THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

# Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

Andrew Owens     Alexei A. Efros

UC Berkeley

Presented by: **Justin Chen, Yulu Pan, Liujie Zheng, Soumitri Chattopadhyay**

Dept. of Computer Science, UNC Chapel Hill

10/11/2023

# Outline

- Motivation
- Methodology
- Applications
    - (a) sound source localization
    - (b) audio-visual action recognition
    - (c) on/off-screen audio-visual source separation
- Qualitative results and discussion

# Motivation

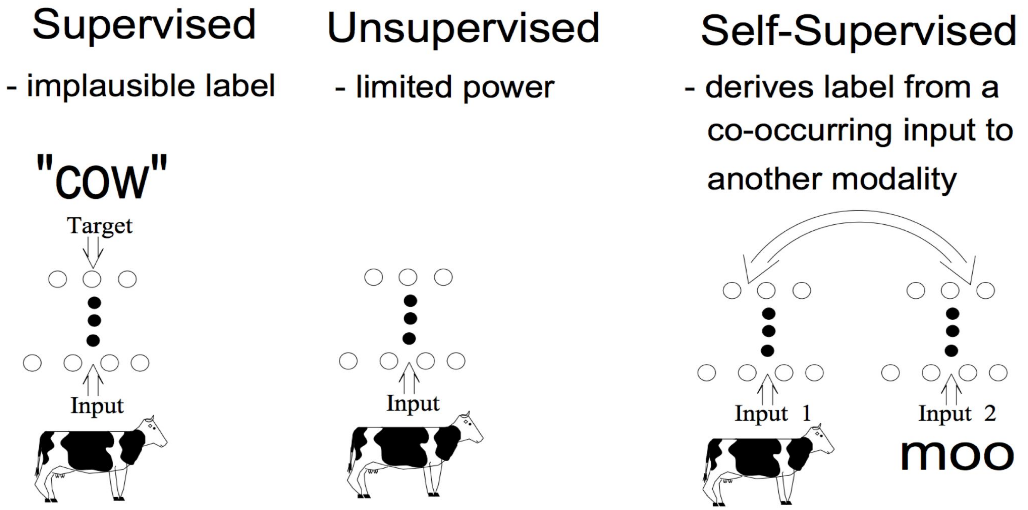
- Why learn audio and visual representations together at all?
  - Well, auditory and visual senses are closely related for perception, and muting any modality can degrade performance, even for humans!

# Motivation

McGurk effect: Humans fuse audio and visual signals at a fairly early stage of processing, the two modalities are used jointly in perceptual grouping

**Idea:** train a model to find audio-visual correspondences in video

# Multisensory self-supervision




Why self-supervised?

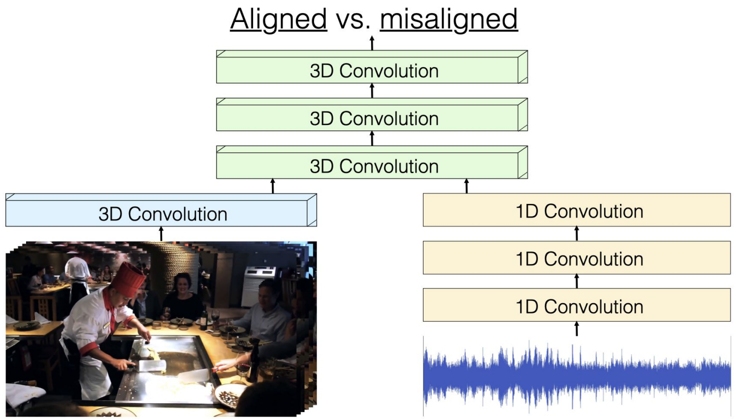
- Manually annotating audio-visual correspondences would be very expensive and difficult to scale

Image credits: Virginia de Sa. Learning Classification with Unlabeled Data. NIPS 1994.

# Self-supervised Multisensory Representation

- Align video with sound
    - Train a network to distinguish aligned and misaligned clips
        - In half of the training data, the vision and sound streams are synchronized; the other half audio is shifted by a few seconds

Fused audio-visual representation

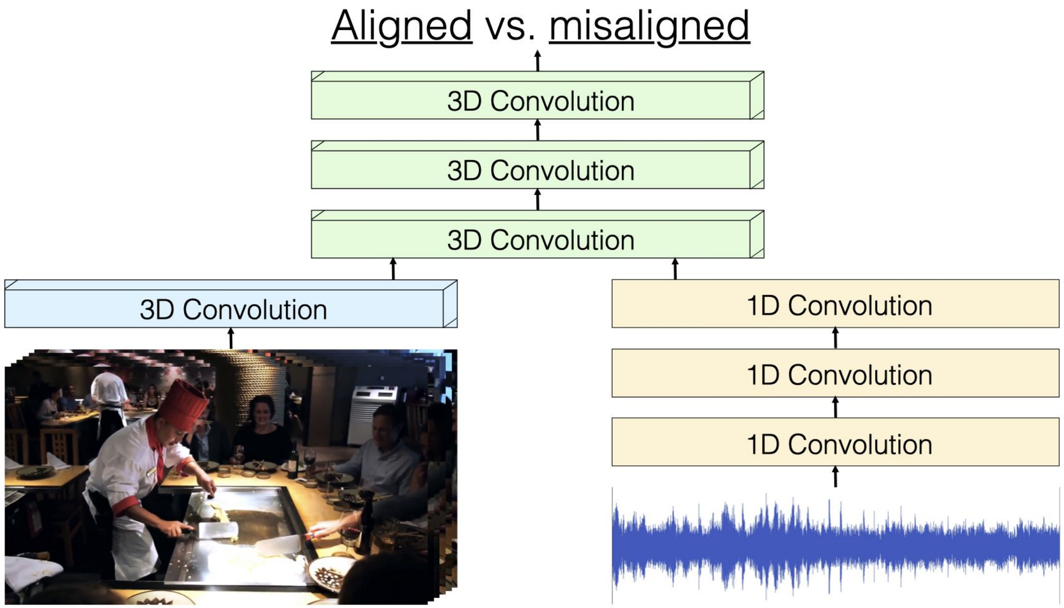



Model:

- 3D ResNet-18
- Early fusion
- 30Hz video + raw waveform

# Self-supervised Multisensory Representation

## Fused audio-visual representation




Training:

- 750K AudioSet videos
- 4.2 sec. clips
- Random 2-5.8 sec. shifts
- 125 frames per example
- 60% accuracy on alignment task

# Self-supervised Multisensory Representation

The task is challenging!

- Audio is shifted by a few seconds vs random pairs of video + audio?


# Self-supervised Multisensory Representation

Evaluated on Kinetics

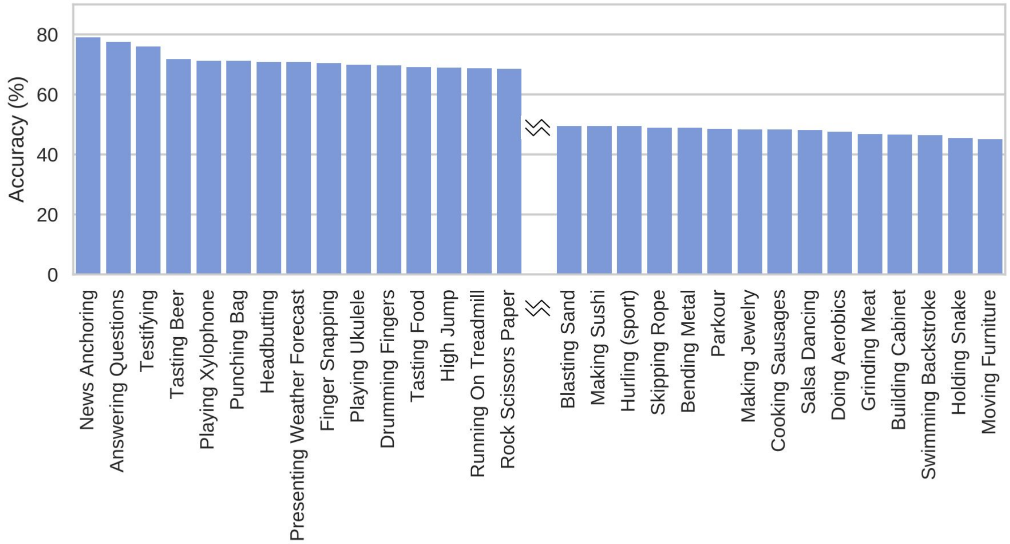
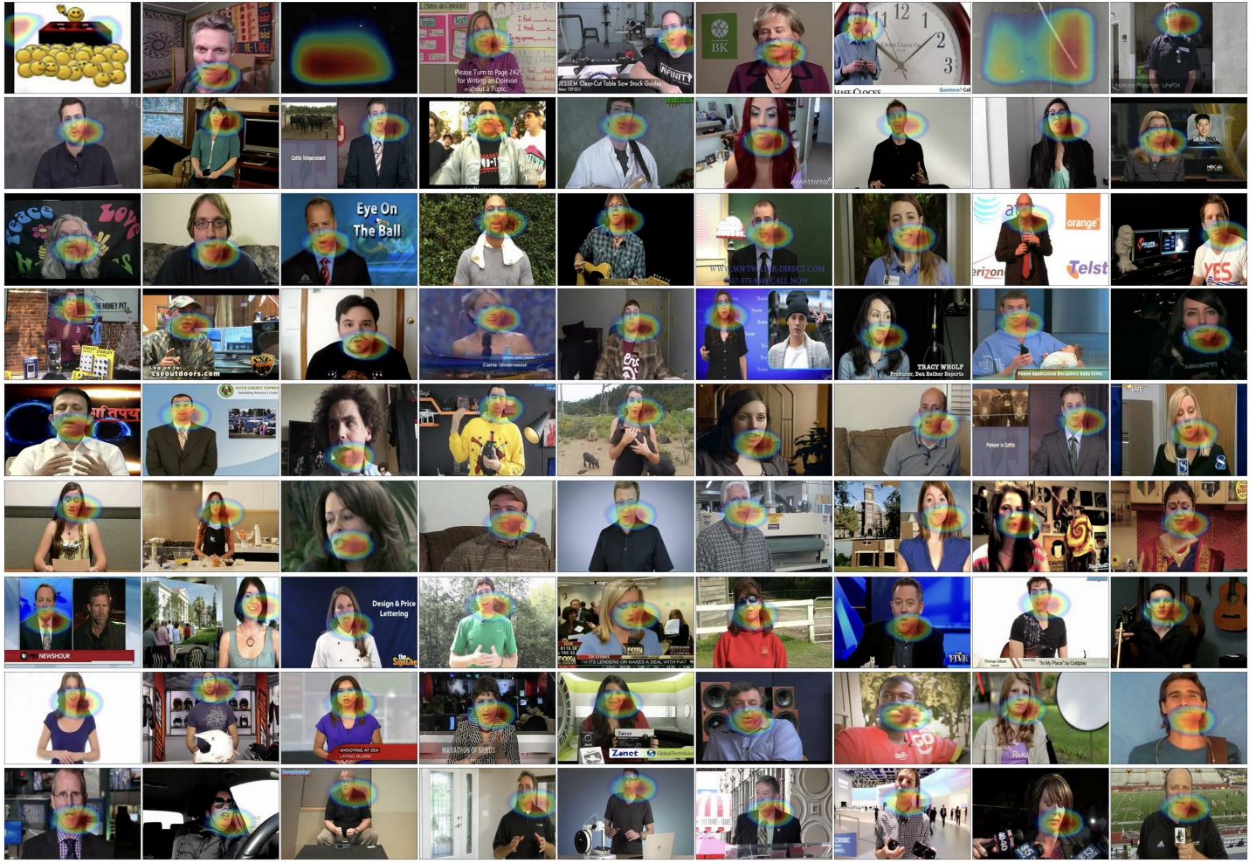



Fig. A1: Accuracy of our model in predicting audio-visual synchronization for the classes in the Kinetics dataset. Chance is 50%.

# Application: Sound source localization

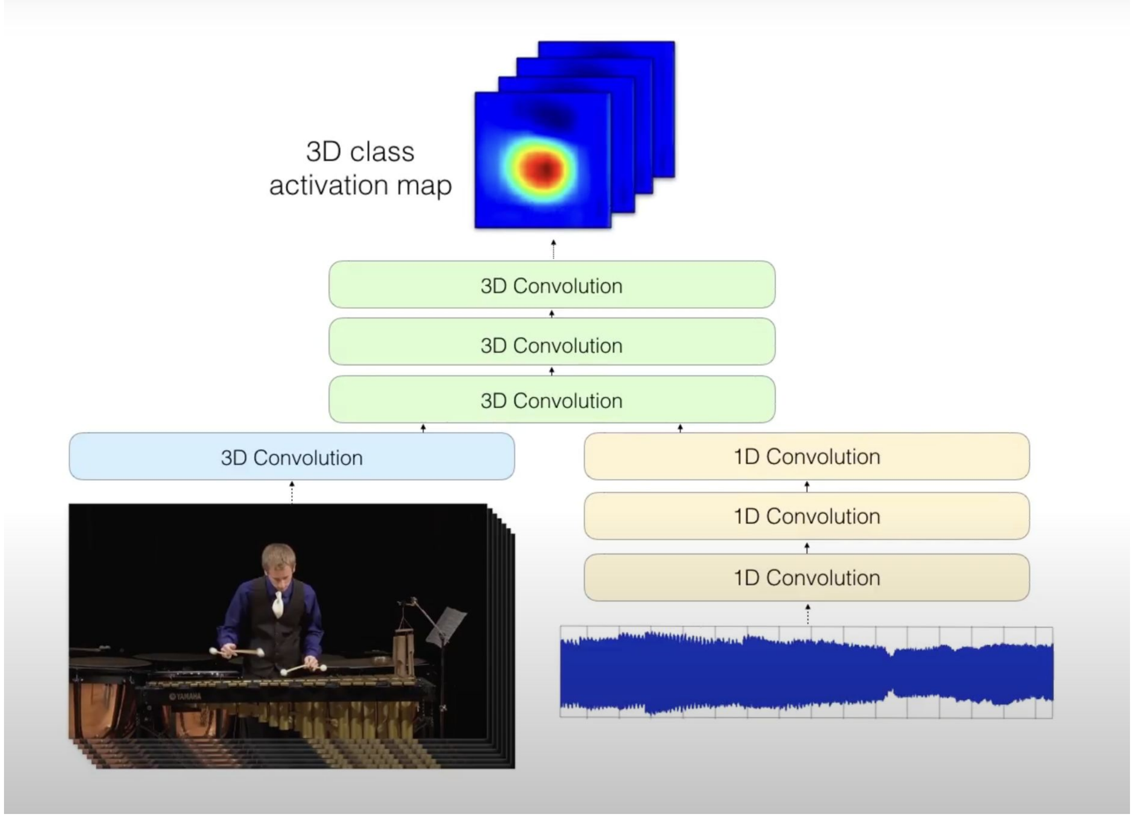


# Application: Sound source localization

# Application: Sound source localization

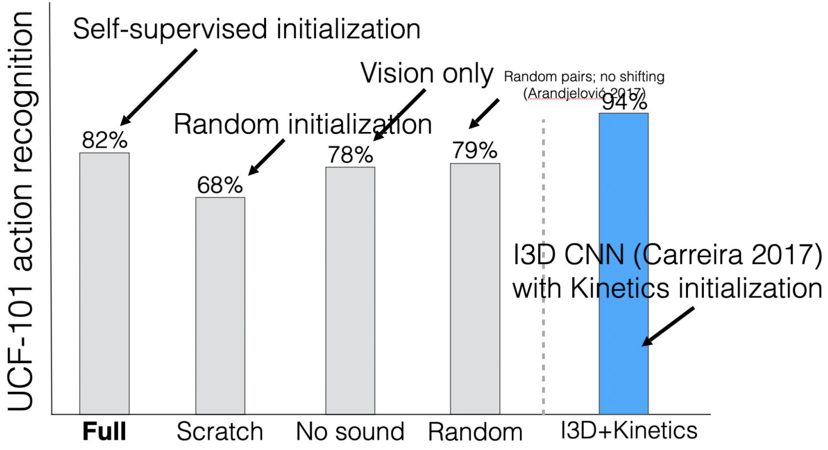
Change the goal from "aligned or not" to predicting a 3D class activation map

Potentially change the goal to other tasks for wider application



# Application: Audio-visual Action Recognition

- Action recognition on UCF-101
- Initialized the weights with those learned from our alignment task, fine-tuned on UCF-101 dataset

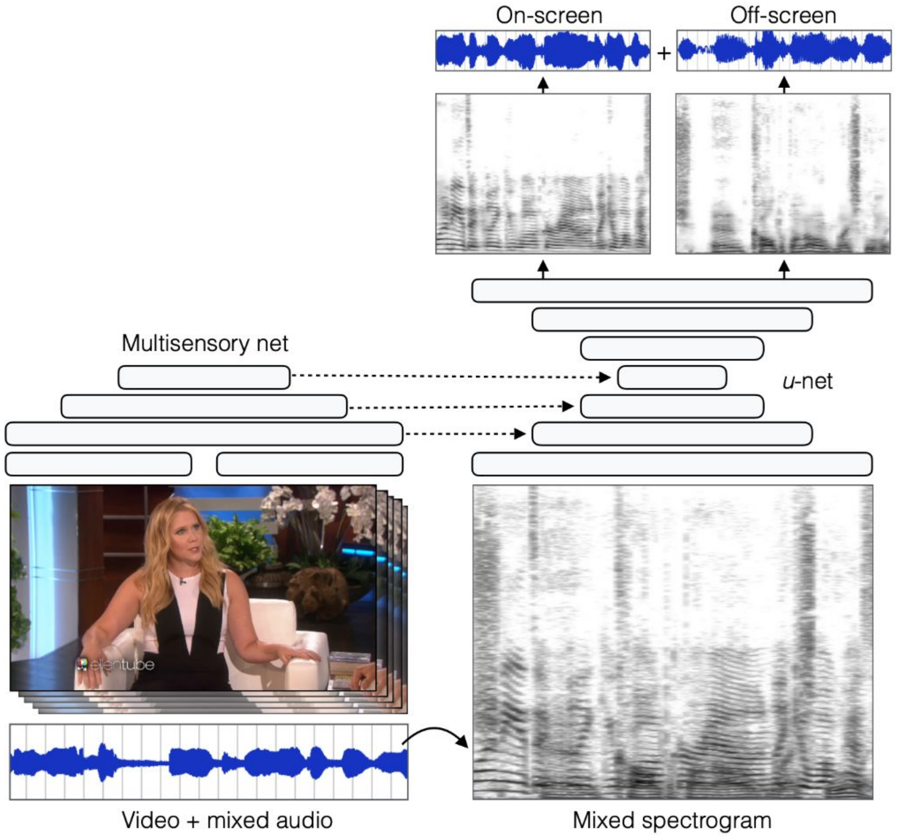



| Model | Acc. |
|---|---|
| Multisensory (full) | 82.1% |
| Multisensory (spectrogram) | 81.1% |
| Multisensory (random pairing [16]) | 78.7% |
| Multisensory (vision only) | 77.6% |
| Multisensory (scratch) | 68.1% |
| I3D-RGB (scratch) [56] | 68.1% |
| O3N [19]* | 60.3% |
| Purushwalkam et al. [61]* | 55.4% |
| C3D [62,56]* | 51.6% |
| Shuffle [17]* | 50.9% |
| Wang et al. [63,61]* | 41.5% |
| I3D-RGB + ImageNet [56] | 84.2% |
| I3D-RGB + ImageNet + Kinetics [56] | 94.5% |

Table 1: Action recognition on UCF-101 (split 1). We compared methods pretrained without labels (top), and with semantic labels (bottom). Our model, trained both with and without sound, significantly outperforms other self-supervised methods. Numbers annotated with "*" were obtained from their corresponding publications; we retrained/evaluated the other models.

# Application: on/off-screen source separation

- Create synthetic sound mixtures by summing an input video's audio track with a randomly chosen track from a random video.
- Train a U-Net that takes in mixed audio spectrogram and input and seperates on-screen and off-screen audios.
- Features from the multisensory encoder are fused at hierarchical levels, ensuring video features match audio sampling rate in concatenation

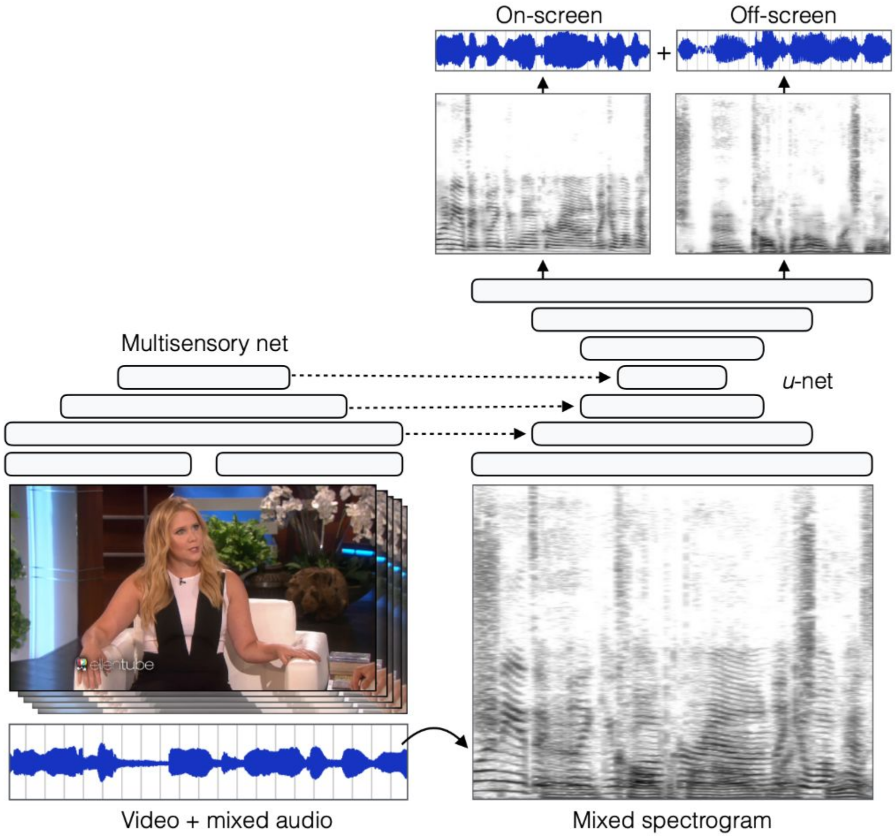


# Application: on/off-screen source separation

Loss function used to train U-Net:

- Simple L1 distance
- Considered two versions –
  - (a) Constraint of on-screen/off-screen identity is enforced (i.e. foreground-background)
  - (b) Treating the sounds as two layers (i.e. permutation invariant)
- Latter version allows on- and off-screen sounds to be swapped in loss term

$$\mathcal{L}_{\mathcal{P}}(x_F, x_B, \hat{x}_1, \hat{x}_2) = \min(L(\hat{x}_1, \hat{x}_2), L(\hat{x}_2, \hat{x}_1)),$$



# Evaluation for Audio-only Separation

| Method | All | | | | Mixed sex | | Same sex | | GRID transfer | |
|---|---|---|---|---|---|---|---|---|---|---|
| | On/off | SDR | SIR | SAR | On/off | SDR | On/off | SDR | On/off | SDR |
| On/off + PIT | 11.2 | 7.6 | 12.1 | 10.2 | 10.6 | 8.8 | 11.8 | 6.5 | 13.0 | 7.8 |
| Full on/off | 11.4 | 7.0 | 11.5 | 9.8 | 10.7 | 8.4 | 11.9 | 5.7 | 13.1 | 7.3 |
| Mono | 11.4 | 6.9 | 11.4 | 9.8 | 10.8 | 8.4 | 11.9 | 5.7 | 13.1 | 7.3 |
| Single frame | 14.8 | 5.0 | 7.8 | 10.3 | 13.2 | 7.2 | 16.2 | 3.1 | 17.8 | 5.7 |
| No early fusion | 11.6 | 7.0 | 11.0 | 10.1 | 11.0 | 8.4 | 12.1 | 5.7 | 13.5 | 6.9 |
| Scratch | 12.9 | 5.8 | 9.7 | 9.4 | 11.8 | 7.6 | 13.9 | 4.2 | 15.2 | 6.3 |
| I3D + Kinetics | 12.3 | 6.6 | 10.7 | 9.7 | 11.6 | 8.2 | 12.9 | 5.1 | 14.4 | 6.6 |
| $u$-net PIT [36] | – | 7.3 | 11.4 | 10.3 | – | 8.8 | – | 5.9 | – | 8.1 |
| Deep Sep. [67] | – | 1.3 | 3.0 | 8.7 | – | 1.9 | – | 0.8 | – | 2.2 |

Table 2: Source separation results on speech mixtures from the VoxCeleb (broken down by gender of speakers in mixture) and transfer to the simple GRID dataset. We evaluate the on/off-screen sound prediction error (On/off) using $\ell_1$ distance to the true log-spectrograms (lower is better). We also use blind source separation metrics (higher is better) [68].

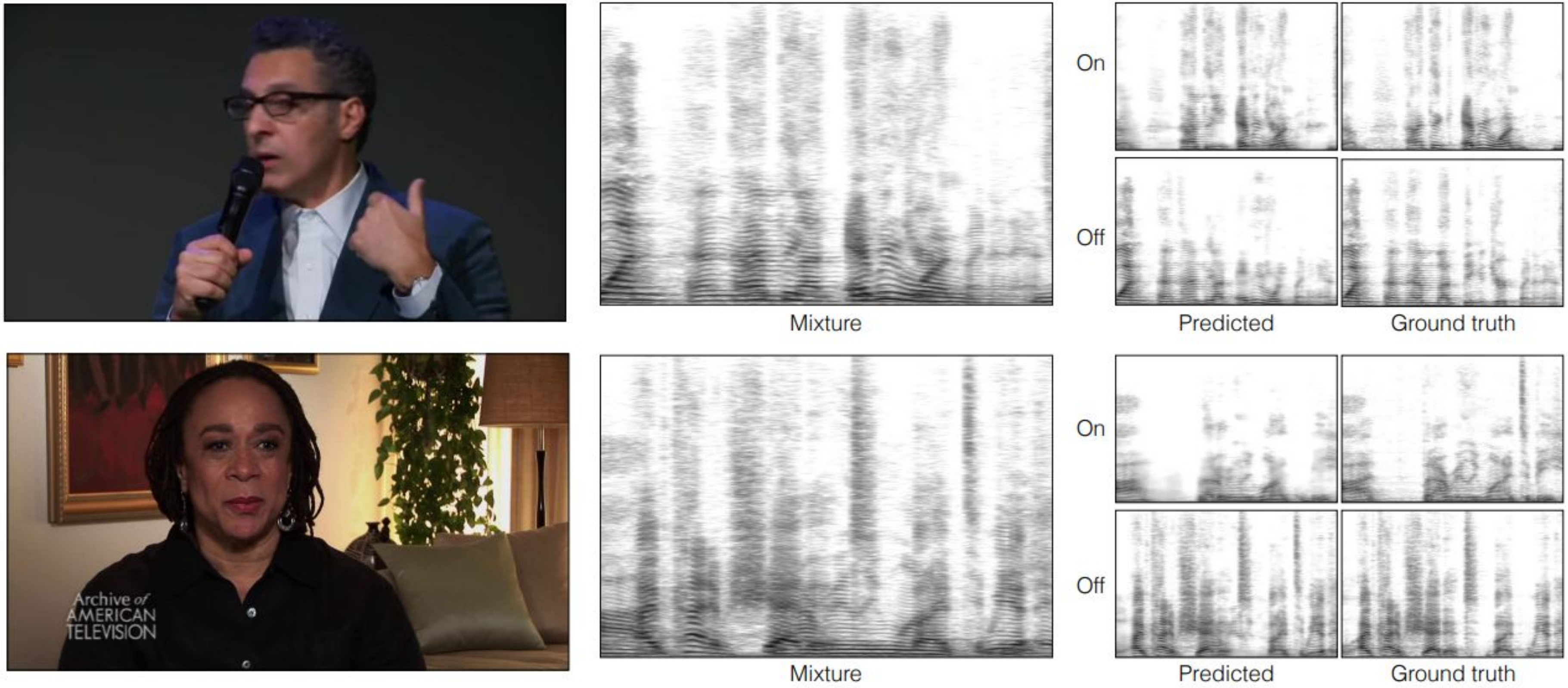
# Evaluation for Audio-Visual Separation

| VoxCeleb short videos (200ms) | On-SDR | SDR | SIR | SAR |
|---|---|---|---|---|
| Ours (on/off) | 7.6 | 5.3 | 7.8 | 10.8 |
| Hou et al. [42] | 4.5 | – | – | – |
| Gabbay et al. [44] | 3.5 | – | – | – |
| PIT-CNN [36] | – | 7.0 | 10.1 | 11.2 |
| $u$-net PIT [36] | – | 7.0 | 10.3 | 11.0 |
| Deep Sep. [67] | – | 2.7 | 4.2 | 10.3 |

Table 3: Comparison of audio-visual and audio-only separation methods on short (200ms) videos. We compare SDR of the on-screen audio prediction (On-SDR) with audio resampled to 2 kHz.

- Adopted our training protocol on the concurrent/closely related prior models
- For the baselines, Viola-Jones face detector was used to crop the mouth region of speakers
- Downsampling to 2 kHz was done to maintain consistency with baselines having small number of frequency bands in their spectrogram

# Qualitative Results for on/off-screen Separation



# Qualitative Results for on/off-screen Separation

# Qualitative Results for on/off-screen Separation

# Thank you!

Questions?

# Arguments

- Our pipeline is simple, intuitive and effective. PixelPlayer's pipeline is way more complicated than ours.
- Their new MUSIC dataset only contains 685 videos
    - Unpopular dataset (101 stars on Github)
    - Only YouTube video IDs, what if the video gets deleted/corrupted?
- Their application is limited (only sound source localization and seperation) while ours has a wide range of applications in the audio-visual community
- They only test on the small MUSIC dataset, while ours test on more popular and large scale dataset. Ours has more quantitative results and more baselines.