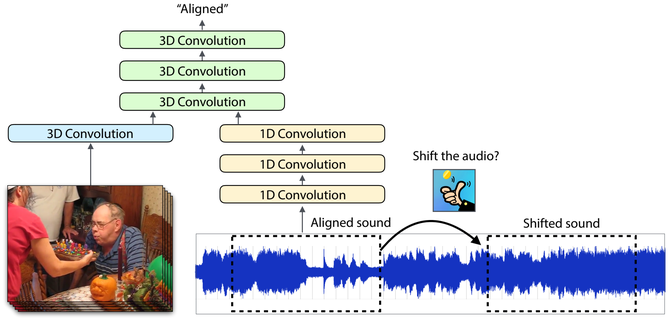
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

Andrew Owens Alexei A. Efros

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Motivation

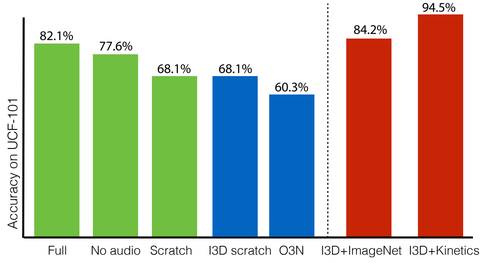
Lots of videos have audio tracks, but today's recognition models mostly ignore them. We learn a rich audio-visual representation via self-supervision, by predicting whether images and sound are temporally aligned.



We train with 750,000 unlabeled videos and apply the features to downstream tasks.

Application: action recognition

- Pretraining & sound are helpful
- Outperforms other self-supervised representations
- Still trails best supervised methods



Related work

[1] A. Owens, P. Isola, J. McDermott, A. Torralba, E.H. Adelson, W.T. Freeman. Visually Indicated Sounds. CVPR 2016

[2] R. Arandjelovic, A. Zisserman. Look, listen and learn. ICCV 2017.

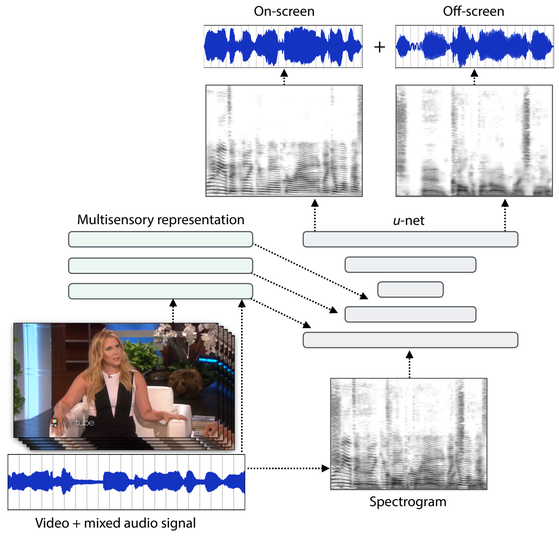
[3] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. Learning deep features for discriminative localization. CVPR 2016.

[4] J. Carreira, A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. CVPR 2016.

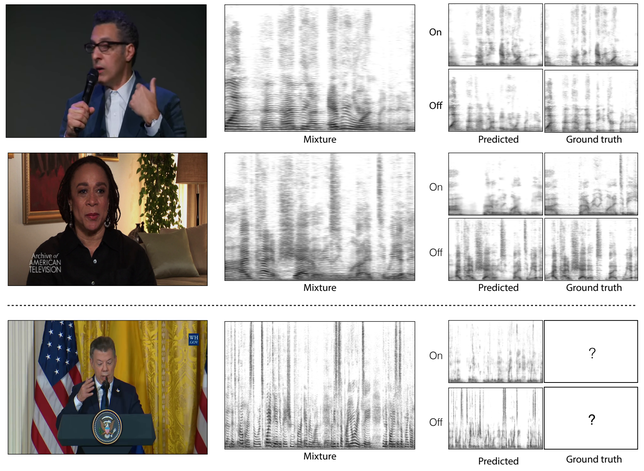
[5] D. Yu, M. Kolbaek, Z. Tan, J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. ICASSP 2017.

Application: source separation

We use our representation to separate on- and off-screen sound.

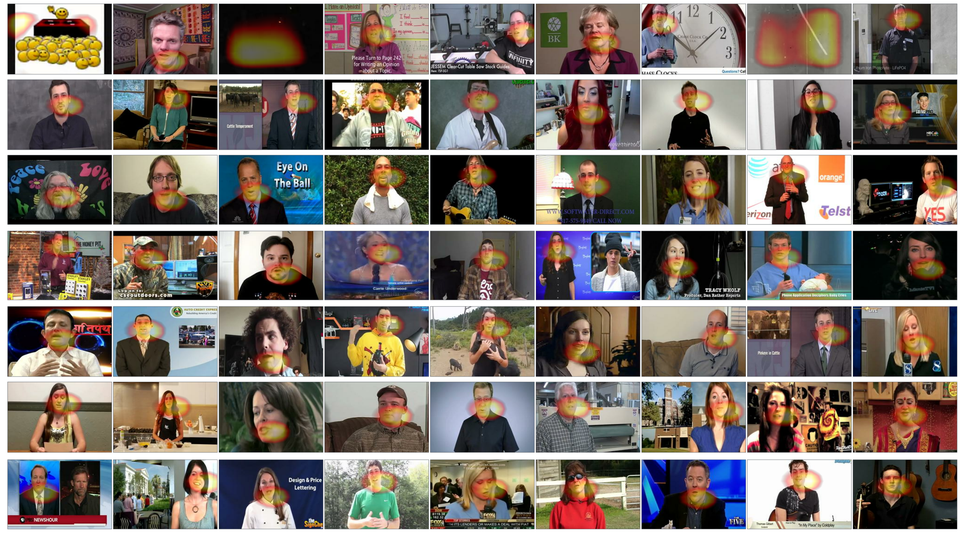


We train the network to recover a video's sound after mixing it with another's. We demonstrate it on real mixtures, too.



What does the network learn?

The network learns to attend to sound sources. We visualize its attention with a class activation map. It responds most strongly to speech:



It attends to other motions, too (top responses from several action categories):