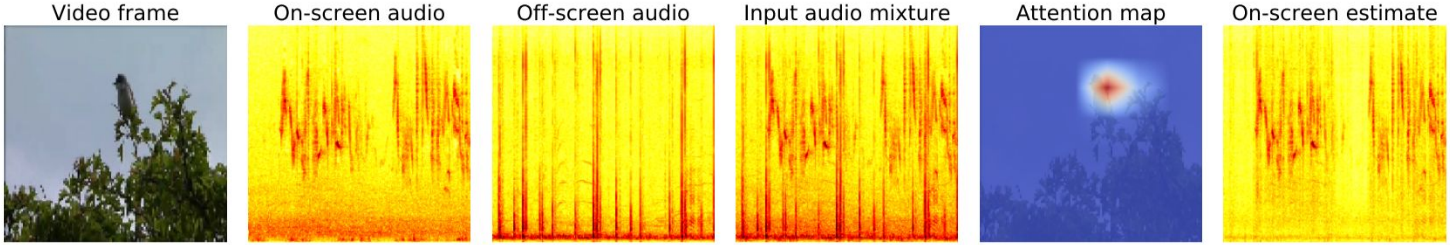
# Into the Wild with AudioScope: Unsupervised Audio-Visual Separation of On-Screen Sounds




**(Resized still with or without overlaid attention map from "Whitethroat" by S. Rae, license: CC BY 2.0.)**

*Efthymios Tzinis*[1,2*], *Scott Wisdom*[2], *Aren Jansen*[2], *Shawn Hershey*[2], *Tal Remez*[2], *Daniel P. W. Ellis*[2], *John R. Hershey*[2]







ICLR

*International Conference on Learning Representations 2021*

1

2 Google Research

* Work done during an internship at Google.

# Ideally we want to automatically separate all types of sounds which appear on screen

**Goal: Capturing sounds which are present on-screen**

Video clip from ["Luchador and Yellow Jumpsuit" by tenaciousme](), license: [CC-BY 2.0]().

# Ideally we want to automatically separate all types of sounds which appear on screen

**Goal: Capturing sounds which are present on-screen**

**Conventional recipe: Train a separation system:**

- Find good data to train with...

**Sound separation in-the-wild:**

**Not easy (nearly impossible) to gather supervised data**

Video clip from "Luchador and Yellow Jumpsuit" by tenaciousme, license: CC-BY 2.0.

# Ideally we want to automatically separate all types of sounds which appear on screen

**Goal: Capturing sounds which are present on-screen**

**Conventional recipe: Train a separation system:**

- Find good data to train with...

**Sound separation in-the-wild:**

- **Not easy (nearly impossible) to gather supervised data**

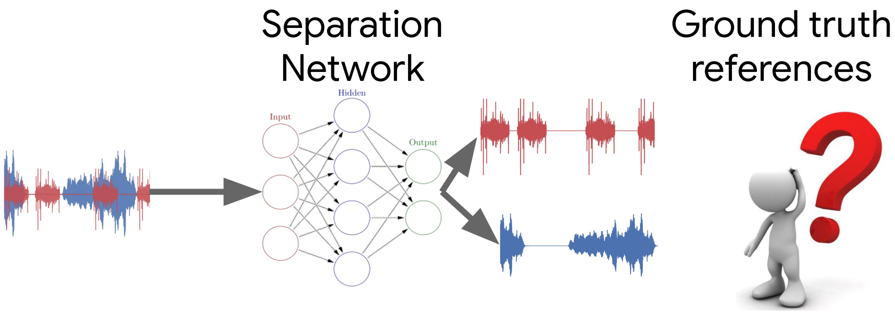
**AudioScope overcomes limitations of prior work:**

- No dependence on **object detection** systems.
- No assumption about **number/class** of sounds.
- No assumption about training with **strictly on-screen-only** mixtures

Video clip from "Luchador and Yellow Jumpsuit" by tenaciousme, license: CC-BY 2.0.

# Our recipe

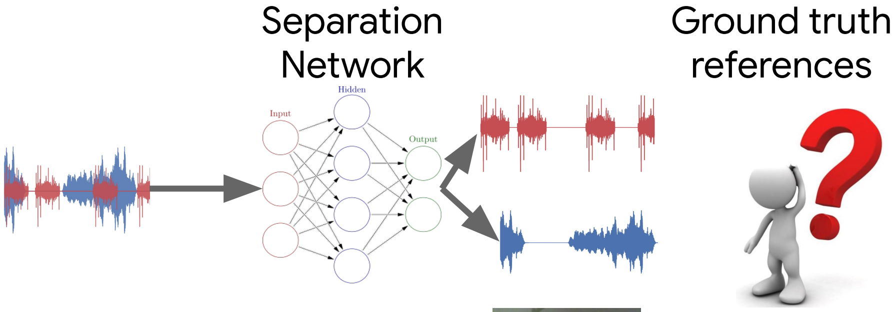
A. Make our sound separation network work with **in-the-wild mixtures** (no ground-truth sources).

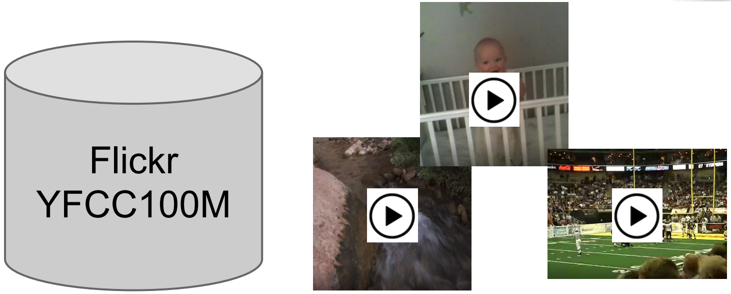


# Our recipe

A. Make our sound separation network work with **in-the-wild mixtures** (no ground-truth sources).

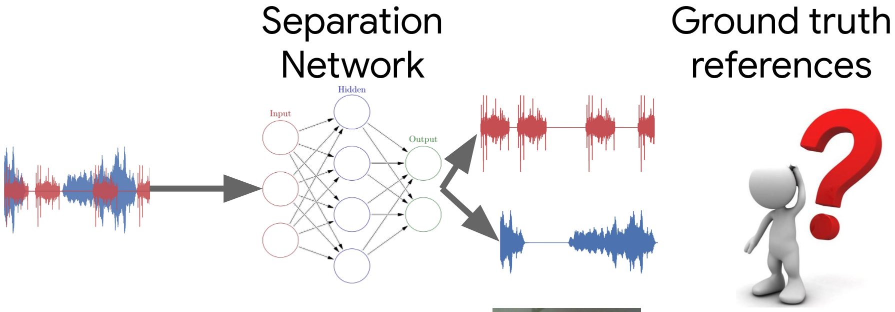
B. Develop a **dataset** with in-the-wild videos with on-screen and off-screen sounds.



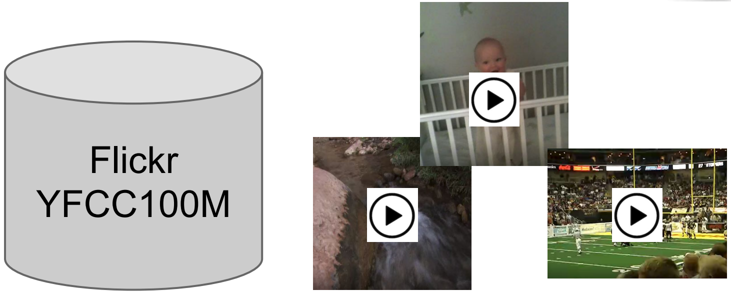



# Our recipe

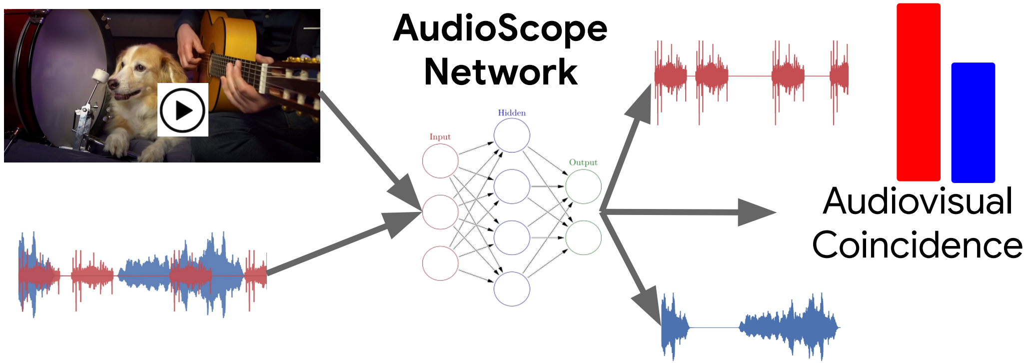
A. Make our sound separation network work with **in-the-wild mixtures** (no ground-truth sources).




B. Develop a **dataset** with in-the-wild videos with on-screen and off-screen sounds.

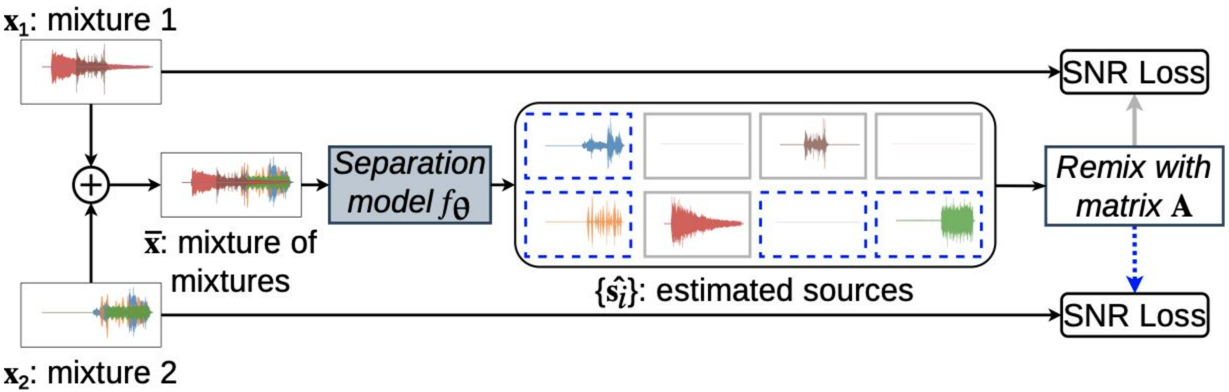



C. Train an **audio-visual coincidence classifier** using self-supervision on audio mixtures from videos.



# A. MixIT (Mixture Invariant Training)
## Unsupervised single-channel audio source separation

- **A self-supervised approach to source separation**
  - Requires **only acoustic mixtures**, not isolated sources.
  - **Competitive with supervised training** in some scenarios.
  - We base our approach on the effectiveness of MixIT




Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R. J., Wilson, K., and Hershey, J. R., "Unsupervised Sound Separation Using Mixtures of Mixtures." in Advances of Neural Information Processing Systems (NeurIPS) 2020.

# B. In-the-wild videos dataset

**YFCC100m videos:**

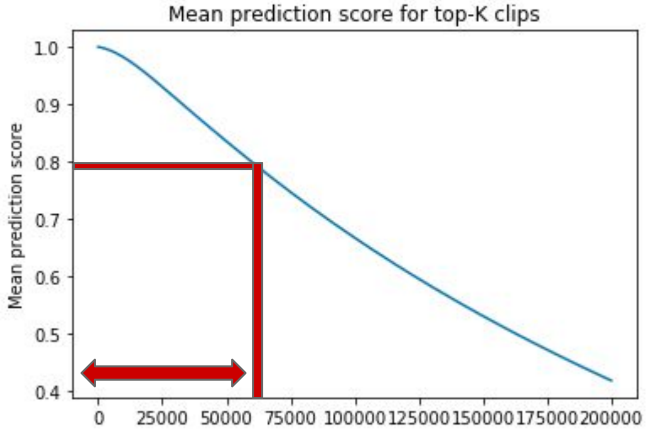
- 200,000 videos (2500 hours)
- Diverse set of sound classes

**Weak annotation of open-domain dataset:**

- Using an unsupervised coincidence model [Jansen et al. 2020], we filtered for video clips with likely on-screen sounds.
- According to weak annotations, ~30% of clips do not contain on-screen sounds / **low audio-visual correspondence.**

**Human annotation of data**

- 40,000 clips (55.5 hours)


Top coincidence videos ~36,000

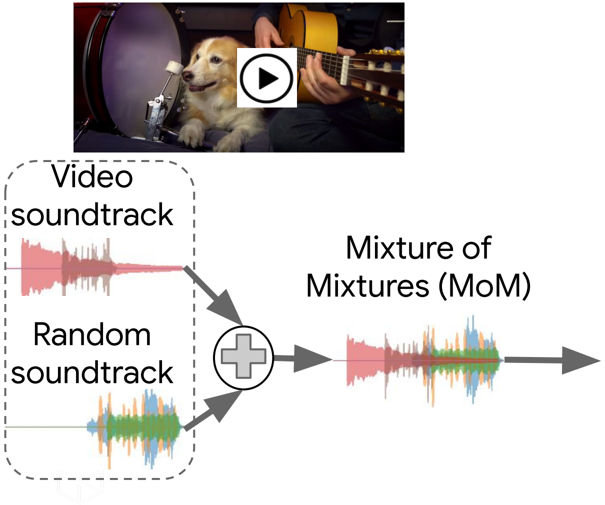
All videos ~200,000

Jansen, A., Ellis, D.P., Hershey, S., Moore, R.C., Plakal, M., Popat, A.C. and Saurous, R.A., 2020, May. Coincidence, Categorization, and Consolidation: Learning to Recognize Sounds with Minimal Supervision. In Proc. ICASSP 2020

# C. Audioscope audio-visual separation and coincidence prediction

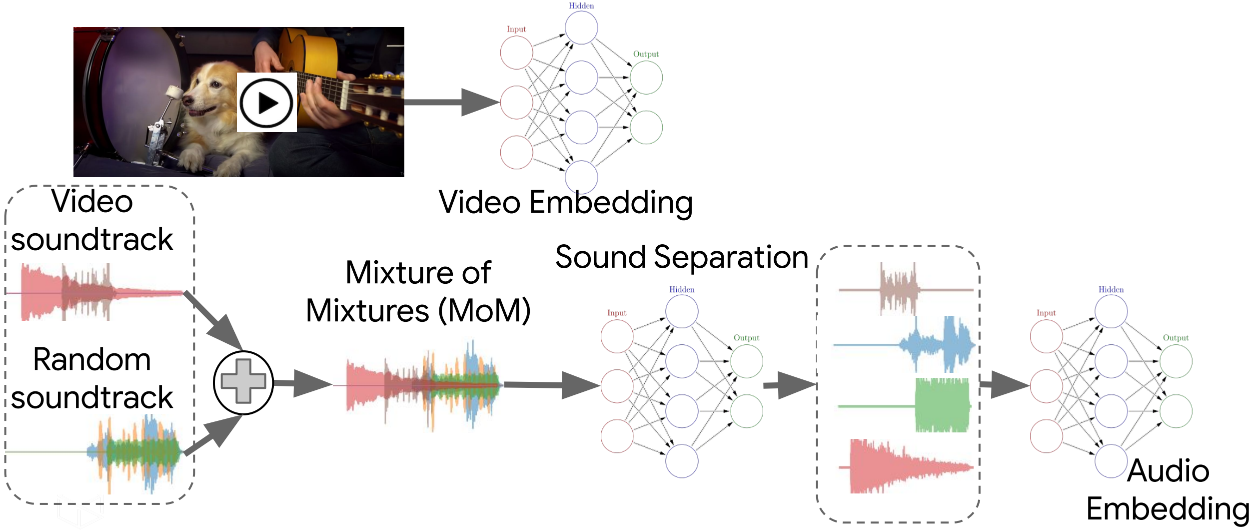
Video soundtrack

# C. Audioscope audio-visual separation and coincidence prediction



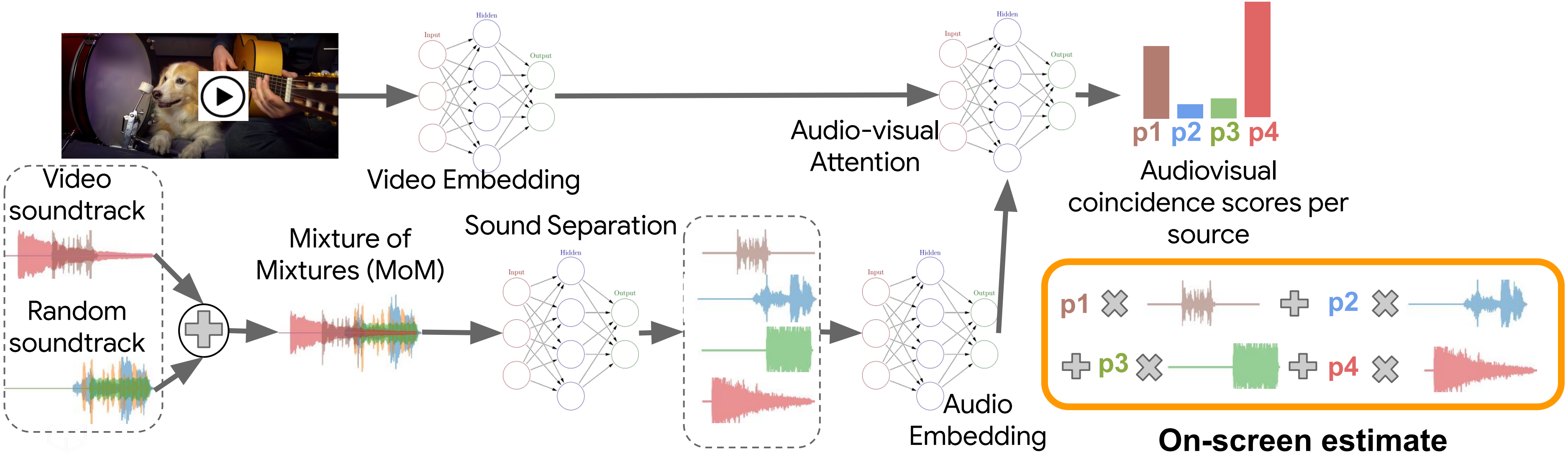
# C. Audioscope audio-visual separation and coincidence prediction

- Separate the sounds and compute video/audio embeddings.



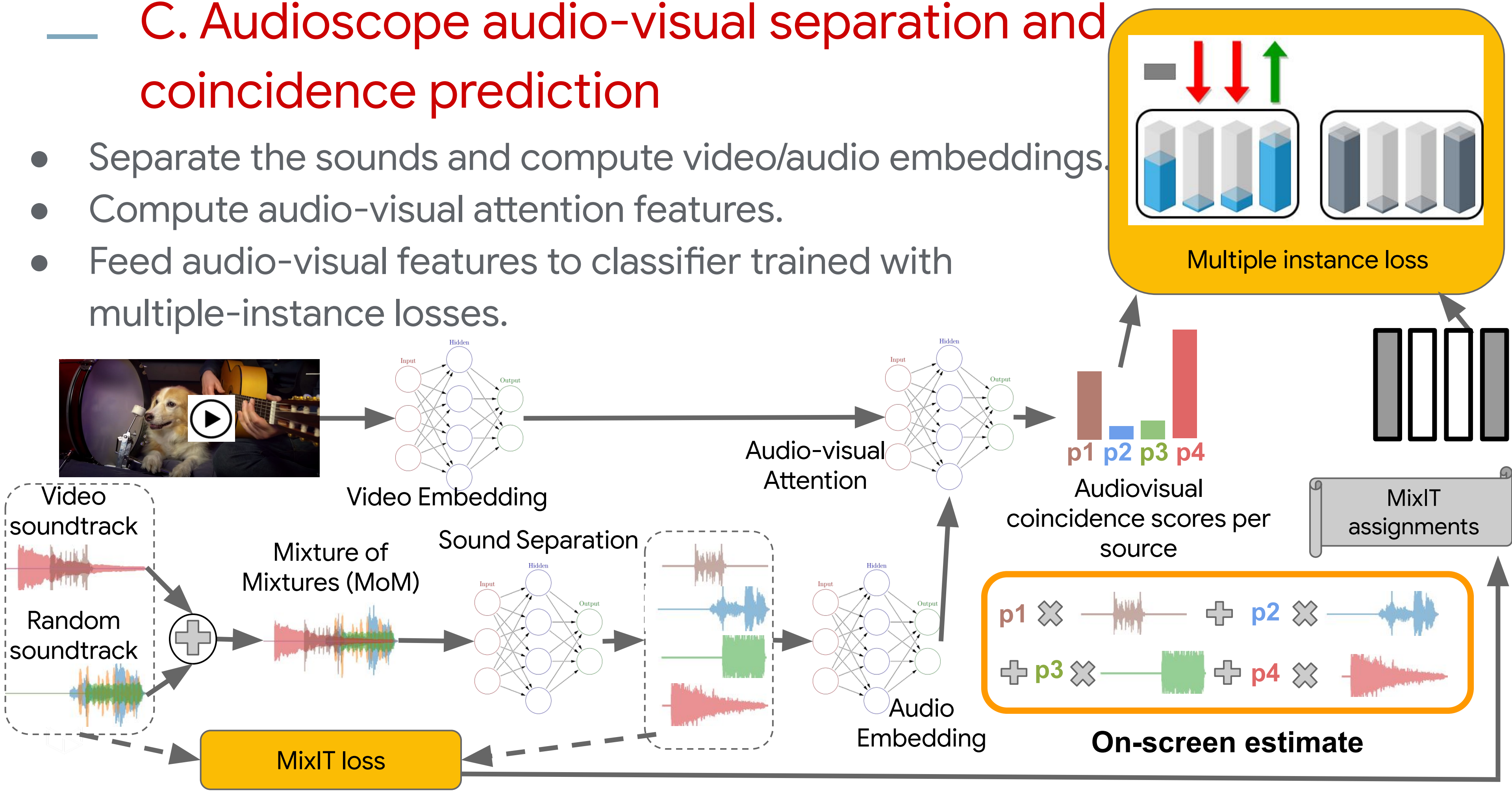
# C. Audioscope audio-visual separation and coincidence prediction

- Separate the sounds and compute video/audio embeddings.
- Compute audio-visual attention features.



# C. Audioscope audio-visual separation and coincidence prediction

- Separate the sounds and compute video/audio embeddings.
- Compute audio-visual attention features.
- Feed audio-visual features to classifier trained with multiple-instance losses.



# Separation results

- Unsupervised training achieves good performance on the in-the-wild on-screen sound separation task.

| | Single mixture | | | Synthetic mixtures of mixtures | | |
|---|---|---|---|---|---|---|
| Supervision | AUC | On-screen reconstruction SI-SNR | Off-screen power suppression | AUC | On-screen reconstruction SI-SNR | Off-screen power suppression |
| Unsupervised | 0.58 | 13.5 dB | 2.5 dB | 0.77 | 6.3 dB | 9.4 dB |
| Semi-supervised | 0.71 | 14.8 dB | 6.6 dB | 0.82 | 6.1 dB | 14.1 dB |
| Relative change | **+22%** | **+10%** | **+64%** | **+7%** | -3% | **+50%** |

# Separation results

- Unsupervised training achieves good performance on the in-the-wild on-screen sound separation task.
- Semi-supervised training significantly further boosts the performance.
    - Small amount of labeled data (~1%) leads to better results in terms of **detection**, **on-screen reconstruction**, and **off-screen suppression**

| | Single mixture | | | Synthetic mixtures of mixtures | | |
|---|---|---|---|---|---|---|
| Supervision | AUC | On-screen reconstruction SI-SNR | Off-screen power suppression | AUC | On-screen reconstruction SI-SNR | Off-screen power suppression |
| Unsupervised | 0.58 | 13.5 dB | 2.5 dB | 0.77 | 6.3 dB | 9.4 dB |
| Semi-supervised | 0.71 | 14.8 dB | 6.6 dB | 0.82 | 6.1 dB | 14.1 dB |
| Relative change | **+22%** | **+10%** | **+64%** | **+7%** | -3% | **+50%** |

# Synthetic mixture example

- Corresponding soundtrack: Bird chirping + wind noise
  - **Only bird appears on-screen**
- Random soundtrack: Fireworks + human laugh

**Input Video**

Video clip from "Whitethroat" by S. Rae, license: CC-BY 2.0, with additional background audio clip from video by deeje, license: CC-BY-SA 2.0.

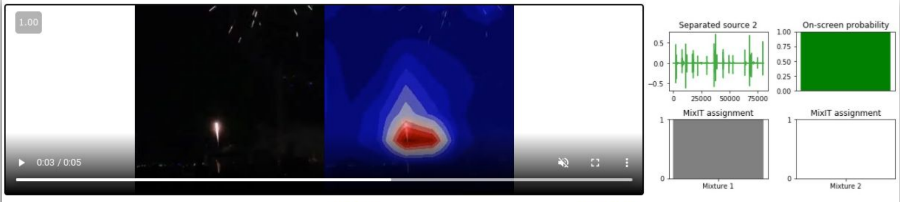
**On-screen estimate from the input video & corresponding attention map**

Video clip from "Whitethroat" by S. Rae with modified audio and overlaid attention map, license: CC-BY 2.0.

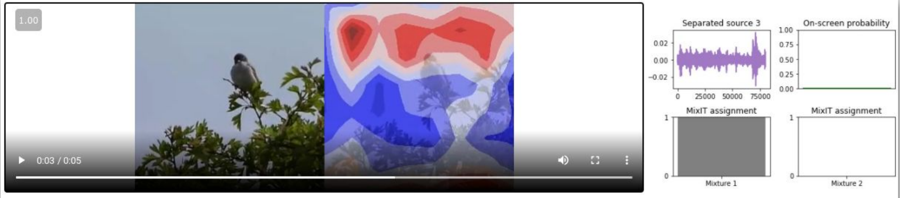
# Thank you!

## Poster Session 2, May 3rd 2021 9am - 11am (PDT)

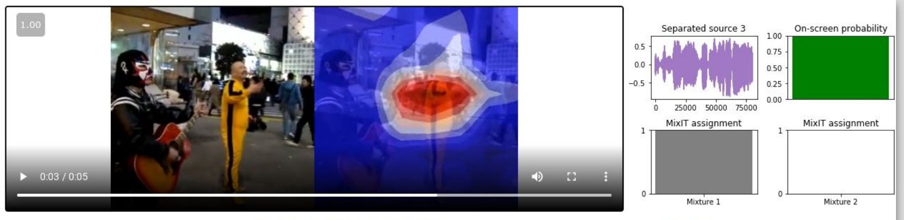
## More examples and dataset available online: audioscope.github.io



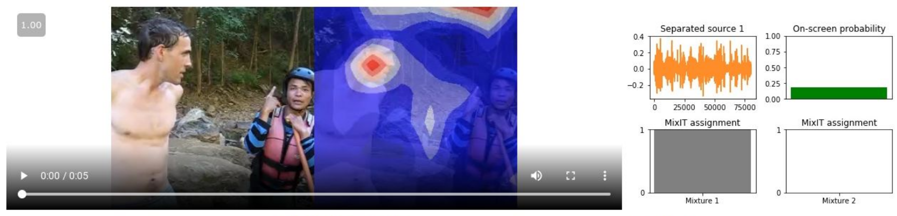
Video clip from video by deeje with modified audio and overlaid attention map, license: CC-BY-SA 2.0.



Video clip from "Whitethroat" by S. Rae with modified audio and overlaid attention map, license: CC-BY 2.0.



Video clip from "Luchador and Yellow Jumpsuit" by tenaciousme with modified audio and overlaid attention map, license: CC-BY 2.0.



Video clip from "MVI_0917" by jeminichronicles with modified audio and overlaid attention map, license: CC-BY-SA 2.0.