g(s^{(1)}(t), O_r^{(2)}), y_{neg}) + f(g(s^{(1)}(t), O_{\hat{n}}^{(1)}), y_{pos})), \quad (3)$$

where $y_{neg} = [0, 1]$ is the negative label. The visual grounding network can be end-to-end optimized with sampled training pairs via l_1.

During inference, we can feed audio-visual pairs $\{O_i^{(1)}, s^{(1)}(t)\}_{i=1}^{N_1}$ and $\{O_i^{(2)}, s^{(2)}(t)\}_{i=1}^{N_2}$ into the trained model to find sounding objects insides the two videos. To facilitate visually indicated sound separation, we need to detect sounding objects from the sound mixture $s_m(t)$ rather than $s^{(1)}(t)$ and $s^{(2)}(t)$, since the individual sounds are unavailable at a testing stage for separation task.

2.3. Visually Indicated Sound Separation

Given detected objects in $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$, we separate sounds for each object using an audio-visual sound separation network from the sound mixture $s_m(t)$ and mute separated sounds of silent objects.

Audio-Visual Sound Separation Network We adopt a similar audio-visual separator as in [11], which consists of three modules: audio network, visual network, and audio-visual sound synthesizing network. The audio network transforms the input audio mixture STFT spectrogram S_m to a C-channel feature map $A_m \in \mathcal{R}^{C\times F\times T}$ with an U-Net [9] structure. We use an ResNet-18 [3] followed by a linear layer to predict a C-dimension object feature $f_{o_i^{(k)}} \in \mathcal{R}^C$ for each object $O_i^{(k)}$. The audio-visual sound synthesizing network takes A_m and $f_{o_i^{(k)}}$ as inputs and output a spectrogram mask $M_i^{(k)} \in \mathcal{R}^{F\times T}$ with a linearly transformed dot product of the audio and visual features. The separated sound spectrogram: $S_i^{(k)} = S_m \times M_i^{(k)}$ can be obtained by masking the sound mixture. The waveform $s_i^{(k)}$ of the object can be reconstructed by the inverse short-time Fourier transform.

Sounding Object-Aware Separation Using the audio-visual sound separation network, we can predict sound spectrograms $\{S_n^{(1)}\}_{n=1}^{N_1}$ and $\{S_n^{(2)}\}_{n=1}^{N_2}$ for objects in $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$, respectively. To learn the separation network, we can optimize it with a L1 loss function:

$$l_2 = \frac{1}{2}(||S^{(1)} - \sum_{n=1}^{N_1} S_n^{(1)}||_1 + ||S^{(2)} - \sum_{n=1}^{N_2} S_n^{(2)}||_1). \quad (4)$$

However, not all objects are audible and spectrograms from different objects contain overlapping content. Therefore, even an object $O_n^{(1)}$ is not sounding, it can also separate non-zero sound spectrogram $S_n^{(1)}$ from S_m, which will introduce errors during training. To address the problem, we propose a sounding object-aware separation loss function:

$$\begin{aligned} l_2^* = \frac{1}{2}(||S^{(1)} - \sum_{n=1}^{N_1} g^*(s_m(t), O_n^{(1)}) S_n^{(1)}||_1 \\ + ||S^{(2)} - \sum_{n=1}^{N_2} g^*(s_m(t), O_n^{(2)}) S_n^{(2)}||_1), \quad (5) \end{aligned}$$

where $g^*(\cdot)$ is a binarized value of $g(\cdot)[0]$. If an object $O_n^{(1)}$ is not a sound source, $g^*(s_m(t), O_n^{(1)})$ will be equal to zero. Thus, the sounding object-aware separation can help to reduce training errors from silent objects in Eq. 4.

In addition, we introduce additional grounding loss terms to guide the grounding model learning from the sound mixture. Since we have no sounding object annotations, we adopt a similar positive sample mining strategy as in Eq. 2 and define a grounding loss as follows:

$$l_3 = \frac{1}{2}(\min_n f(g(s_m(t), O_n^{(1)}), y_{pos}) + \min_n f(g(s_m(t), O_n^{(2)}), y_{pos})). \quad (6)$$

2.4. Co-learning in a Cycle

Combing grounding and separation loss terms, we can learn the two tasks in a unified way with a co-learning objective function: $l_{col} = l_1 + l_2^* + l_3$.

Although our sounding object visual grounding and visually indicated sound separation models can be learned together, the two tasks loosely interact in l_2^*. Clearly, a good grounding network can help improve the separation task. However, the grounding task might not be able to benefit from co-learning training since there is no strong feedback from separation to guide learning the grounding model. To further strengthen the interaction between the two tasks, we propose a cyclic co-learning strategy, which can make them benefit from each other.

If an object $O_n^{(k)}$ makes sound in video $V^{(k)}$, the separated spectrogram $S_n^{(k)}$ should be close to $S^{(k)}$; otherwise, the difference between $S_n^{(k)}$ and $S^{(k)}$ should be larger than a separated sound spectrogram from an sounding object and $S_v^{(k)}$. We use L1 distance to measure dissimilarity of spectrograms:

$$d_n^{(k)} = ||S_n^{(k)} - S^{(k)}||_1, \quad (7)$$

where $d_n^{(k)}$ will be small for a sounding object $O_n^{(k)}$. Based on the observation, we select the object $O_n^{(k)}$ with the minimum $d_n^{(k)}$ make the dominant sound in V^k to compose positive samples for sounding object visual grounding. Let $\hat{n}_1 = \text{argmin}_n\, d_n^{(1)}$ and $\hat{n}_2 = \text{argmin}_n\, d_n^{(2)}$. We can reformulate grounding loss terms in Eq. 3 and 7 as:

$$l_1^* = \frac{1}{2}(f(g(s^{(1)}(t), O_r^{(2)}), y_{neg}) + f(g(s^{(1)}(t), O_{\hat{n}_1}^{(1)}), y_{pos})), \quad (8)$$

$$l_3^* = \frac{1}{2}(f(g(s_m(t), O_{\hat{n}_1}^{(1)}), y_{pos}) + f(g(s_m(t), O_{\hat{n}_2}^{(2)}), y_{pos})). \quad (9)$$

In addition, if $d_n^{(k)}$ is very large, the object $O_n^{(k)}$ is very likely not be audible, which can help us mine potential negative samples for mixed sound grounding. Specifically, we select the objects that are associated with the largest $d_n^{(k)}$, and $d_n^{(k)}$ must be larger than a threshold ϵ. Let $n_1^* = \text{argmax}_n\, d_n^{(1)}$, s.t. $d_n^{(1)} > \epsilon$ and $n_2^* = \text{argmax}_n\, d_n^{(2)}$, s.t. $d_n^{(2)} > \epsilon$. We can update l_3^* with learning from potential negative samples:

$$l_3^* = \frac{1}{4}\sum_{k=1}^{2}(f(g(s_m(t), O_{\hat{n}_k}^{(k)}), y_{pos}) + f(g(s_m(t), O_{n_k^*}^{(k)}), y_{neg})). \quad (10)$$

Finally, we can co-learn the two tasks in a cycle with optimizing the joint cyclic co-learning loss function: $l_{ccol} = l_1^* + l_2^* + l_3^*$. Inside cyclic co-learning, we use visual grounding to improve sound separation and enhance visual grounding based on feedback from sound separation. The learning strategy can make the tasks help each other in a cycle and significantly improve performance for both tasks.

3. Experiments

3.1. Experimental Setting

Dataset In our experiments, 520 online available musical solo videos from the widely-used MIT MUSIC dataset [11] is used. The dataset includes 11 musical instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin, and xylophone. The dataset is relatively clean and sounding instruments are usually visible in videos. We split it into training/validation/testing sets, which have 468, 26, 26 videos from different categories, respectively. To train and test our cyclic co-learning model, we randomly select three other videos for each video to compose training and testing samples. Let's denote the four videos as A, B, C, D. We compose A, B together as $V^{(1)}$ and C, D together as $V^{(2)}$, while sounds of $V^{(1)}$ and $V^{(2)}$ are only from A and C, respectively. Thus, objects from B and D in the composed samples are inaudible. Finally, we have 18720/260/260 composed samples in our training/validation/testing sets for visual grounding and sound separation tasks.

Evaluation Metrics For sounding object visual grounding, we feeding detected sounding and silent objects in videos into different grounding models and evaluate their binary classification accuracy. For sound separation, we use the commonly used mir eval library [7] to measure performance in terms of two standard metrics: Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR).

Implementation Details We sub-sample audio signals at 11kHz, and each video sample is approximately 6 seconds. The STFT is calculated using a Hann window size of 1022 and a hop length of 256 and each 1D audio waveform is transformed to a 512 × 256 Time- Frequency spectrogram. Then, it is re-sampled to $T, F = 256$. The video frame rate is set as $1fps$ and we randomly select 3 frames per $6s$ video. Objects extracted from video frames are resized to 256 × 256 and then randomly cropped to 224 × 224 as inputs to

Table 1. Sounding object visual grounding performance (%). Top-2 results are highlighted.

Methods	OTS [1]	DMC [5]	Grounding only	CoL	CCoL
Single Sound	58.7	65.3	**72.0**	67.0	**84.5**
Mixed Sound	51.8	52.6	**61.4**	58.2	**75.9**

Table 2. Visually indicated audio separation performance. Top-2 results are highlighted.

Methods	SoP [11]	CoSep [2]	CoL	CCoL	Oracle
SDR	3.42	1.41	**6.50**	**7.27**	7.71
SIR	4.98	4.26	**11.81**	**12.77**	11.42

our network.

Sounding Object Visual Grounding We compare our methods to two recent methods: OTS [1] and DMC [5]. In addition, we make an ablation study to investigate the proposed models. The Grounding only model is trained only with grounding losses: l_1; the co-learning (CoL) model jointly learn visual grounding and sound separation using the l_{col}; and the cyclic co-learning (CCoL) further strengthens the interaction between the two tasks optimized via l_{ccol}. We evaluate sounding object visual grounding performance on both solo and mixed sounds.

Table 1 show sounding object visual grounding results. Even our grounding only has already outperformed the OTS and DMC, which can validate the effectiveness of the proposed MIL-based positive sample mining approach. Then, we can see that the CoL with jointly learning grounding and separation achieves worse performance than the Grounding only model. It demonstrates that the weak interaction inside CoL cannot let the grounding task benefit from the separation task. However, with introducing separation results to help the grounding sample mining, our CCoL is significantly superior over both Grounding only and CoL models. The results can demonstrate the sounding object visual grounding can benefit from visually indicated sound separation with our cyclic learning.

Audio-Visual Sound Separation Results To demonstrate the effectiveness of the proposed CCoL framework on audio-visual sound separation, we compare it to two recent state-of-the-art methods: SoP [11] and CoSep [2] and the CoL baseline model in Table 2. Note that SoP and CoSep are trained using source code provided by the authors and the same training data as ours. Moreover, we show separation results of an Oracle model, which directly feeds ground truth grounding labels of mixed sounds to train the audio-visual separation network.

We can see that our CoL outperforms both SoP and CoSep, and CCoL is better than CoL. The results demonstrate that sounding object visual grounding in the co-learning can help to mitigate training errors from silent video objects in separation, and separation performance can be further improved with the help of enhanced grounding model by cyclic co-learning. Compared to the Oracle model, it is reasonable to see that CCoL has slightly lower SDR. A surprising observation is that CoL and CCoL achieve better results in terms of SIR. One possible reason is that our separation networks can explore various visual objects as inputs during joint grounding and separation learning, which might make the models more robust on SIR.

From the sounding object visual grounding and audio-visual sound separation results, we can conclude that our cyclic co-learning framework can make the two tasks benefit from each other and significantly improve both visual grounding and sound separation performance.

Acknowledgments

This work was supported in part by NSF 1741472, 1813709, and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] R. Arandjelovic and A. Zisserman. Objects that sound. In *ECCV*, 2018. 4

[2] R. Gao and K. Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 1, 4

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, pages 31–35. IEEE, 2016. 1

[5] D. Hu, F. Nie, and X. Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019. 4

[6] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2017. 1

[7] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, 2014. 3

[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[9] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2014. 2

[11] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *ECCV*, 2018. 1, 2, 3, 4