WEAKLY-SUPERVISED AUDIO-VISUAL SOUND SOURCE DETECTION AND SEPARATION

Tanzila Rahman[1,2] *and Leonid Sigal*[1,2,3]

[1]University of British Columbia [2]Vector Institute for AI [3]Canada CIFAR AI Chair

{trahman8, lsigal}@cs.ubc.ca

ABSTRACT

Learning how to localize and separate individual object sounds in the audio channel of the video is a difficult task. Current state-of-the-art methods predict audio masks from artificially mixed spectrograms, known as Mix-and-Separate framework. We propose an audio-visual co-segmentation, where the network learns both what individual objects look and sound like, from videos labeled with only object labels. Unlike other recent visually-guided audio source separation frameworks, our architecture can be learned in an end-to-end manner and requires no additional supervision or bounding box proposals. Specifically, we introduce weakly-supervised object segmentation in the context of sound separation. We also formulate spectrogram mask prediction using a set of learned mask bases, which combine using coefficients conditioned on the output of object segmentation — a design that facilitates separation. Extensive experiments on the MUSIC dataset show that our proposed approach outperforms state-of-the-art methods on visually guided sound source separation and sound denoising.

Index Terms— Co-segmentation, spectrogram, mix-and-separate framework, mask coefficient

1. INTRODUCTION

Multi-modal, visual and auditory, perception is an important research topic. Human brain has remarkable ability to isolate specific conversation from a noisy environment, as noted by Cherry through "cocktail party effect" [1]. At the same time, we can recognize objects and segment regions corresponding to those objects using our visual and auditory systems. We can also imagine how a particular, visually depicted, object may sound. Each object has unique physical properties, some of which can be visually observed, which leads it to generate a unique sound modulated by interactions with other objects and the environment. Therefore, working jointly with auditory and visual cues can be very useful for recognition of objects, localization of object regions and separation of sounds they make. Separating sounds of each object from a video has wide range of applications including audio denoising, hearing aids, automated transcription of speech and music, equalization, audio event remixing and dialog following.

Recent methods for audio-visual source separation [2, 3, 4] utilize "mix-and-separate" approach to train neural network architectures using self-supervision. The paradigm is simple, given a video, mix the audio track by combining audio channel with one from another video, and train the network to recover the original audio back, conditioned on the visual encoding of corresponding video content. This paradigm effectively synthesizes "cocktail party effect" by mixing clean sound(s) with others not present in the scene. While effective in training models for variety of tasks, such as sound source separation [2, 4] and on-/off-screen audio identification [5], this approach implicitly assumes that videos contain single-source sounds and attempts to correlate regions of the video with spectrograms [4]. Co-separation approach recently introduced by Gao *et al.* [2] addresses single-source limitation, but relies on object detectors trained with an external dataset annotated with bounding boxes for potential audible objects. In addition, while audio classes and corresponding spectrogram segmentations, that correspond to detected regions, are "discovered" during training, the model has no capacity to refine object detectors themselves to be optimal for sound-source separation task; *e.g.*, an entire object region is implicitly assumed to produce the sound.

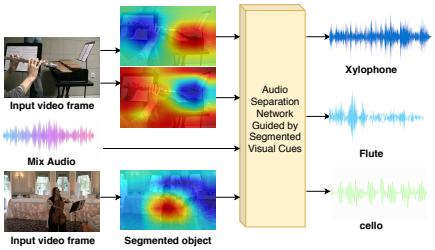

Fig. 1. **Visually-guided Sound Source Separation.** Our method first detects and segments object(s) which make sound(s), in a weakly-supervised manner, and then separates their respective audio signals.

Inspired by prior work, we aim to address aforementioned limitations. Specifically, we propose a weakly-supervised audio-visual detection and separation method. Our approach, similar to [2], does not assume single-source video; but, unlike [2], also does not rely on externally trained object detection module or object-level annotations of any kind. Instead, we leverage weak video-level labels to jointly learn visual and auditory segmentors that depend on one another. Our architecture has two paths: (1) a *video frame semantic segmenta-*

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

tion path designed to segment a frame into a set of regions using an attention mechanism that generates per-object-class attention map trained using weak frame-level classification objective; and (2) a *spectrogram mask prediction path* which takes both mixed spectrogram and pooled object-class image features and outputs a dense spectrogram mask with an objective to mask out the mixed-in sound. The spectrogram mask prediction branch is implemented using attention U-Net architecture [6], similar to [2, 4]. However, importantly, unlike prior methods, we train U-Net to produce a set of base masks from which a final mask is constructed using a set of sparse coefficients predicted from multi-modal audio-visual features. This architecture design takes inspiration from [7]. We find that such bi-linear decomposition is very useful in practice, allowing spectrograms to collaborate in learning a set of auditory sound bases, while relying on coefficient predictor to figure out how those should mix for a specific object type. Finally, despite having weaker supervision (no object annotations), compared to [2], we illustrate superior performance on the benchmark MUSIC and cross-dataset performance on AudioSet datasets.

Contributions. Our main contribution is an audio-visual co-segmentation approach for sound source separation, where the network learns both what individual objects look like and sound like, from videos labeled with only, one or more, object labels. This formulation and architecture has a number of appealing properties. Mainly, it does not assume single sound source input data, can be learned in an end-to-end manner, and requires no additional supervision or bounding box proposals. On the technical side, we introduce weakly-supervised object segmentation in the context of sound separation. We also formulate spectrogram mask prediction using a set of learned mask bases which are combined using sparse coefficients conditioned on multi-modal (visual object and auditory) features. Extensive experiments on the MUSIC dataset [4] show that our proposed approach outperforms state-of-the-art methods on visually guided sound source separation and sound denoising.

2. RELATED WORK

Audio-only sound source separation. Sound source separation is a challenging problem in speech processing and was first illustrated by "cocktail party effect" [1]. Classical approaches for the task include local Gaussian modeling [8], and Non-negative Matrix Factorization (NMF) [9]. Recently, deep learning based approaches [2, 3, 4] have gained popularity and most of recent methods use "Mix-and-Separate" framework to train the network by artificially mixing multiple audio streams first and then learning to separate each audio from the mixture. We also use mix-and-separate idea, but use visual features to guide audio separation.

Audio-visual source separation. Multi-modal learning has recently become a popular topic in the computer vision community. Auditory signal is used to supervise vision model during training in [10]. Similarly, in [11] visual features are used to guide sound models. Following [4, 12], we use audio-visual features to perform the separation. Unlike [2], we do not use any pre-trained object detector and propose an end-to-end approach to detect, localize and separate sound sources.

Weakly supervised visual learning. Given a video, our approach is able to detect which audio signals correspond to which objects and localize those objects within the video frames in a weakly-supervised manner. Earlier approaches address weakly-supervised object detection and segmentation using Multiple Instance Learning (MIL) [13]. More recently, pseudo-annotation generation [14] has gained popularity. Motivated by [15], we generate pseudo-annotations for weakly-supervised segmentation; and use these to visually guide the network to separate sound in an end-to-end fashion instead of using pre-trained object detectors [2].

3. APPROACH

We introduce a method for visually-guided sound separation, which leverages segmented object regions predicted to make sounds. In this section we first formalize our audio-visual sound source separation and detection task (Section 3.1) and then focus on describing the proposed deep neural network architecture for solving it (Section 3.2).

3.1. Problem Formulation

In this work, we use "Mix-and-Separate" framework [16, 2, 4], a well known approach for the task of sound source separation. The idea is to generate an artificially complex auditory signal by mixing multiple individual audio signals and learn to separate each individual sound of interest from the composition (see Figure 1 for illustration).

Given two input videos V_1 and V_2 with accompanying audio $A_1(t)$ and $A_2(t)$, respectively, we detect and segment objects, that make sound, from each video using a weakly supervised segmentation network. Then we generate a complex mixed auditory signal $A_m(t) = A_1(t) + A_2(t)$ by mixing two audio signals $A_1(t)$ and $A_2(t)$. Using a short-time Fourier transform (STFT) [17] with F-frequency bins, we transformed the mixed signal $A_m(t)$ into a magnitude spectrogram $A^M \in \mathbb{R}_+^{F \times N}$. A^M represents the change of frequency and phase over the time in mixed auditory signal. Suppose V_1 contains two objects O_1' and O_1'' and corresponding audios $A_1(t)'$ and $A_1(t)''$ accordingly. Similarly, V_2 contains one object O_2 with accompanying audio $A_2(t)$. Now our goal is to separate sounds $A_1(t)'$, $A_1(t)''$ and $A_2(t)$ of each detected object O_1', O_1'' and O_2 by predicting a spectrogram mask μ_n with the supervision of visual cues. To train the network one can use either ratio or binary mask and obtain object level magnitude spectrogram by $A_n = A^M \times \mu_n$. Finally, one can apply Inverse Short-Time Fourier transform (ISTFT) [17] to reconstruct object level wave-form sounds.

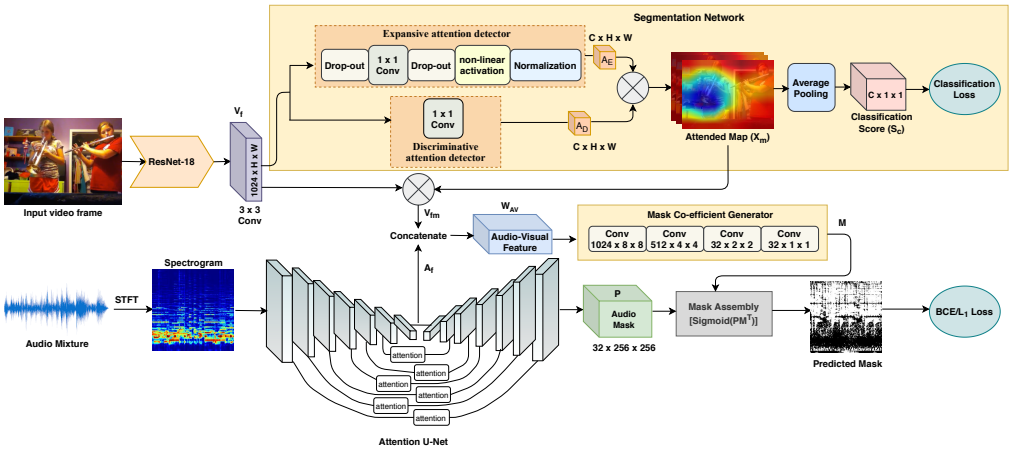

Fig. 2. **Weakly-supervised Audio-Visual Architecture.** ResNet-18 followed by a 3×3 convolution layer is used to extract visual feature ($\mathbf{V}_f$) from input video frames and fed to the segmentation network to detect the sound sources. Depending on the classification scores from the segmentation network, we generate soft semantic segmentations by producing class-specific attention map ($\mathbf{X}_m$). We use this attention map to pool features from respective image regions of $\mathbf{V}_f$ generating $\mathbf{V}_{fm}$. The resultant feature is concatenated with the bottleneck features of attention u-net to generate audio-visual feature ($\mathbf{W}_{AV}$). $\mathbf{W}_{AV}$ is passed to the mask coefficient generator to generate k mask coefficients ($\mathbf{M}$). At the same time, attention U-Net generates k audio channels ($\mathbf{P}$) and combined linearly ($\sigma(\mathbf{PM}^T)$) to predict final audio spectrogram mask guided by visual feature.

3.2. Weakly-supervised Audio-Visual Architecture

We propose a weakly-supervised audio-visual detection and separation architecture illustrated in Figure 2. Our architecture has two paths: (1) a *video frame semantic segmentation path* designed to detect objects that have potential to make sounds and segment them out in the frame, using an attention mechanism that generates per-object-class attention map, trained using weak frame-level classification objective (top block in yellow in Figure 2); and (2) a *spectrogram mask prediction path* which takes both mixed audio and pooled object-class image features and outputs a dense mask with an objective to mask out the mixed-in sound (bottom block, Figure 2).

We propose an end-to-end approach, unlike [2], to detect and segment objects from the input video frame. The input to our video frame segmenter is an RGB image/frame. The output is two fold – (i) a one-channel semantic segmentation attention map, per object class, that highlights regions where this object is present and (ii) probability of this object being present in the first place. Note, that (i) is only meaningful for objects that are present (probability of presence is high, above a threshold τ).

The spectrogram mask prediction path is trained to generate a (binary or real-valued) mask that masks-out the mixed-in sound. Prior approaches decode the multi-modal encoding of the mixed-audio and visual representation of attended frame [4], or an object region in the frame [2], into a mask directly. Instead, we utilize an attention U-Net architecture to first dynamically generate auditory mask bases from the mixed spectrogram itself. We then generate coefficients for these bases conditioned on the multi-modal features. The final mask is constructed as a coefficient-weighted combination of predicted bases. This decomposition allows shared learning of bases, and focuses visual conditioning on a few coefficients; this, we find, significantly improves the performance.

Video frame semantic segmentation. We use ResNet-18 [18] as backbone network followed by a 3×3 *convolution* to extract $H \times W$ spatial visual features $\mathbf{V}_f \in \mathbb{R}^{1024 \times H \times W}$ from the input video frame. These features are feed to the segmentation network to detect and segment objects. Following [15], our object detection network uses a decoupled spatial neural attention to detect and localize salient object regions simultaneously. The segmentation network contains two branches: (1) Expansive attention detector which identifies object regions and generates expansive attention map $\mathbf{A}_E \in \mathbb{R}^{C \times H \times W}$; and (2) Discriminative attention detector that predicts the discriminative parts and generates discriminative attention map $\mathbf{A}_D \in \mathbb{R}^{C \times H \times W}$. Expansive attention detector consists of a drop-out layer, 1×1 *convolution layer*, another drop-out layer, a non-linear activation layer (Eq. 1) and a spatial-normalization step (Eq. 2). Each element in $\mathbf{A}_E$ is defined as follows:

$$\alpha^c_{(i,j)} = F(\mathbf{W}_c^T \mathbf{V}_f(:,i,j) + b^c), \tag{1}$$

$$\alpha^c_{(i,j)} = \frac{\alpha^c_{(i,j)}}{\sum_i^H \sum_j^W \alpha^c_{(i,j)}}, \tag{2}$$

where $c \in \mathcal{C}$ and $F(\cdot)$ denote channel/class and non-linear activation respectively. Discriminative attention detector contains a 1×1 *convolution layer* and directly outputs a class-specific object attention map $\mathbf{A}_D$. We combine both attentions and generate final attention maps as follows: $\mathbf{X}_m = \mathbf{A}_E \odot \mathbf{A}_D$, where $\odot$ is the element-wise multiplication. Each depth channel of $\mathbf{X}_m$ is passed through a *spatial average pooling layer* to generate classification score for corresponding class; this results in $\mathbf{S} \in \mathbb{R}^{|\mathcal{C}|}$ class scores. Then we apply

a multi-label classification loss (*c-loss*) denoted as follows:

$$\mathcal{L}_{c-loss} = -\sum_{c}^{\mathcal{C}} \left[\mathbf{y}_c \log \frac{1}{1+e^{-\mathbf{S}_c}} + (1-\mathbf{y}_c) \log \frac{e^{-\mathbf{S}_c}}{1+e^{-\mathbf{S}_c}} \right], \tag{3}$$

where $\mathbf{y}_c$ denotes binary GT label for corresponding c-th class and $|\mathcal{C}|$ is the number of object classes.

Note that $\mathbf{X}_m$ can be interpreted as soft semantic segmentation (segmentation can be obtained by thresholding $\mathbf{X}_m^c$), with each channel corresponding to a specific object type. We can detect which objects are present, at test time, in a given video frame, by thresholding the classification scores $\mathbf{S}$.

Attention U-Net for audio processing. Motivated by [4], in this work, we use time-frequency representation of sound. Therefore, first we apply STFT on the input mixture sound to generate corresponding spectrogram. Then magnitude of spectrogram is transformed into log-frequency scale and used for further processing. Following [6], we use attention U-Net to extract audio features from the log magnitude of spectrogram. Attention U-Net uses attention gate (AG) to highlight discriminative features while passing through the skip connection. We use 7 convolutions (or down-convolutions) and 7 de-convolutions (or up-convolution) with skip connections in between for attention U-Net. The size of input spectrogram is $1 \times 256 \times 256$ and the final output of attention U-Net are audio mask bases ($\mathbf{P} \in \mathbb{R}^{k \times 256 \times 256}$) with k channels/bases. In this work, we use 32 as the value of k.

Mask Coefficient Generator. Following [7], the goal of mask coefficient generator is to predict k mask coefficients: $\mathbf{M} \in \mathbb{R}^k$. In this work, we use audio-visual feature to generate mask coefficient. Based on classification scores, $\mathbf{S}_c$, that are above a certain threshold, τ, from the segmentation network, we select corresponding class-specific attention channel(s) of $\mathbf{X}_m$ and apply weighted pooling on the visual feature $\mathbf{V}_f$ to generate attended visual feature for a corresponding object – $\mathbf{V}_{fm}$. The attended visual feature is concatenated with bottleneck U-Net feature, $\mathbf{A}_f$, to produce audio-visual feature vector $\mathbf{W}_{AV}$. $\mathbf{W}_{AV}$ is fed to the mask coefficient generator to predict k mask coefficient ($\mathbf{M}$). The mask coefficient generator consists of a series of *convolution layers* with non-linear activations and batch-normalization. In this paper, we use *ReLU* as non-linear activation function. We predict final magnitude of spectrogram, μ_A, by linearly combining k audio mask bases from $\mathbf{P}$ with the mask coefficient $\mathbf{M}$:

$$\mu_A = \sigma(\mathbf{P}\mathbf{M}^T). \tag{4}$$

The predicted magnitude of spectrogram μ_A is combined with the phase of input spectrogram. Then we use the inverse STFT to get a wave-form of the prediction. Our ultimate goal is to learn spectrogram masks of two types: *binary* or *ratio*. Following [4], in case of binary mask we use per-pixel sigmoid cross entropy loss (*i.e.*, BCE Loss, $\mathcal{L}_{BCE}$, to train the network). Similarly, per-pixel $\mathcal{L}_1$ loss [20] is used to train the network when we use ratio mask.

4. EXPERIMENTS

4.1. Datasets

MUSIC dataset. We evaluate our method using MUSIC dataset [4] which contains 685 untrimmed videos of musical solos and duets. We find that 31 videos are now missing from YouTube. The train/val/test split of MUSIC dataset is unavailable. Therefore, we follow the train/val/test split of [2] where the first/second video in each category is considered as the validation/test data, and the rest used for training data.

AudioSet-SingleSource. This is a small dataset, assembled in [21], which we only use for evaluation. The dataset consists of 15 musical instruments plus additional sounds produced by animals and vehicles. For our cross-dataset experiment we randomly select 11 out of 15 musical instruments for evaluation. Note, number of the instruments are *unseen* by the model – not in the MUSIC dataset that we use for training.

4.2. Pre-processing and implementation details

Following [4], to reduce the computational cost, we subsampled the audio signals to 11kHz and sample approximately 6 secs audio by random cropping from each untrimmed video. A Hann window size of 1022 and a hop length of 256 is used to compute STFT and generate a 512×256 Time-Frequency audio spectrogram which is further re-sampled on a log-frequency scale to obtain a 256×256 Time-Frequency representation. This representation is used as input to the attention U-Net. We obtain an output predicted mask and apply an inverse sampling step to convert the mask back to linear frequency scale of size 512×256 followed by an inverse STFT to recover wave-form signal. Following [2], we randomly sample 1-frame to train the model. To process the input video frame, we use ResNet-18 [18] pre-trained on ImageNet. We follow the experimental protocol of [4] and randomly sample 2 videos from MUSIC dataset to generate mixed audio for training and testing.

4.3. Sound source separation and detection

Evaluation Metrics. To measure performance we use three widely used metrics for sound separation: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). All the results are reported using widely used mir eval library [22]. The baselines used to quantitatively compare our results (in Table 1) are described in supplementary material.

Visually guided sound source separation. Table 1 shows quantitative evaluation of experimental results on MUSIC dataset, using both binary and ratio masks. We also include sound separation results with and without weakly-supervised segmentation network, as an ablation, to show the importance of that module in our architecture. We note that additionally removing the mask-coefficient component effectively reduces our model to the Sound-of-Pixels [4] baseline – the reason we do not include this variant. We note that improvements due to our decomposible construction of the mask are very signifi-

Table 1. Audio separation results on MUSIC test set. Performance reported using SDR/SIR/SAR. SDR and SIR capture separation accuracy; SAR only captures the absence of artifacts.

	Ratio Mask			Binary Mask		
Methods	SDR (↑)	SIR (↑)	SAR (↑)	SDR (↑)	SIR (↑)	SAR (↑)
NMF-MFCC [19]	0.92	5.68	6.84	-	-	-
AV-Mix-and-Separate [2]	3.23	7.01	9.14	-	-	-
Sound-of-Pixels [4]	7.81	11.06	14.05	7.26	12.25	11.11
CO-SEPARATION [2]	7.64	13.8	11.3	-	-	-
Ours (Mask coefficient)	8.40	12.53	14.11	9.25	**15.98**	**12.45**
Ours (Mask coefficient + Seg. Net)	**9.14**	**13.35**	**14.18**	**9.29**	15.09	12.43

Table 2. Multi-label object classification accuracy. Performance in (%) on the MUSIC test set.

Threshold value(τ)	0.1	0.2	0.3	0.4	0.5
Binary mask	80.30	91.41	93.18	92.68	88.89
Ratio mask	**83.08**	**91.92**	**93.69**	**93.18**	**89.65**

Table 3. Cross dataset evaluation of audio separation. Evaluation of the model trained using the MUSIC dataset on the AudioSet-SingleSource dataset; using ratio mask. SAR is responsible for capturing absence of artifacts, therefore, it can be higher even when separation results are poor.

Methods	SDR (↑)	SIR (↑)	SAR (↑)
Sound-of-Pixels [4]	0.72	20.14	**16.10**
Ours(Mask coeff.)	5.11	26.57	13.95
Ours(Mask coeff. + Seg. Net)	**7.19**	**29.98**	12.15

cant (7.26 vs. 9.25 in SDR using binary mask). The improvements due to weakly-supervised detection and segmentation is slightly more modest (8.40 vs. 9.14 in SDR using ratio mask) but are still substantial. Consistent with [2], we find SDR and SIR metric to be most informative.

Figure 3 shows corresponding qualitative results. The first and second rows illustrate randomly sampled video mixture pairs and corresponding spectrograms of the mixed sound. The third and fourth rows show ground truth and predicted separated spectrograms. Finally fifth row illustrated predicted spectrogram generated by running pre-train model from [2][1]. One can clearly see that our method outperforms the state-of-the-art [2] in both quality and sharpness of resulting spectrograms. See supplementary material for additional ablations.

Sound object detection and segmentation. Our object detection and segmentation utilizes a weakly-supervised network. Importantly, in addition to weakly-supervised loss, audio separation pathway, that depends on the resulting segmentations, provides additional regularization. We measure accuracy of our object detection network by computing multi-class classification accuracy on the MUSIC test set, as reported in Table 2 as a function of the threshold τ. Results illustrate that we can achieve high accuracy of up to 93.69% and that regularization with ratio mask variant of the audio network is consistently better for visual object detection. We visualize segmentation localization qualitatively (dataset does not contain spatial annotations for quantitative analysis) in Figure 4.

[1] https://github.com/rhgao/co-separation

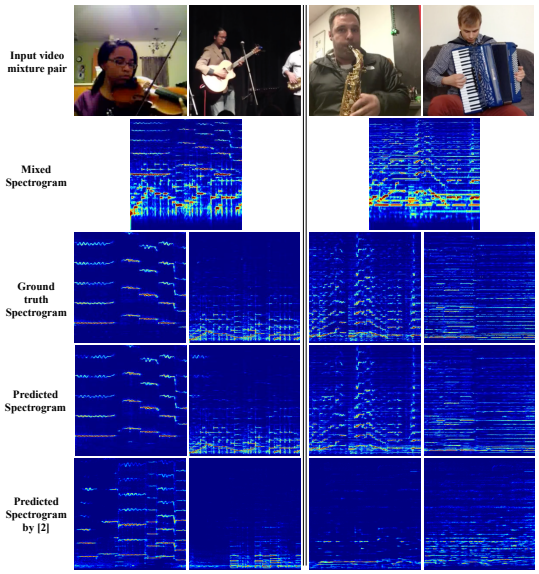

Fig. 3. Qualitative audio separation results on MUSIC test set. Test samples, our results and comparison with [2] are shown. See text for details and discussion.

Cross-dataset experiments. We also perform cross dataset testing to evaluate the generality of our method. We do so by measuring the performance of our proposed model, trained on MUSIC dataset, by applying it on the AudioSet-SingleSource dataset. The results are presented in Table 3. Note the nearly $10\times$ performance increase in SDR as compared to [4].

Audio separation for unseen objects. We also conduct a small experiment to see how the models perform for separating objects/instruments that the model has not seen during training. The results are presented in Table 4. Here, the model never seen some instruments (*e.g.*, Banjo, Marimba) during training on MUSIC dataset but evaluated on those instruments from AudioSet-SingleSource dataset. In this case the model is relying on similarity of novel instruments to those used in training our model.

5. CONCLUSION

In this paper, we introduce an end-to-end audio-visual co-segmentation network to separate and detect sound source without requiring additional supervision or bounding box proposal and solve the problem in a weakly supervised manner from large-scale unlabeled videos. Moreover, our mask co-

Table 4. **Audio separation for unseen objects.** Toy experiment with cross dataset setting where the model never seen some instruments during training on MUSIC dataset.

	Sound-of-Pixels [4]			Ours		
Instruments	SDR (↑)	SIR (↑)	SAR (↑)	SDR (↑)	SIR (↑)	SAR (↑)
Banjo/Electric Guitar	0.03	0.08	**24.82**	**1.08**	**2.16**	12.20
Saxophone/Marimba	3.64	5.30	11.07	**12.19**	**19.97**	**16.00**
Cello/Electric Guitar	0.79	0.81	**28.20**	**2.07**	**3.57**	9.33

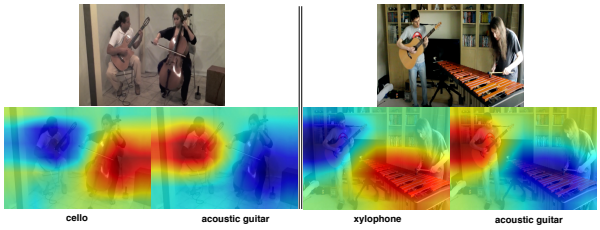

Fig. 4. **Attended object map.** Attended maps (red higher attention) that correspond to object classes (in bottom). Result from our learned weakly-supervised segmentation network.

efficient generator facilitates separation conditioned on the output from the segmentation network. Both quantitative and qualitative results show the effectiveness of our proposed method compared to the existing state-of-the-art methods for sound source separation.

6. REFERENCES

[1] E Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *JASA*, vol. 25, no. 5, pp. 975–979, 1953.

[2] Ruohan Gao and Kristen Grauman, "Co-separating sounds of visual objects," in *ICCV*, 2019, pp. 3879–3888.

[3] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba, "The sound of motions," in *ICCV*, 2019, pp. 1735–1744.

[4] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *ECCV*, 2018, pp. 570–586.

[5] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018, pp. 631–648.

[6] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv*, 2018.

[7] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee, "Yolact: real-time instance segmentation," in *ICCV*, 2019, pp. 9157–9166.

[8] Derry Fitzgerald, Antoine Liutkus, and Roland Badeau, "Projet—spatial audio separation using projections," in *ICASSP*, 2016, pp. 36–40.

[9] Jonathan Le Roux, John R Hershey, and Felix Weninger, "Deep nmf for speech separation," in *ICASSP*, 2015, pp. 66–70.

[10] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba, "Ambient sound provides supervision for visual learning," in *ECCV*, 2016, pp. 801–816.

[11] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016, pp. 892–900.

[12] Xudong Xu, Bo Dai, and Dahua Lin, "Recursive visual sound separation using minus-plus net," in *ICCV*, 2019, pp. 882–891.

[13] Pedro O Pinheiro and Ronan Collobert, "From image-level to pixel-level labeling with convolutional networks," in *CVPR*, 2015, pp. 1713–1721.

[14] Tong Shen, Guosheng Lin, Lingqiao Liu, Chunhua Shen, and Ian D Reid, "Weakly supervised semantic segmentation based on co-segmentation.," in *BMVC*, 2017.

[15] Tianyi Zhang, Guosheng Lin, Jianfei Cai, Tong Shen, Chunhua Shen, and Alex C Kot, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2930–2941, 2019.

[16] Ruohan Gao and Kristen Grauman, "2.5 d visual sound," in *CVPR*, 2019, pp. 324–333.

[17] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] Martin Spiertz and Volker Gnann, "Source-filter based clustering for monaural blind source separation," in *DAFx*, 2009.

[20] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[21] Ruohan Gao, Rogerio Feris, and Kristen Grauman, "Learning to separate object sounds by watching unlabeled video," in *ECCV*, 2018, pp. 35–53.

[22] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel PW Ellis, "mir_eval: A transparent implementation of common mir metrics," in *ISMIR*, 2014.