Co-Separating Sounds of Visual Objects

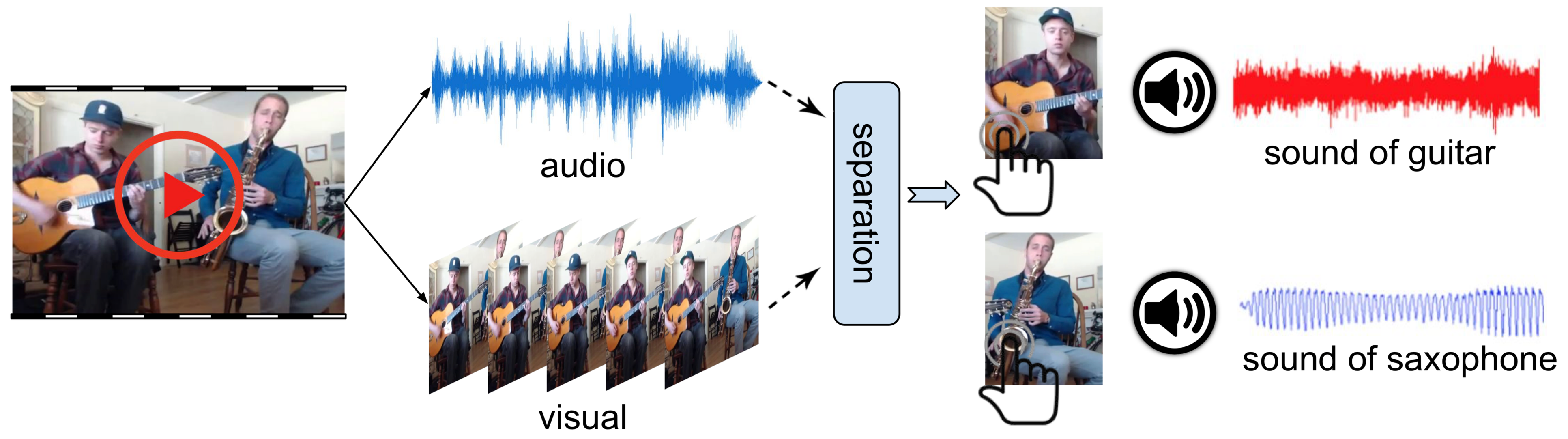
Ruohan Gao[1] Kristen Grauman[1,2]

[1]The University of Texas at Austin, [2]Facebook AI Research

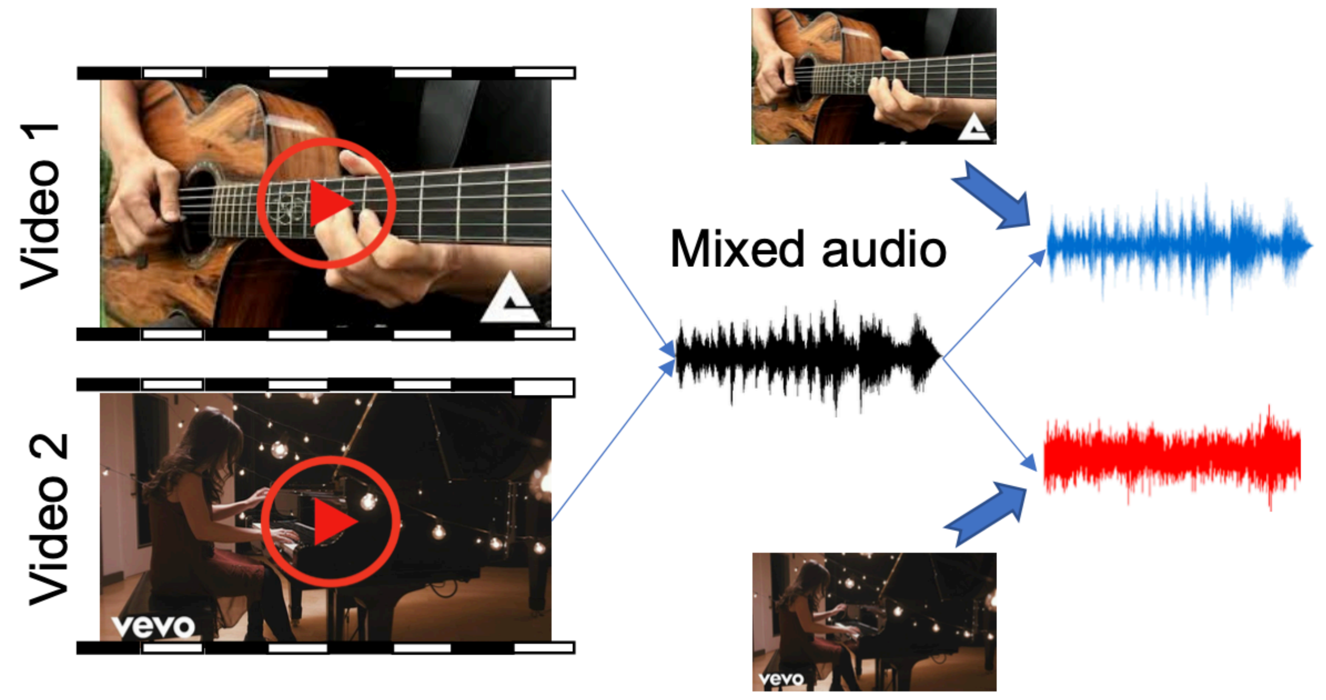
ICCV 2019 Seoul, Korea

Audio-Visual Source Separation

Goal: audio-visual object source separation in videos



Current approaches: Mix-and-Separate



Simpson et al. 2015; Huang et al. 2015; Yu et al. 2017; Ephrat et al. 2018; Owens & Efros 2018; Zhao et al. 2018; Afouras et al. 2018; Gao & Grauman 2019; Zhao et al. 2019

Limitations:

- Require single-source training clips
- Assume sources in a recording are independent

Motivation: Image Co-Segmentation

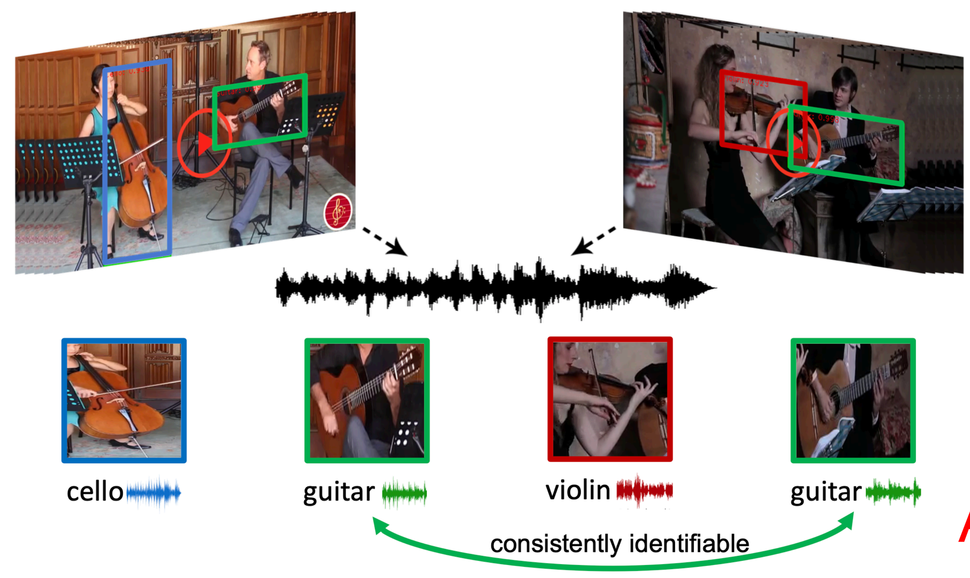
Input image pair Co-segmentation

Jointly segmenting two related images can be easier than segmenting them separately

Rother et al. CVPR 2006

Our Idea: Co-Separation

Co-separation: separate sounds for pairs of training videos

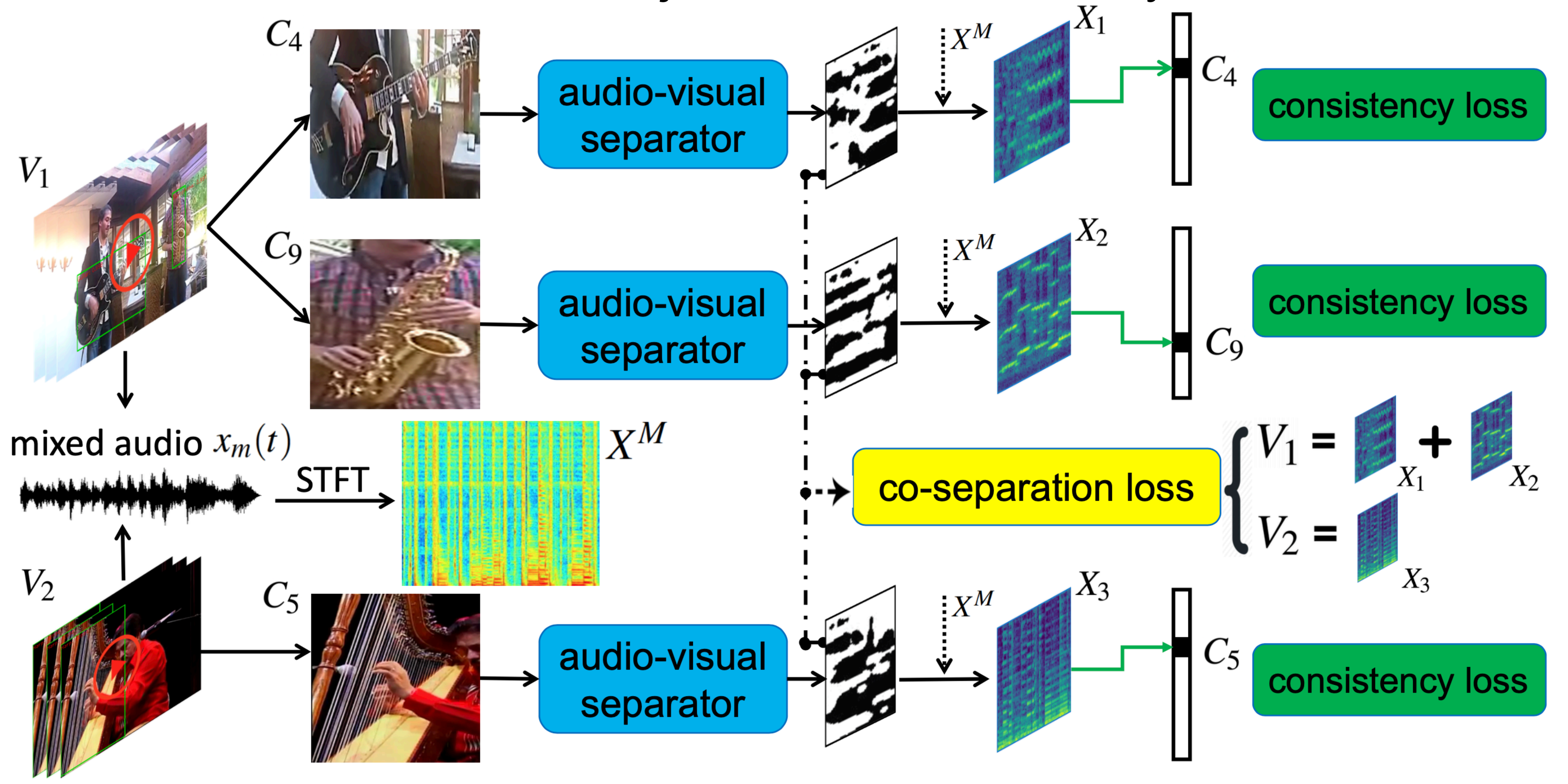


Co-separate the sounds in multi-source training samples by learning to associate similar sounds with detected objects.

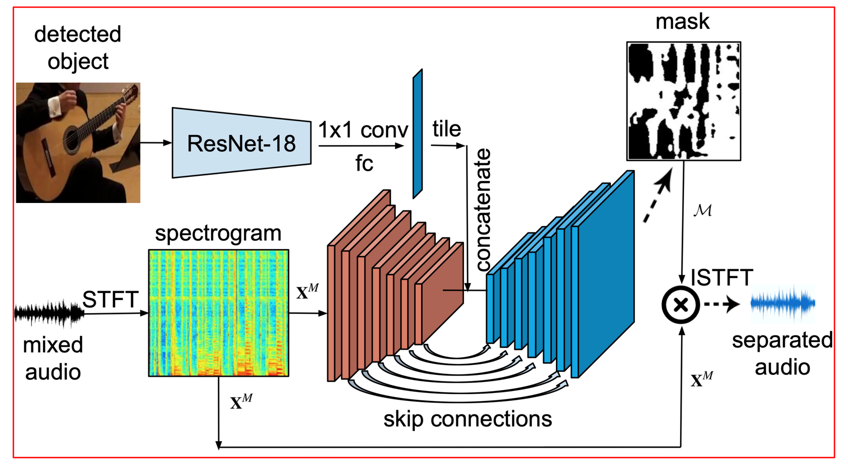
At test time, input is single video.

Training paradigm:

We detect objects in a pair of videos, and require separated sounds from detected objects to be consistently identifiable.



Audio-visual separator→



- Co-separation loss:

$$L_{co\text{-}separation\_spect} = \Big\|\sum_{i=1}^{|\mathcal{V}_1|} X_i - X^{V_1}\Big\|_1 + \Big\|\sum_{i=1}^{|\mathcal{V}_2|} X_i - X^{V_2}\Big\|_1$$

- Object-consistency loss:

$$L_{object\text{-}consistency} = \frac{1}{|\mathcal{V}_1|+|\mathcal{V}_2|}\sum_{i=1}^{|\mathcal{V}_1|+|\mathcal{V}_2|}\sum_{c=1}^{C} -y_{i,c}\log(p_{i,c})$$

- Final objective: $L = L_{co\text{-}separation} + \lambda L_{object\text{-}consistency}$

Experimental Results

Datasets:

MUSIC (Zhao et al. 2018, 536 solos and 149 duet videos, 11 categories)

AudioSet-Unlabeled (Gemmeke et al. 2017, >100k clips of 15 categories)

MIT MUSIC

	Single-Source		Multi-Source	
	SDR	SIR	SDR	SIR
Sound-of-Pixels (Zhao *et al.* 2018)	7.30	11.9	6.05	9.81
AV-Mix-and-Separate	3.16	6.74	3.23	7.01
NMF-MFCC (Spiertz *et al.* 2009)	0.92	5.68	0.92	5.68
CO-SEPARATION (Ours)	**7.38**	**13.7**	**7.64**	**13.8**

AudioSet-Unlabeled

	SDR	SIR
Sound-of-Pixels (Zhao *et al.* 2018)	1.66	3.58
AV-MIML (Gao *et al.* 2018)	1.83	-
AV-Mix-and-Separate	1.68	3.30
NMF-MFCC (Spiertz *et al.* 2009)	0.25	4.19
CO-SEPARATION (Ours)	**4.26**	**7.07**

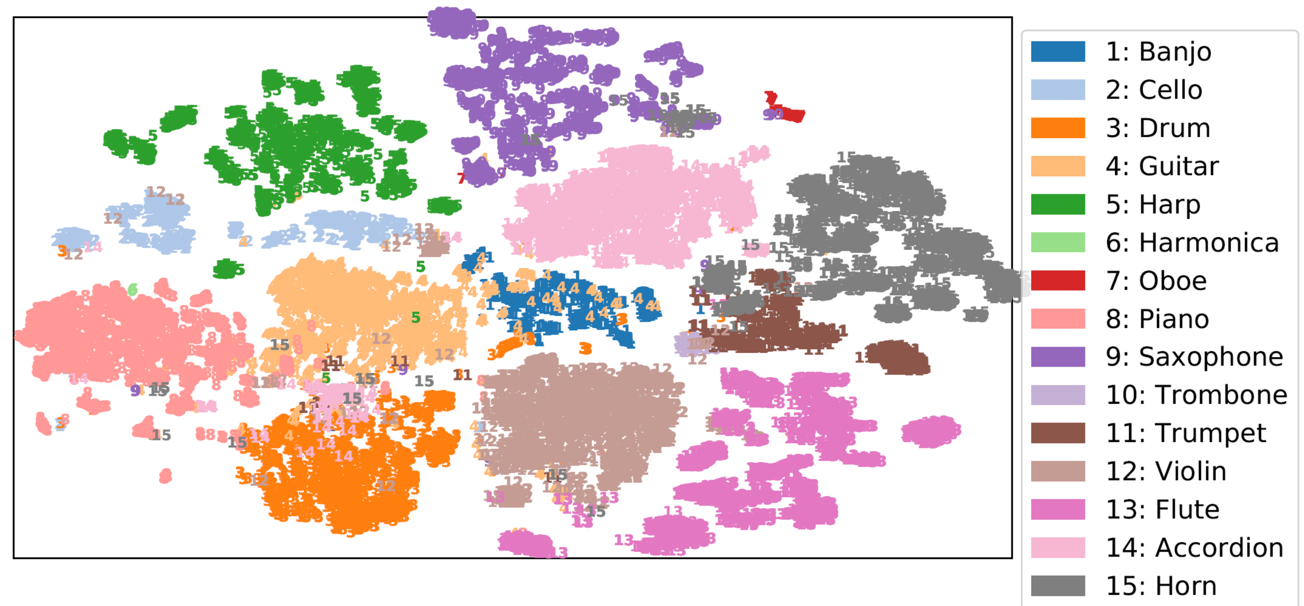
What if we train with only duets?

	Sound-of-Pixels		Ours	
	SDR	SIR	SDR	SIR
Violin/Saxophone	1.52	1.48	8.10	11.7
Violin/Guitar	6.95	11.2	10.6	16.7
Saxophone/Guitar	0.57	0.90	5.08	7.90

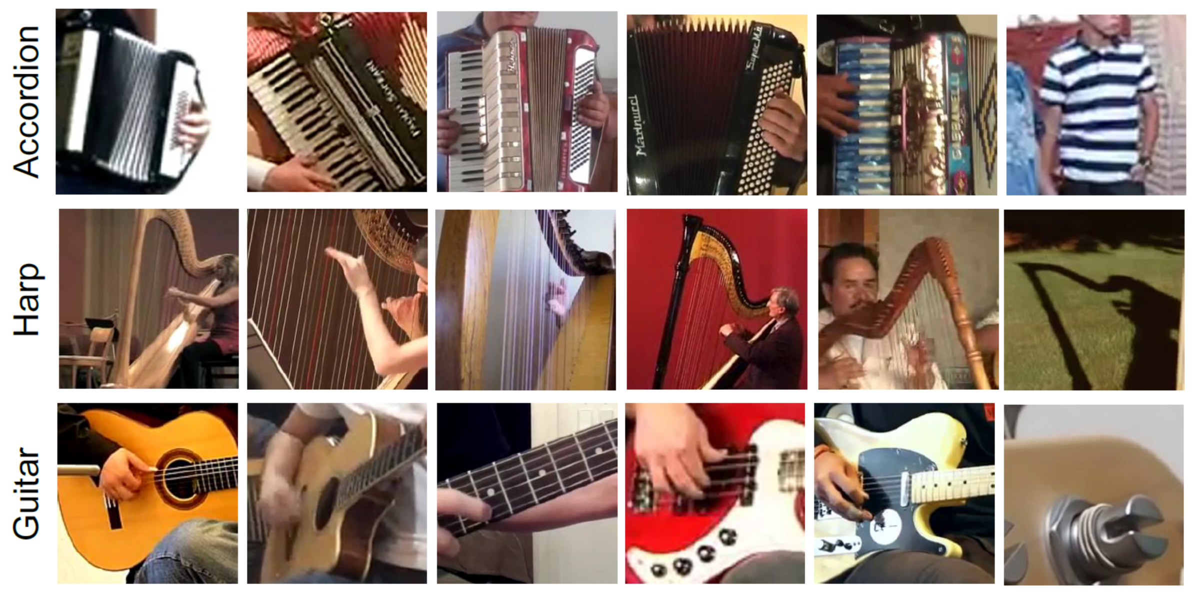
Co-separation overcomes the limitation of mix-and-separate when presented with multi-source training videos.

Discover object sounds:

Trained with multi-source videos, our learned audio embedding discovers object sounds in AudioSet.



Localize what is heard:



Object proposals associated with highest confidence scores

Please visit our project page for video results:

vision.cs.utexas.edu/projects/coseparation/