2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Self-Supervised Learning for Alignment of Objects and Sound

Xinzhu Liu, Xiaoyu Liu, Di Guo, Huaping Liu, Fuchun Sun and Haibo Min

***Abstract*— The sound source separation problem has many useful applications in the field of robotics, such as human-robot interaction, scene understanding, etc. However, it remains a very challenging problem. In this paper, we utilize both visual and audio information of videos to perform the sound source separation task. A self-supervised learning framework is proposed to implement the object detection and sound separation modules simultaneously. Such an approach is designed to better find the alignment between the detected objects and separated sound components. Our experiments, conducted on both the synthetic and real datasets, validate this approach and demonstrate the effectiveness of the proposed model in the task of object and sound alignment.**

I. Introduction

Most audio signals are mixtures of several audio sources (speech, music, noises). The task of sound source separation requires these different sources to be separated. It has a wide range of applications in the field of robotics [1]. In a human-robot interaction scene, the robot needs to distinguish the human voice from the background noise to better understand what the human is saying [2]. In a rescue situation, the robot needs to detect human voice from the mixed sound picked up by its sensors so as to better rescue people from dangerous situations [3]. Sound separation is also useful in the tasks of speech recognition and dialogue management [4], search of the closest sound source [5] and auditory scene understanding [6]. In addition, sound separation is a vital component in sound source detection and localization tasks [7]–[9].

Even though the study of sound source separation has a long and rich history, the variant of the problem with only audio modality available (an example being the famous "cocktail party problem" [10]) remains to this day a very difficult task [11]. Early literature in this field began by focusing on signal processing methods [12]. For instance, sparse coding methods are proposed to extract sources from real-world sound signals [13]. Non-negative matrix factorization is a similarly popular signal processing method [14]. More recently, some literature has started to provide a supervised learning treatment to sound separation, applying deep learning methods to this problem [12]. In [15], a deep learning framework is proposed to assign contrastive embedding vectors to regions of different time-frequency in a spectrogram to predict the target spectrogram from the input mixtures [15]. In another recent paper, permutation invariant training is proposed for speaker-independent speech separation [16]. However, all the above-proposed methods require restrictive assumptions and the performance is still not satisfying.

XZ. Liu, D. Guo, HP. Liu and FC. Sun are with Beijing National Research Center for Information Science and Technology, Institute for Artificial Intelligence, Tsinghua University, and the Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China. XY. Liu is with School of Information and Electronics, Beijing Institute of Technology, China. H. Min is with Beijing HiBingo Hi-Tech Co. Ltd. Corresponding author: Huaping Liu(hpliu@tsinghua.edu.cn).

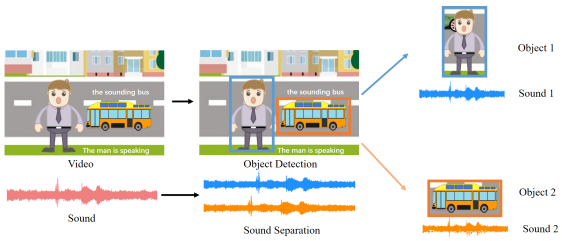

Fig. 1. Our model simultaneously detects visual objects and separates the sound of each object from the sound of the video. In the above video, two types of visual objects are detected and the sound is divided into two parts. Then, the objects and the sound are matched, so that the sound of each object can be obtained.

On the other hand, videos contain both visual images and audio information at the same time. Since the visual and audio information are complementary to each other, we can resort to both the visual and audio information contained in the video clip for sound separation task [17]. Ravulapalli et al. [18] propose to separate multiple concurrent audio and video events with a feature-based approach. To identify sources in video and separate them into different audios, a canonical correlation analysis method is employed for audio-visual source separation [19]. Gao et al. utilize a deep multi-instance multi-label learning framework to separate sound source in videos [20]. Ephrat et al. present a model to isolate a single speech signal from a mixture of other speakers' sound and background noise [21]. At the same time, scholars also conduct research on the sound localization problem. In [22], a low-rank and sparsity method is represented to localize visual objects associated with an audio source and separate the audio signal meanwhile. A deep visual-audio speech enhancement network is proposed to separate a speakers voice with the information of lip regions in the corresponding video [23]. Zhao et al. present the PixelPlayer, a system that learns to separate the sound of each pixel from the input sounds [24]. Furthermore, they modify the model to do independent image segmentation and sound source separation for input videos [25]. For learning representation, Owens et al. propose a self-supervised neural network to learn aligned visual and audio representation,

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

which is applied to the on/off-screen audio sound separation [26]. Some semantic information from vision is also very useful for sound separation. For example, Gao et al. propose a model to generate the sound of instruments with object-category labels [27]. The object detection results are used to help separate the on-screen and off-screen sounds [28], [29].

However, in real-life scenarios when working with robots, the sound separation results in pixel-level is not sufficient for applications, because robots cannot know what each pixel represents in the real environment. What are of greater interest to robots are the actual objects and the sound generated by each object, not individual pixels. Therefore, we propose a self-supervised model to separate the sound in object-level without any label information. In practice, if several objects are sounding simultaneously, as is shown in Fig.1, the robot can detect every object in the environment and separate the sound of every object from the mixed sound. The main contributions are summarized as follows:

1) A novel object detection and sound separation framework is established to perform object detection and sound source separation tasks simultaneously.
2) A self-supervised learning network is developed to build the association between detected objects and sound components to predict the emitted sound for every object.
3) Evaluative experiment is conducted to measure the performance of the proposed model, with data from realistic scenes used for model validation.

This paper is organized as follows. In section II, we illustrate the problem of buliding the alignment between visual objects and sound sources. In section III, the network structure of our model is presented. In section IV, we further explain the process of object detection and sound separation with our model. Finally, experiments to evaluate our model are presented in section V.

II. Problem Formulation

For a short video clip $\mathcal{V}$, usually a key-frame $\mathcal{I}$ is sufficient to capture most of the visual objects. We therefore use

$$\mathcal{O}_I(\mathcal{I}) = \{\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_N\},$$

to represent the set of the detected visual objects in the video $\mathcal{V}$. Please note that $\boldsymbol{O}_n$ represents the n-th detected object and N is the number of the detected objects in the key-frame, which is simply set as the middle one of the video frames.

For the video associated sound track $\mathcal{S}$,

$$\mathcal{O}_S(\mathcal{S}) = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_K\},$$

where $\boldsymbol{S}_k$ represents the k-th sound of an isolated source and K is the number of the sound sources.

In many cases, even if we can successfully detect the objects in the visual images and separate the sound sources for the concerned video, it is still very challenging to construct the alignment between visual objects and sound sources. The main reason is that not all of the visual objects emit sound (e.g., building), while the sound-emitting objects may be invisible in the image (e.g. wind). In this paper, we try to recover these alignment relationship and use this relation to help separating the sound sources. Specifically, the goal is to generate the objects sets $\mathcal{O}_I(\mathcal{I})$, the sound sets $\mathcal{O}_S(\mathcal{S})$, and more importantly, to get the alignment sets

$$\mathcal{A}(\mathcal{V}) = \{(\boldsymbol{O}_n, \boldsymbol{S}_k) | (n,k) \subseteq \{1,2,\cdots,N\} \times \{1,2,\cdots,K\}\},$$

where the sound source $\boldsymbol{S}_{k^*}$ is emitted by the visual object $\boldsymbol{O}_{n^*}$, for $(\boldsymbol{O}_{n^*}, \boldsymbol{S}_{k^*}) \in \mathcal{A}(\mathcal{V})$.

To solve this problem from the viewpoint of machine learning, we collect a lot of unlabeled video clips which contain objects of interest and their sounds, and construct a self-supervised deep learning architecture to learn the intrinsic relationship between visual objects and sound sources. It should be noted this work is novel compared with existing work [24], [25], [20]. In fact, Refs. [24] and [25] study the pixel-level and segmentation-level sound, but do not consider the object-level alignment. The most relevant work is [20], which introduces video-level object prediction and constructs a multi-instance multi-label network. In their work, sound separation is performed using Non-negative Matrix Factorization (NMF), while our work recovers the audio-visual alignment using deep learning architecture, and leverages the advantages of visual object detection and sound separation.

III. Network Architecture

The model is composed of three main modules: visual object detection module, sound feature extraction module, and sound separation module, which is presented in Fig. 2. The visual object detection module detects visual objects in the key-frame of every video, and extracts visual features of each detected object. The sound feature extraction module is utilized to extract features of the sound, and divide the features into several components that may contain specific semantic meanings. The sound separation module integrates the information obtained from visual object detector and sound feature extractor to generate the sound of each detected object.

A. Visual Object Detection

Visual object detection module takes the key-frame as input, which is the mid-point frame of videos. The Faster-RCNN [30] network is utilized to detect objects and simultaneously extract visual features. For each key-frame $\mathcal{I}$, the detection boxes and visual features of each object are recorded, and the set of visual features of objects is represented as

$$\{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_N\},$$

where $\boldsymbol{f}_n$ is the visual feature of the n-th detected object from the last feature layer of the Faster-RCNN network.

B. Sound Feature Extraction

As for the input sound $\mathcal{S}$, the Short-Time Fourier Transform (STFT) is used to convert the audio signal into the spectrogram. This spectrogram is resampled on a log-frequency scale to obtain a T-F spectrogram $\mathcal{P}$. Then, the U-Net [31] structure is applied to the spectrogram to extract the sound

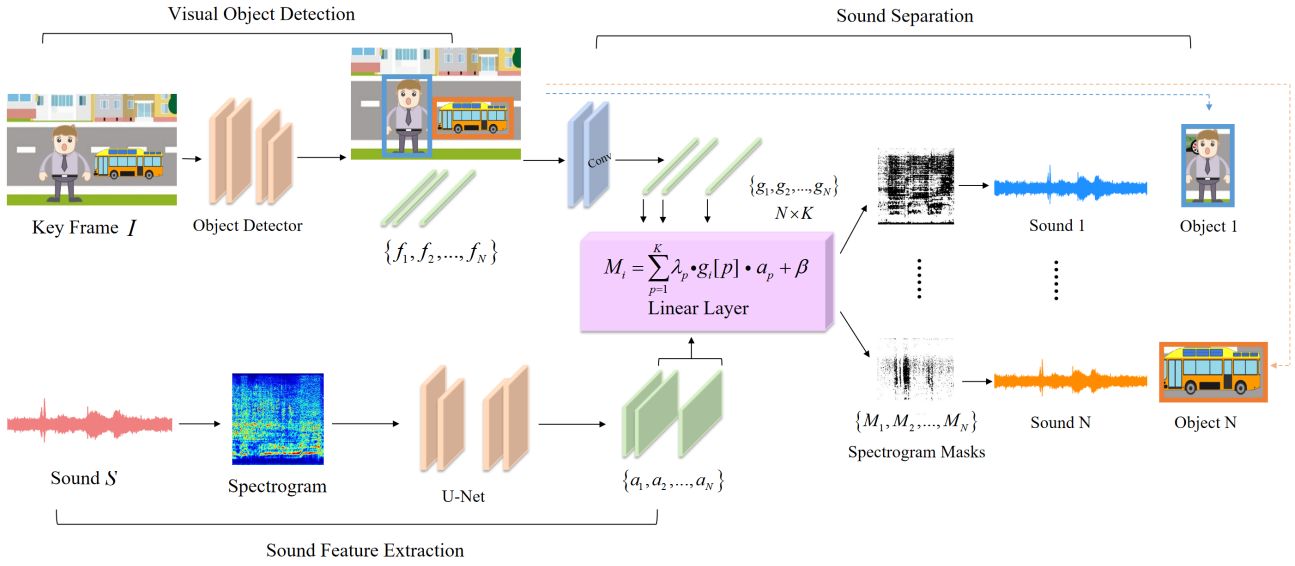

Fig. 2. The structure of the proposed model.

feature and divide the sound into K components, which is represented as the set $\mathcal{C}$

$$\mathcal{C} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K\}.$$

These K components can be regarded as the basic units of sound sources, and the feature of every sound source can be obtained by weighting the feature of those components together. The components can have some semantic information, so that combining different components in a specific way can get the sound of a specific object.

C. Sound Separation Network

The visual features of objects and sound components belong to different modalities, so it is difficult to build connection directly between visual and auditory modalities. In order to establish the association between two modalities, a two-layer convolution module is utilized to project the visual features first into the space, in which the mapped features can be compared and processed jointly with sound components. The set of visual features $\mathcal{G}$ after projection is represented as

$$\mathcal{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_N\},$$

where $\boldsymbol{g}_n \in \mathbb{R}^K$ is the n-th object's mapped visual feature.

Then, the sound separation network matches every object with a specific sound source by processing the projected visual features and the sound components jointly. To make the training process easier, the separation network is utilized to predict the mask $\boldsymbol{M}_n$ for the visual object $\boldsymbol{O}_n$, which can separate the sound of an object from the whole video sound. The set of the mask for every video is obtained as

$$\{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_N\}.$$

In fact, the separation network learns the function $\mathcal{F}(\mathcal{C}, \boldsymbol{g}_n)$ to predict the mask $\boldsymbol{M}_n$. In the proposed model, the function $\mathcal{F}$ is defined as a linear layer, which combines the K-dimensional visual feature vectors with K sound components to perform the spectrogram masks prediction.

After separation, the mixed sound spectrogram and the mask are used to recover the sound waveform of every object. Finally, the model can detect every object in the video, and separate the sound of every object.

D. Training Process

In fact, training directly on individual video data is difficult to implement, due to the much little extra information obtained during the self-supervised learning process. For those videos with only one object, the model cannot learn the separation function $\mathcal{F}$ by training on individual videos because there is nothing useful to learn for sound separation. For those videos with several objects, it is difficult to obtain precise ground truth of object sound from the video itself. To train the model more effectively, we generate training samples manually. We mix the sound of two videos, and formulate the objective of the training process to separate the sound of individual video from the mixed sound of two videos.

In the training process, in order to generate training samples, we select two videos $\mathcal{V}^{(1)}, \mathcal{V}^{(2)}$ from the training dataset randomly. Then we mix the sound of each video to generate the mixed sound $\mathcal{S}_{mix}$

$$\mathcal{S}_{mix} = \mathcal{S}^{(1)} \circ \mathcal{S}^{(2)},$$

where $\circ$ represents the mix operation of two sounds. The set of sound component features of sound $\mathcal{S}_{mix}$ obtained by the sound feature extraction is $\mathcal{C}_{mix}$. Given the key-frame $\mathcal{I}^{(1)}, \mathcal{I}^{(2)}$ of the two videos, the set of projected visual features are $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$. Our model is trained to minimize the difference between predicted spectrogram masks $Mask^{(1)}$, $Mask^{(2)}$ obtained by function $\mathcal{F}$ and real masks of the two videos.

The ground truth of the spectrogram masks of the two videos is defined as

$$Gt^{(1)} = \frac{\mathcal{P}^{(1)}}{\mathcal{P}_{mix}} \qquad Gt^{(2)} = \frac{\mathcal{P}^{(2)}}{\mathcal{P}_{mix}},$$

where $\mathcal{P}^{(1)}, \mathcal{P}^{(2)}, \mathcal{P}_{mix}$ represent the spectrogram of the input sound and the mixed sound respectively. The loss function is defined as the distance between predicted masks and the ground truth masks,

$$Loss = Dis(Gt^{(1)}, Mask^{(1)}) + Dis(Gt^{(2)}, Mask^{(2)}),$$

where Dis is the pixelwise distance between the separated spectrogram masks and its corresponding ground truth. The distance needs to be as small as possible.

It should be noted that the training process is a form of self-supervised learning process. Although ground truth masks are provided for our model to learn, we do not use any video label information or other extra information except the videos themselves. We use the manually mixed sound as training data, and the ground truth sound, which is exactly the sound of individual videos, is naturally available without any additional supervision information. Hence, the ground truth can be obtained easily and no extra information is used during the self-supervised learning procedure.

IV. Object Detection and Sound Separation

The goal of our model is to find all the visual objects in the video and separate the sound of every visual object from the mixed sound signal of the video simultaneously. Specifically, this model is designed to address the problem of particular kinds of noise that occur frequently in real-life during human speech. Therefore, we mainly focus on separating the sounds of people and other interfering sounds.

After training, the model should be applied to videos with multiple sound sources in the prediction process, to perform object detection and sound separation tasks simultaneously. The model predicts the spectrogram mask for each object in the video, and recovers the sound waveform of each object with the mask and the mixed spectrogram. As the ground truth of object sound in videos with multiple sound sources is difficult to obtain, only qualitative analysis is completed in the prediction phase.

The difference between the prediction process and the training process is the type of processed data and the target. In the prediction process, the model is utilized to separate the object sound from an individual video sound. However, in the training process, the model tries to separate the sound of individual videos from the mixed sound of two videos. In fact, what the model learns in the training process is effective and useful in object sound separation in the prediction process.

V. Experiment

A. Dataset

We build a new video dataset for our experiment. Our dataset is a subset of AudioSet [32], which contains 10000 videos with 10 seconds. According to goals we want to achieve, we choose 8 categories, containing *Male Speech*, *Female Speech*, *Baby cry*, *Bus*, *Truck*, *Motorcycle*, *Train horn*, and *Race car*. The dataset is divided into two splits: 9000 videos in the training set and 1000 videos in the testing set. Some sample frames are shown in Fig.3.

Fig. 3. Sample frames from the selected dataset.

All the video data can be divided into two more general categories, the speaking people and sounding vehicles. During the training and testing process, we choose two videos randomly from those two general categories respectively and mix their sound. Then our model separates the sound of each video from the whole sound. Among 9000 videos in the training set, 4500 videos belong to the category of person speech, and the other 4500 videos belong to the category of vehicle's sound. Besides, both 500 videos are belonging to the category of person speech and vehicle's sound respectively in the testing set.

B. Data Processing

The data is preprocessed to obtain audio-visual pairs for training, The middle frame of videos is extracted as the key-frame of each video. Each audio sample is about 10 seconds, and we resample the audio signals to 11kHz. After that, the signals are converted into spectrograms using the Short-Time Fourier Transform (STFT) with a window size of 1022 and a hop length of 256, which results in 256 $\times$ 512 Time-Frequency(T-F) representations of sound. To speed up, we further resample signals on a log-frequency scale to obtain a 256 $\times$ 256 T-F spectrograms, which is similar to the application of Mel-Frequency scale.

C. Evaluation Criteria

To evaluate the trained model, we use the model to separate the sound that is manually mixed from two different video samples. Quantitative analysis and qualitative analysis are conducted to assess our method.

In the testing process, we randomly choose two videos in the testing set and mix the sound of each video manually. Then, we use our model to separate the sound, and some sound separation metrics are applied to evaluate the performance. For each video sample representing human speech, we randomly select a video of vehicle type and mix the sound of two videos. We conduct 300 times of mixing totally, where 4500 pairs of people samples and vehicle samples are mixed each time. To quantitatively evaluate the effect of sound source separation, we used three common metrics: The

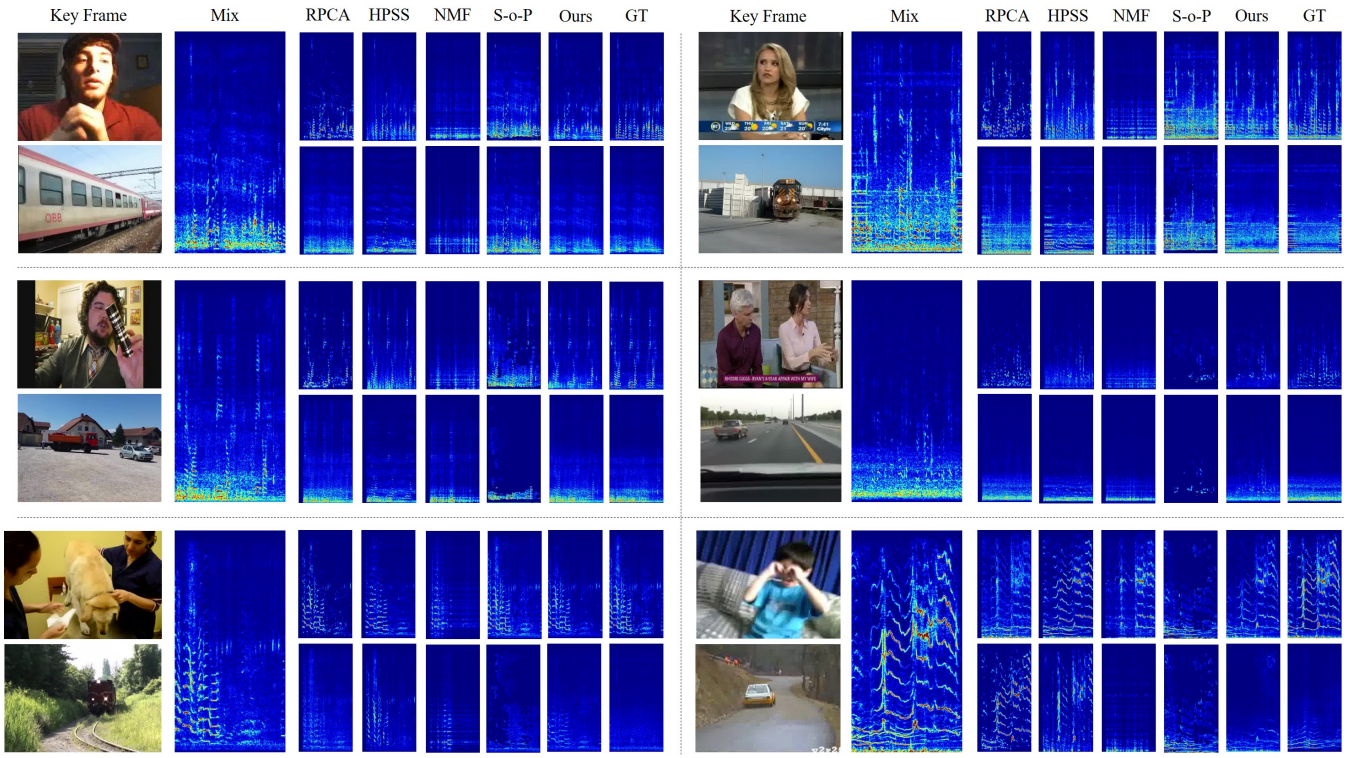

Fig. 4. Qualitative analysis results of several samples. We show the key frame of the mixed two videos, the spectrogram of mixed sound and the spectrogram of separated sound with different sound source separation methods. For each sample, the first column is the key frame of two videos. The third, forth and fifth columns are the separated spectrogram with RPCA, HPSS, and NMF methods respectively. The sixth column is the separated sound spectrogram with the Sound-of-Pixels model. The seventh column is the separated spectrogram with our model. The eighth column is the ground truth sound spectrogram, which is the sound spectrogram of each video's audio clip.

Source-to-Distortion Ratio (SDR), Source-to-Interferences Ratio (SIR) and Source-to-Artifacts Ratio (SAR) proposed in [33]. The metrics are calculated using The BSS-EVAL Toolbox [33].

With the ground truth of the sound, we can decompose the output sound into several parts: target, interferences, noise and artifacts

$$s = s_{target} + e_{interf} + e_{noise} + e_{artif}$$

SDR is the matrix to evaluate the extent of general distortion, SIR is used to evaluate the amount of interferences caused by other sources, while SAR is to evaluate the amount of artifacts errors terms. The definitions are as follows [33]:

The Source-to-Distortion Ratio (SDR):

$$SDR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2}.$$

The Source-to-Interferences Ratio (SIR):

$$SIR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2}.$$

The Source-to-Artifacts Ratio (SAR):

$$SAR = 10log_{10}\frac{\|s_{target} + e_{noise} + e_{interf}\|^2}{\|e_{artif}\|^2}.$$

From the definition, we can know that SIR and SDR can evaluate the accuracy of sound separation results. SAR measures the absence of artifacts, so actually SAR is not precise enough to reflect the separation accuracy [27].

The separated sound segment and predicted sound spectrogram are shown for qualitative analysis.

D. Quantitative Analysis

We compare the quantitative performance of our methods with several audio-only sound source separation baselines including RPCA, HPSS, and NMF methods. At the same time, we also compare our method with the sound source separation model proposed in [24], which utilizes visual-audio features. We refer to that model as *Sound-of-Pixels* in the following parts of this paper.

The results are shown in Table I. Compared with those audio-only sound separation baselines, our model achieves the highest in SDR. However, in terms of SIR, NMF and HPSS methods outperform our model. From the definition of SDR and SIR, it is noticed that SIR compares the separated sound with only the amount of interferences, while SDR compares the separated sound with the sum of all the unrelated signals including the amount of interferences, noise, etc. Hence, SDR is a better metrics reflecting the accuracy of the sound separation task. Although SIR of NMF and HPSS methods are higher, SDR of those methods are low, which indicates that there is a large noise in the separated target sound obtained by NMF and HPSS methods. Furthermore, in the part of qualitative analysis,

TABLE I
QUANTITATIVE ANALYSIS IN DIFFERENT METHODS

Methods	SAR	SIR	SDR
RPCA	4.3025	-0.1598	-4.2977
HPSS	9.4705	2.8645	0.1160
NMF	0.6633	4.6153	-4.2616
Sound-of-Pixels	6.6016	1.0348	-3.3530
Our Model	15.0934	1.2884	0.1342

we show the spectrogram of each separated sound with different sound separation methods. The results show that our model performs better in sound separation than other audio-only sound separation models. Our model can extract visual features, and those features can guide and support the process of sound separation to obtain better results.

In addition, our method also outperforms the visual-audio Sound-of-Pixels model in SAR, SIR and SDR. That pixel-level sound separation model utilizes the visual features of each pixel and the segmented audio features to accomplish the sound separation together. In fact, the model cannot know what each pixel represents in the real scene, and cannot build the connection between adjacent pixels or related pixels. Our model extracts the object-level visual features for every detected object. The object-level features contain more semantic information than the pixel-level features, hence it can better guide the sound source separation.

E. Qualitative Analysis

The qualitative results are shown in Fig. 4. We give 6 samples of the sound separation task in the testing set. For each sample, we show the separated sound spectrogram of each video with different sound separation methods mentioned in quantitative analysis. From those performance, we can obtain results as follows:

1) The performance of sound separation with our model is better than other methods, and the difference between separated spectrogram with our model and ground truth spectrogram is minimal.
2) The difference between the separated sound spectrogram with the Sound-of-Pixels model and ground truth is larger than that between our model's result and ground truth. The separated sound gained by the Sound-of-Pixels model lose some information of the origin sound sometimes.
3) The NMF method does not separate two sound effectively, and the two sound spectrograms are similar sometimes.
4) The performance of HPSS and RPCA method is a little better than NMF method, but still not as good as our method. The separated sound results of person speech with RPCA method lose information sometimes.

The qualitative results demonstrate the actual performance of sound separation with different models. Though SIR of NMF and HPSS methods is a little higher, the actual sound separation performance indicates they have worse separation results. Through comprehensively analyzing the experiment results in quantitative and qualitative ways, our model performs better than other methods.

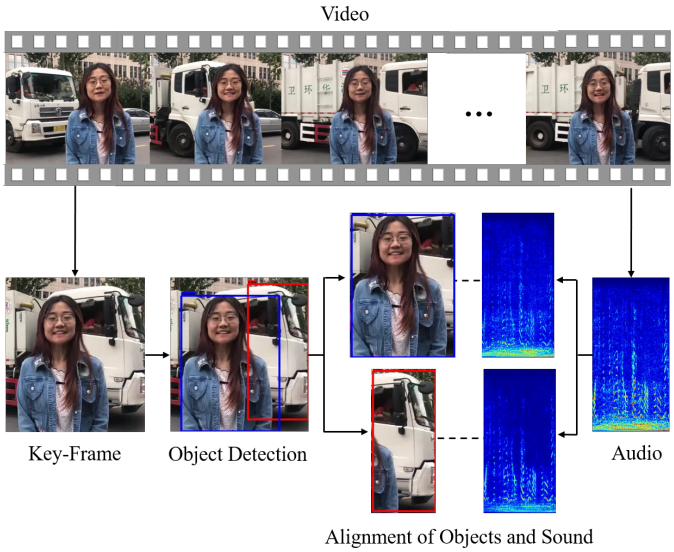

Fig. 5. Results in real-world scenarios. We show the results of object detection and sound separation tasks. Then, the detected objects and the separated audio clips are aligned.

F. Real-World Scenarios

To further illustrate the effectiveness of our model, we capture some real-life videos. In one of the videos we capture, a student is standing on the street reporting the weather condition, with a large noisy vehicle passing behind her. The sound of that vehicle's engine mixes with the speech of that student. With our model, the sound of that student can be extracted clearly, which is shown in Fig. 5. The detected student and that vehicle are marked with boxes. And the sound of the whole video is separated into 2 parts: the sound of the student and the sound of the vehicle. The results of separated audios are shown in the attached video.

VI. CONCLUSION

In this paper, we develop a self-supervised learning model to perform object detection and sound source separation tasks for videos in real-world scenarios simultaneously. The results of experiments on manually synthetic data and captured videos in real scenes indicate that the model can build alignment between objects and sounds by exploiting the visual and audio information at the same time. However, the model also have limitations in a more complex setting. For example, the proposed model cannot separate the sounds in instance-level. We will try to solve that task in the future.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants U1613212 and 91848206.

REFERENCES

[1] H. Liu, F. Sun, and X. Zhang, "Robotic material perception using active multimodal fusion," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9878–9886, 2018.

[2] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5610–5614.

[3] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially shared deep neural network in sound source separation and identification using a uav-embedded microphone array," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1299–1304.

[4] M. Frchette, D. Ltourneau, J.-M. Valin, and F. Michaud, "Integration of sound source localization and separation to improve dialogue management on a robot," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*, 2012, pp. 2358–2363.

[5] Q. Nguyen and authorJongSuk Choi, "Selection of the closest sound source for robot auditory attention in multi-source scenarios," in *Journal of Intelligent Robotic Systems*, 2016, pp. 239–251.

[6] R. Kojima, O. Sugiyama, R. Suzuki, K. Nakadai, and C. E. Taylor, "Semi-automatic bird song analysis by spatial-cue-based integration of sound source detection, localization, separation, and identification," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1287–1292.

[7] H. Liu, Z. Zhang, Y. Zhu, and S.-C. Zhu, "Self-supervised incremental learning for sound source localization in complex indoor environment," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2599–2605.

[8] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 74–79.

[9] F. Grondin and F. Michaud, "Noise mask for tdoa sound source localization of speech on mobile robots in noisy environments," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 4530–4535.

[10] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," in *The Journal of the Acoustical Society of America*, 1953, p. 975979.

[11] J. H. McDermott, "The cocktail party problem," in *Current Biology*, 2009, p. R1024R1027.

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct 2018.

[13] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," *Proc Icmc*, vol. 2003, 2003.

[14] ——, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[15] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[16] D. Yu, M. Kolbk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[17] H. Liu, D. Guo, X. Zhang, W. Zhu, and F. Sun, "Toward image-to-tactile cross-modal perception for visually-impaired people," *IEEE Transactions on Automation Science and Engineering*, DOI: 10.1109/TASE.2020.2971713.

[18] S. Ravulapalli and S. Sarkar, "Association of sound to motion in video using perceptual organization," in *18th International Conference on Pattern Recognition (ICPR'06)*, 2006, pp. 1216–1219.

[19] C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann, "Nonnegative cca for audiovisual source separation," in *2007 IEEE Workshop on Machine Learning for Signal Processing*, 2007, pp. 253–258.

[20] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[21] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 37, 2018.

[22] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic, "Audio-visual object localization and separation using low-rank and sparsity," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2901–2905.

[23] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[24] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[25] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2357–2361.

[26] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[27] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *ICCV*, 2019.

[28] F. Wang, D. Guo, H. Liu, J. Zhou, and F. Sun, "Sound-indicated visual object detection for robotic exploration," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8070–8076.

[29] J. Zhou, F. Wang, D. Guo, H. Liu, and F. Sun, "Video-guided sound source separation," *ICIRA 2019: Intelligent Robotics and Applications*, pp. 415–426, 2019.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015, pp. 234–241.

[32] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.