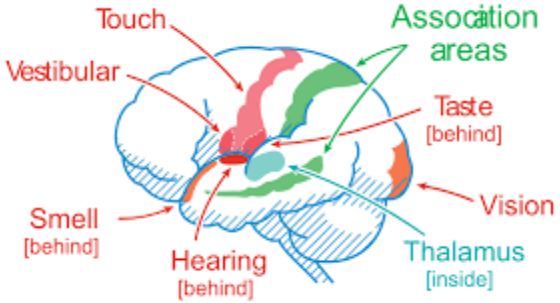
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

Eric Semeniuc, Jan Hansen, Yuan Yao, Yuchi Zhang

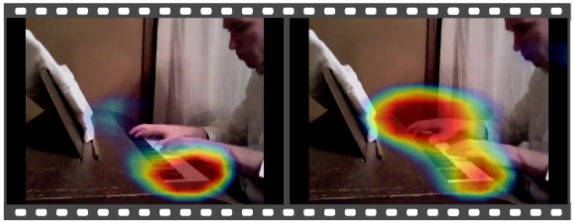
Cross-Modality



Goal: Use the multiple modalities of an event as a learning signal

Image source: CC and University of Utah

Sound Localization

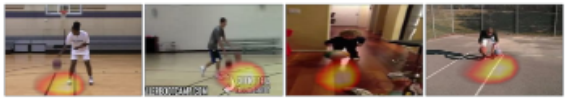


Action Recognition

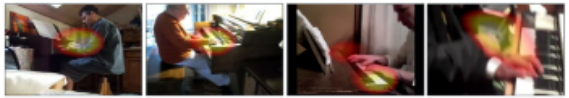

Chopping wood


Shuffling cards


Using keyboard

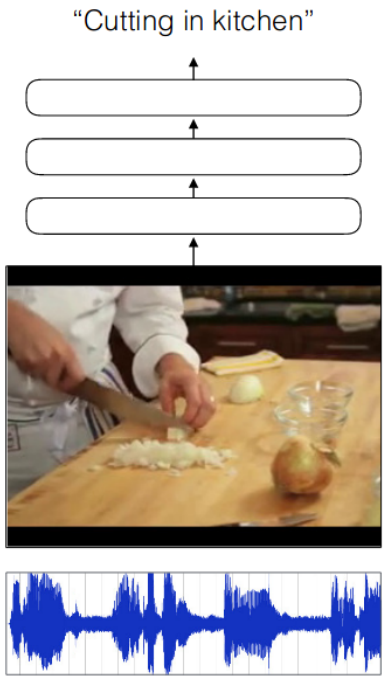

Dribbling basketball


Playing bagpipe

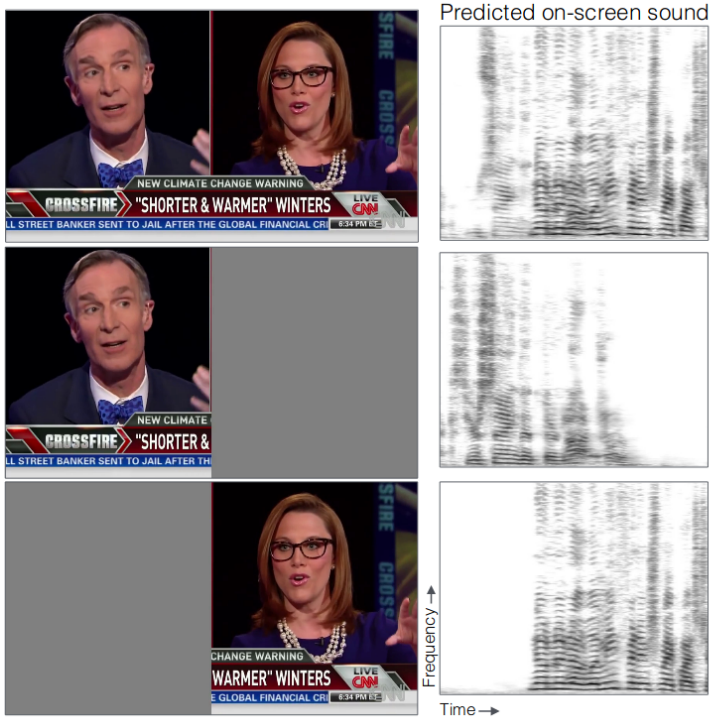

Playing organ

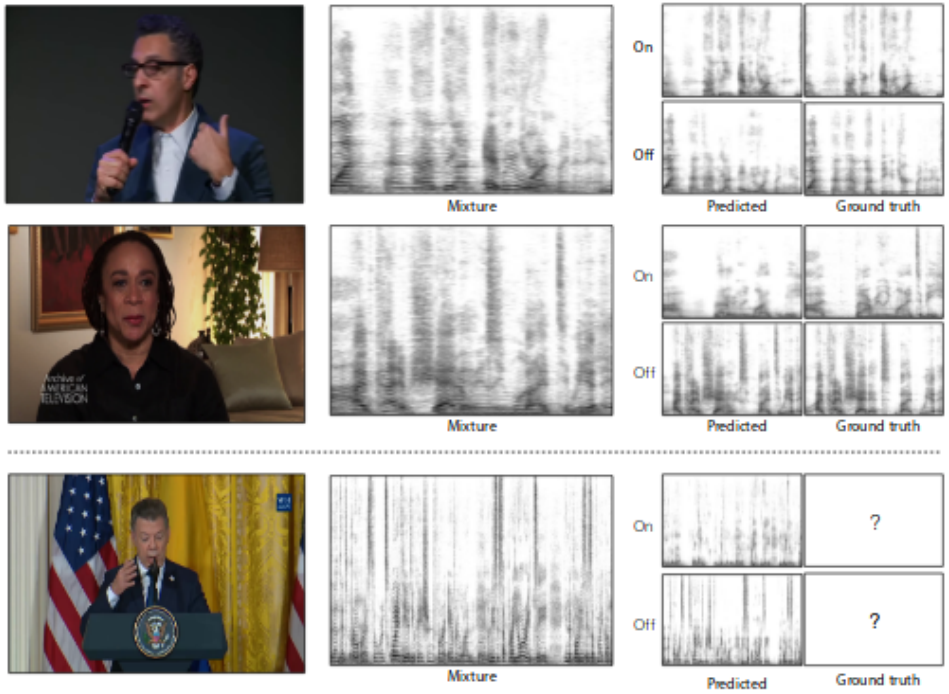

Playing bass guitar


Tap dancing



Stream separation



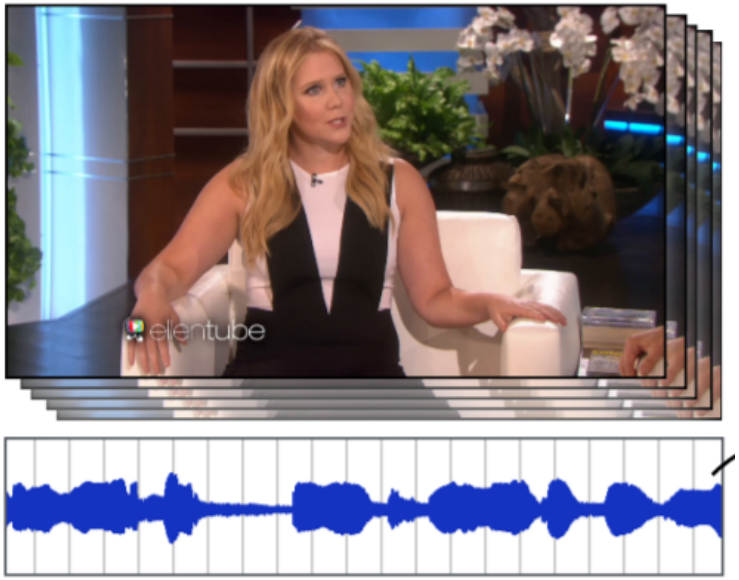


Related work

- Audio-visual scene analysis
  - McGurk effect VIDEO
- Self-supervised learning (no human labeling)
  - Image and audio source coherency (de Sa, Arandjelovic)
- Audio-visual alignment
  - Lip reading (Chung et al)
- Sound localization
  - Associate motion and audio, sound of pixels (...et al.), Zhou
- Blind source/Audio-Visual separation
  - Cocktail party problem
  - Face detection and beam forming

Learn a multisensory representation

Key idea: train a model to predict whether video's audio and visual streams are synchronized.



Align sight with sound

Input: **video clips**, half of the data are synchronized, the others are shifted.

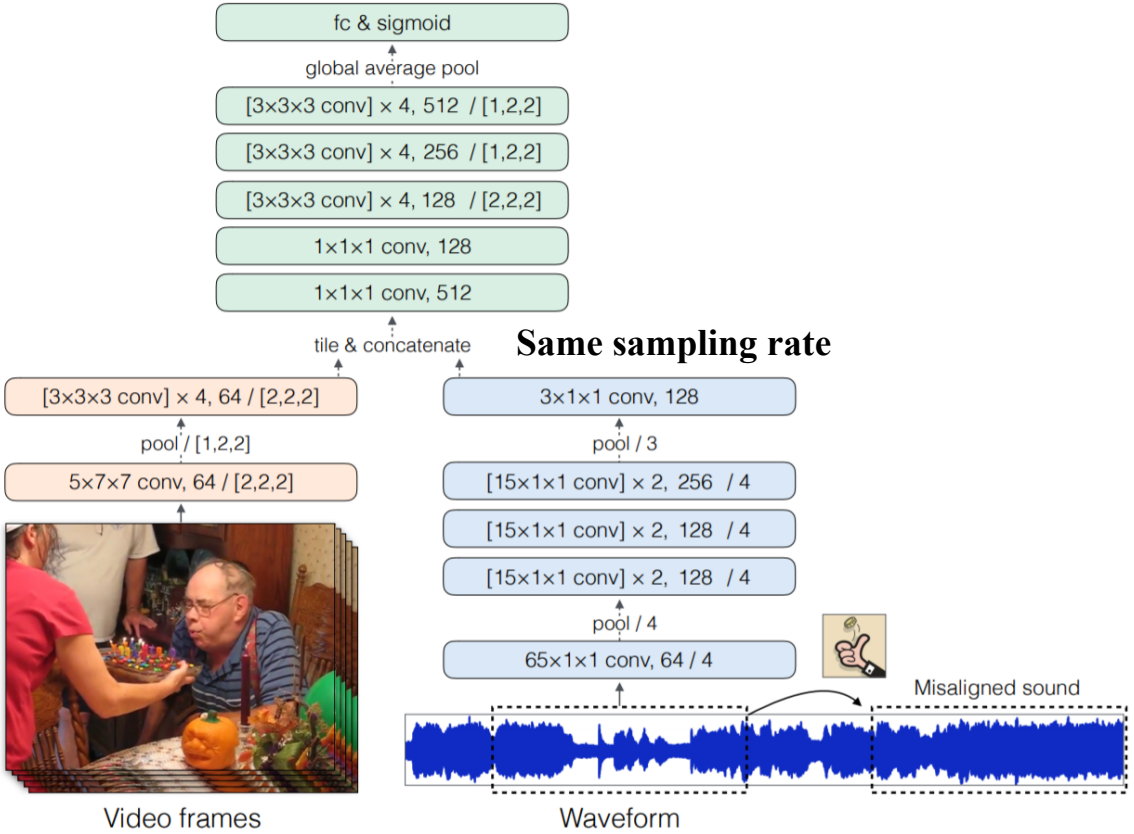
Output: y = {**0,1**} means whether the audio and video are synchronized.

Model: $p_\theta(y \mid I, A)$, I is visual stream, A is audio stream

Objective: (Maximize log-likelihood)

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}_{I,A,t}[\log(p_\theta(y=1 \mid I, A_0)) + \log(p_\theta(y=0 \mid I, A_t))]$$

Fused audio-visual network



Experiment

Data:

- Input 4.2 sec. videos, randomly shifted by 2.0 to 5.8 seconds
- 750,000 videos sampled from AudioSet
- 29.97 Hz
- Random crop + flipping

Performance:

- 59.9% accuracy
- User study 66.6% accuracy

Visualizing sound sources

Zhou et al. 2015:

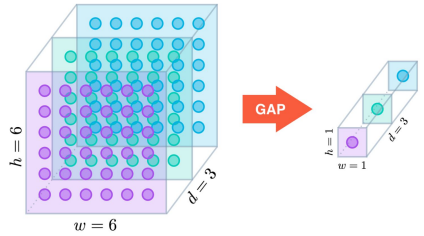
- convolutional units of CNN’s are object detectors in an unsupervised setting
- This is lost when the final fully connected layer is used for classification.

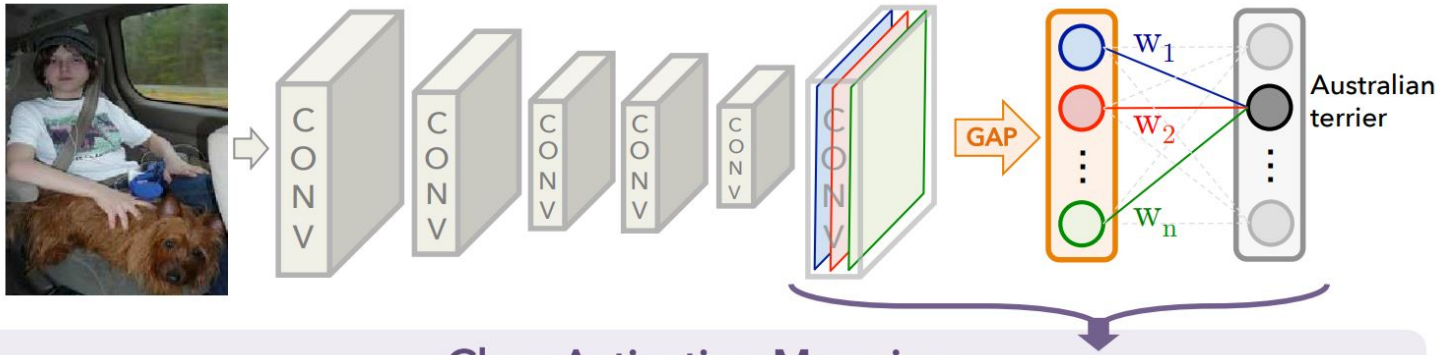
  At the same time GoogLeNet was trying to avoid the final fully connected layer to minimize the number of parameters

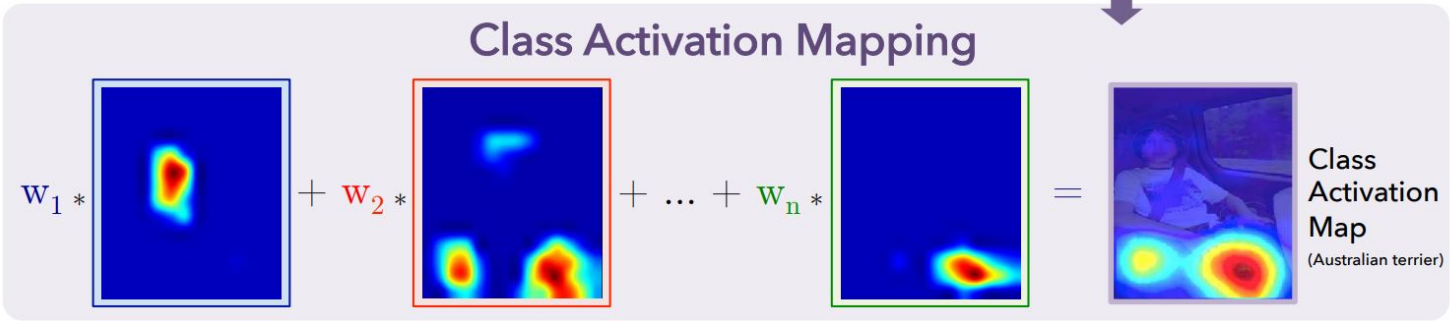
Zhou et al. 2015 - Class Activation Mapping

GAP - Global Average Pooling

GMP - Global Maximum Pooling



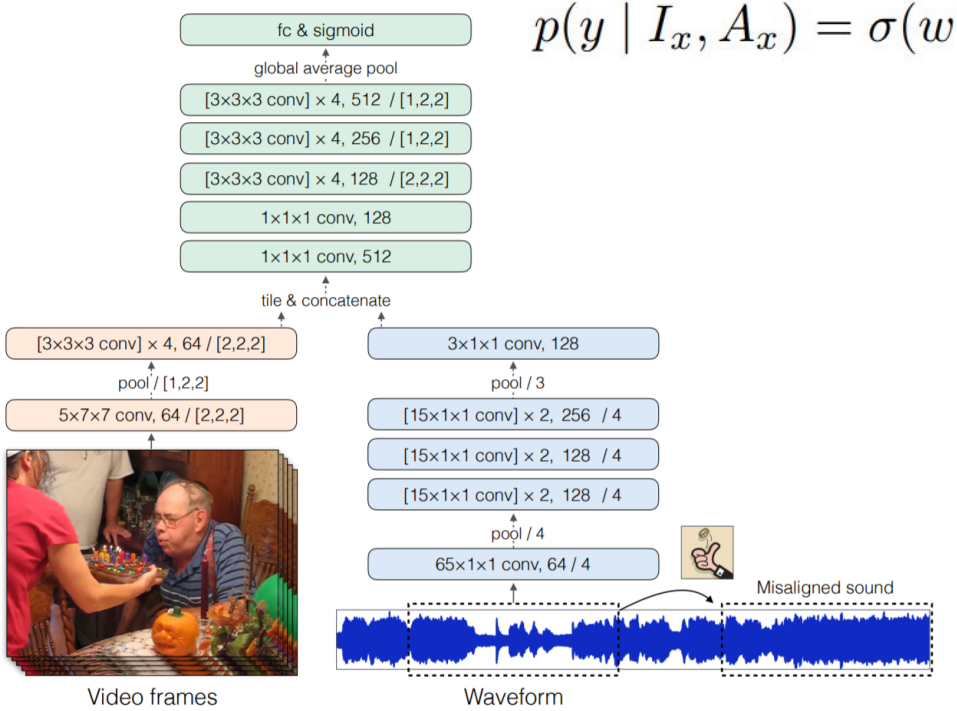




Zhou et al. 2015

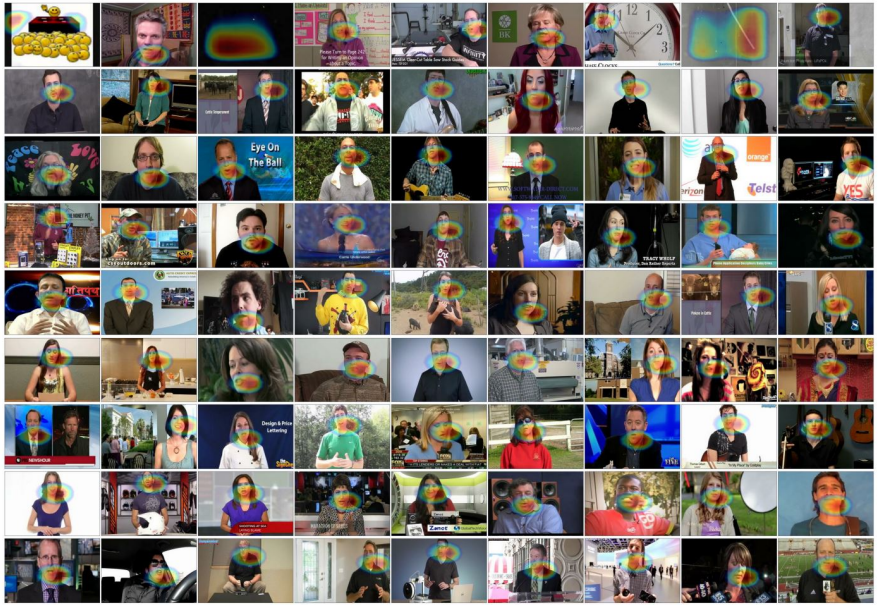
Visualizing sound sources - Class Activation Mapping

$$p(y \mid I_x, A_x) = \sigma(w^\top f(I_x, A_x))$$



Hypothesis: A good audio-visual representation (early fusion) will pay special attention to visual sound sources

Visualizing sound sources

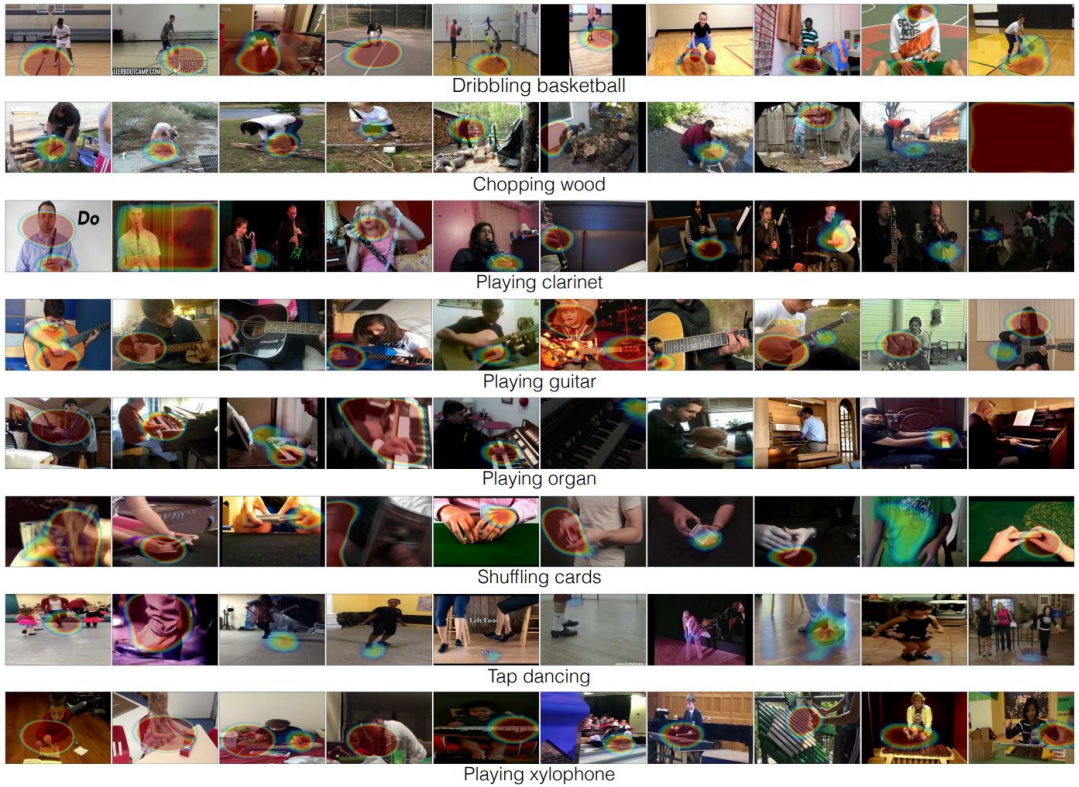


strong response - faces

weak response - no faces

Visualizing sound sources

Kinetics sound dataset: no speech



One sound source: GMP

Multiple sources: GAP

Action recognition

1. We've seen that the representation localizes sound sources
2. Can it also be used in an unsupervised recognition task?

UCF-101 dataset

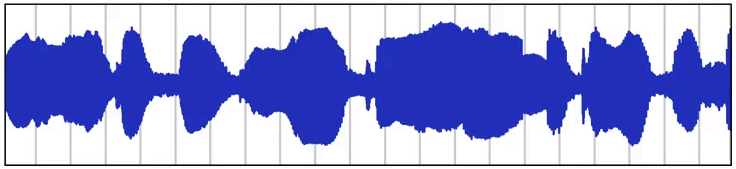
Results

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%

Application: on/off-screen source separation

Adapted from Owens and Efros (2018)

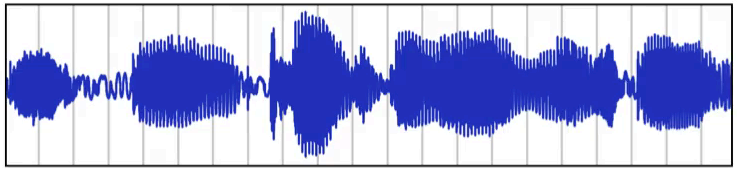
Creating training data



On-screen

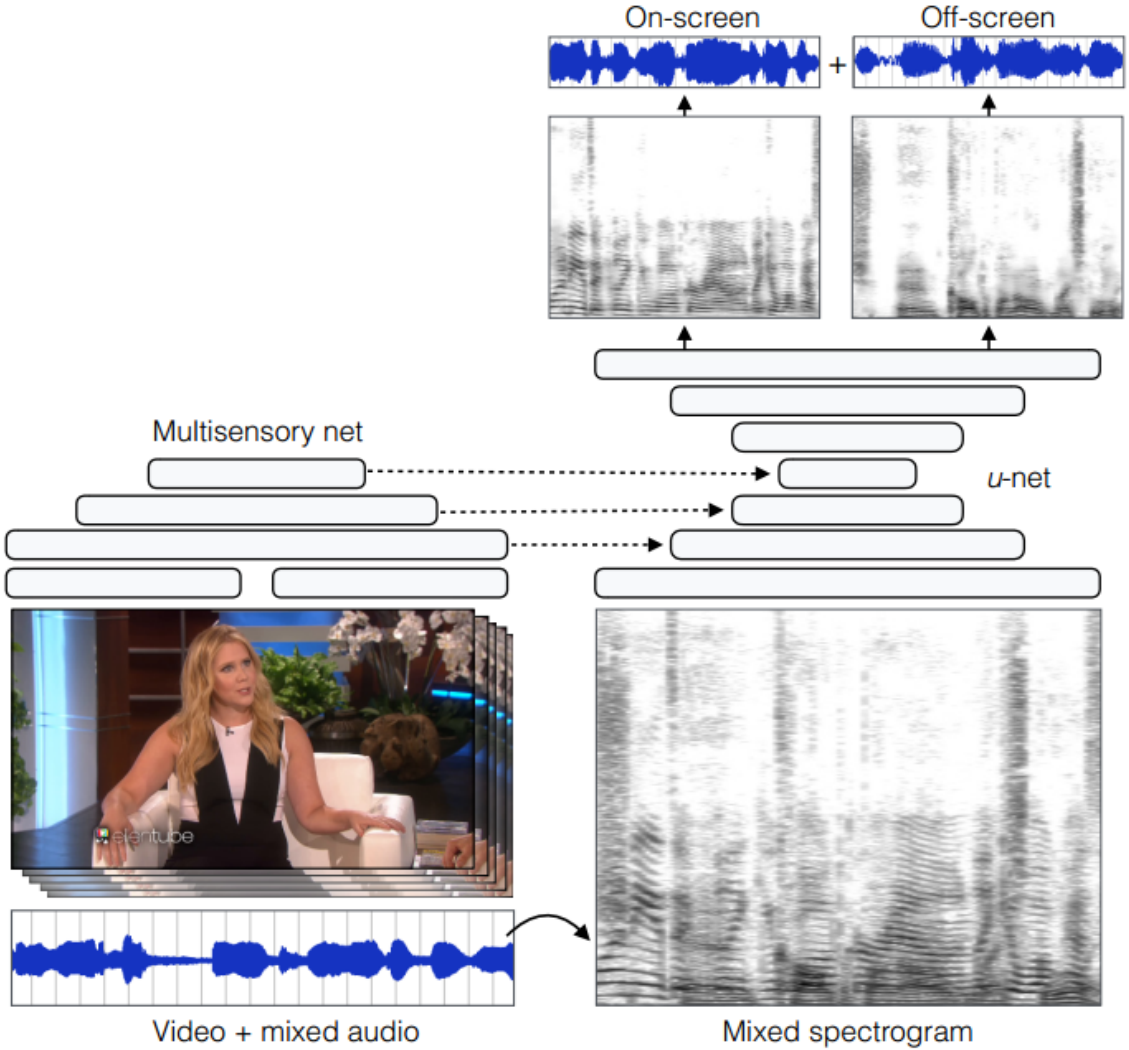
Off-screen

+



Adapted from Owens and Efros (2018)

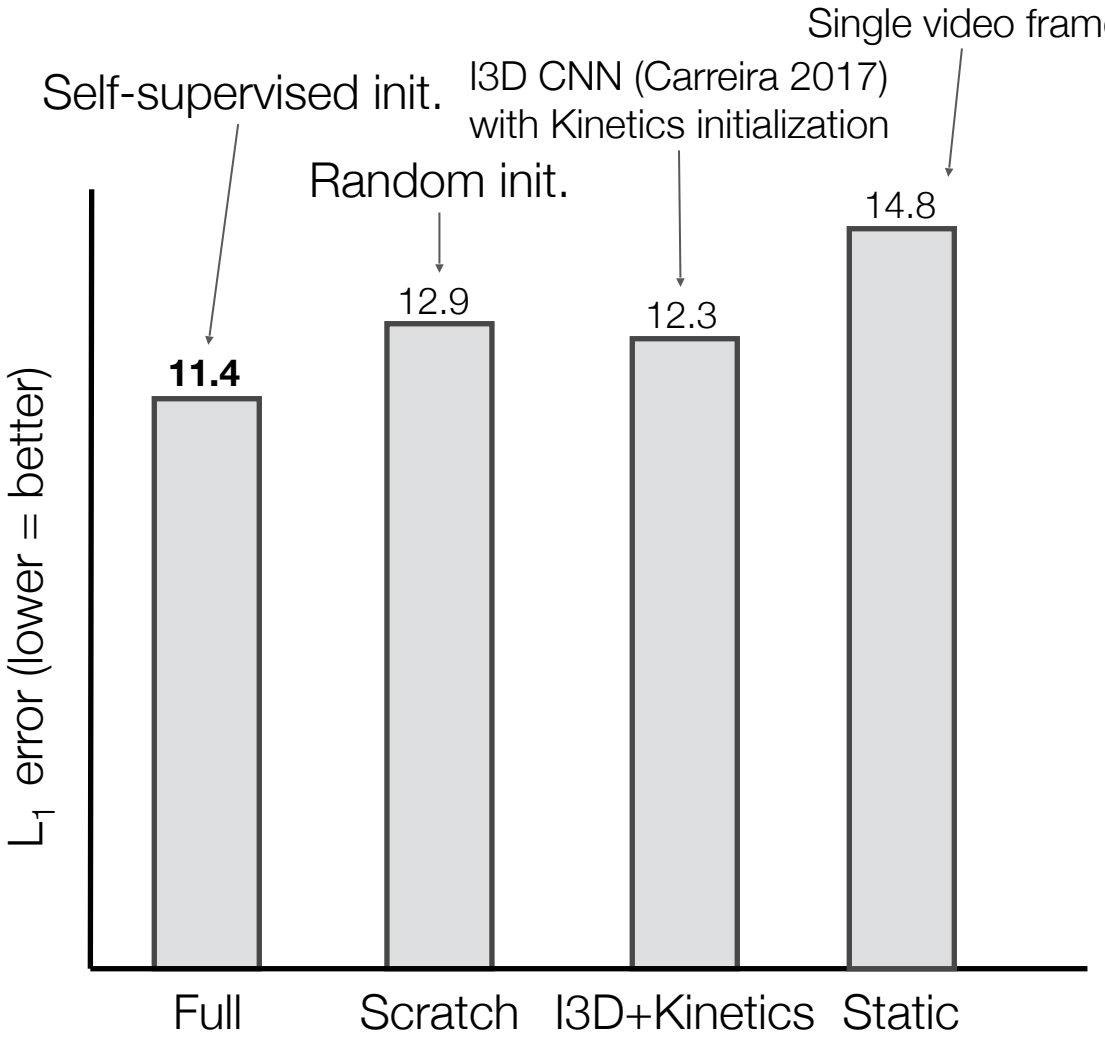
Model

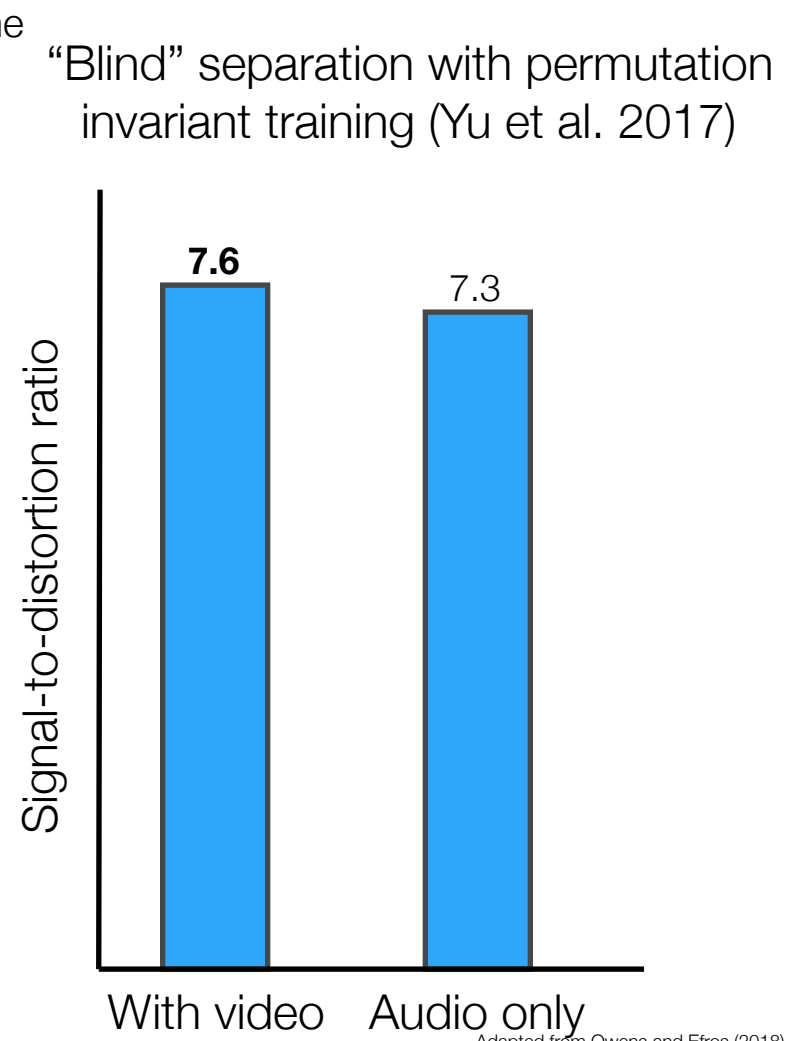


Adapted from Owens and Efros (2018)

Training

- 4 sec. videos from VoxCeleb + AudioSet
- L1-Loss on log spectrograms
- No labels or face detection





Adapted from Owens and Efros (2018)

Input video

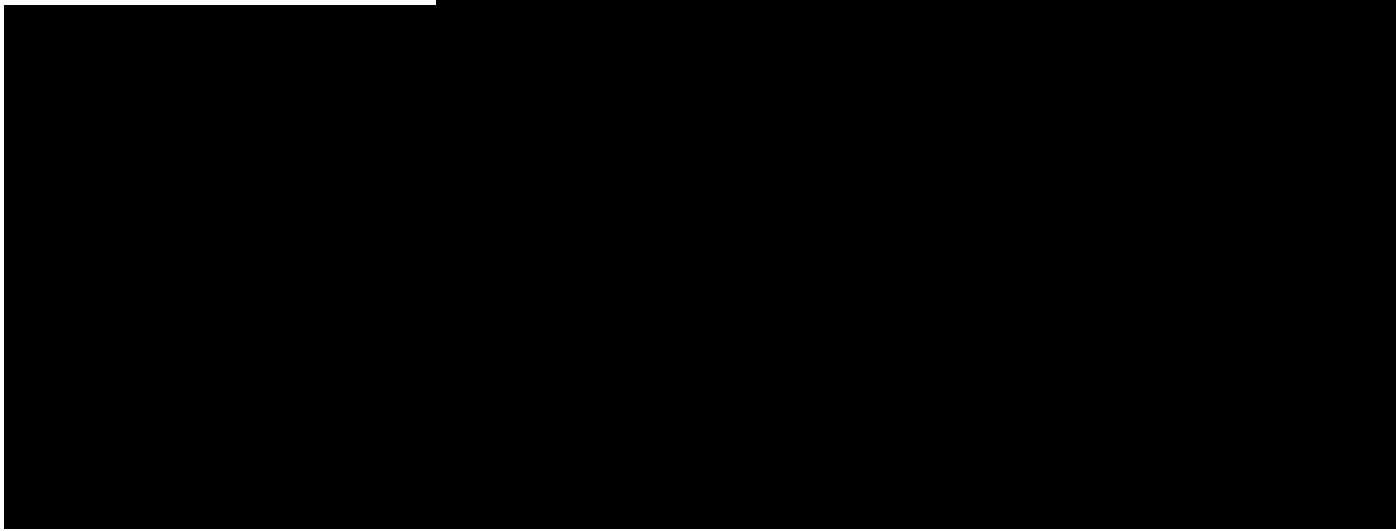
On-screen prediction

Off-screen prediction

Adapted from Owens and Efros (2018)

Multiple on screen sound sources

Mask one side of the screen



Adapted from Owens and Efros (2018)

Discussion

- Pros
  - Multisensory feature learned with self-supervision
  - Three potential applications
    - Sound localization
    - Action recognition
    - Audio-visual source separation
- Cons
  - Action recognition unclear
  - Issue with shot cuts
  - Sound localization, ventriloquist
    - Not for multiple on-screen sound sources
    - Sound localization + source separation? => Sound of Pixels

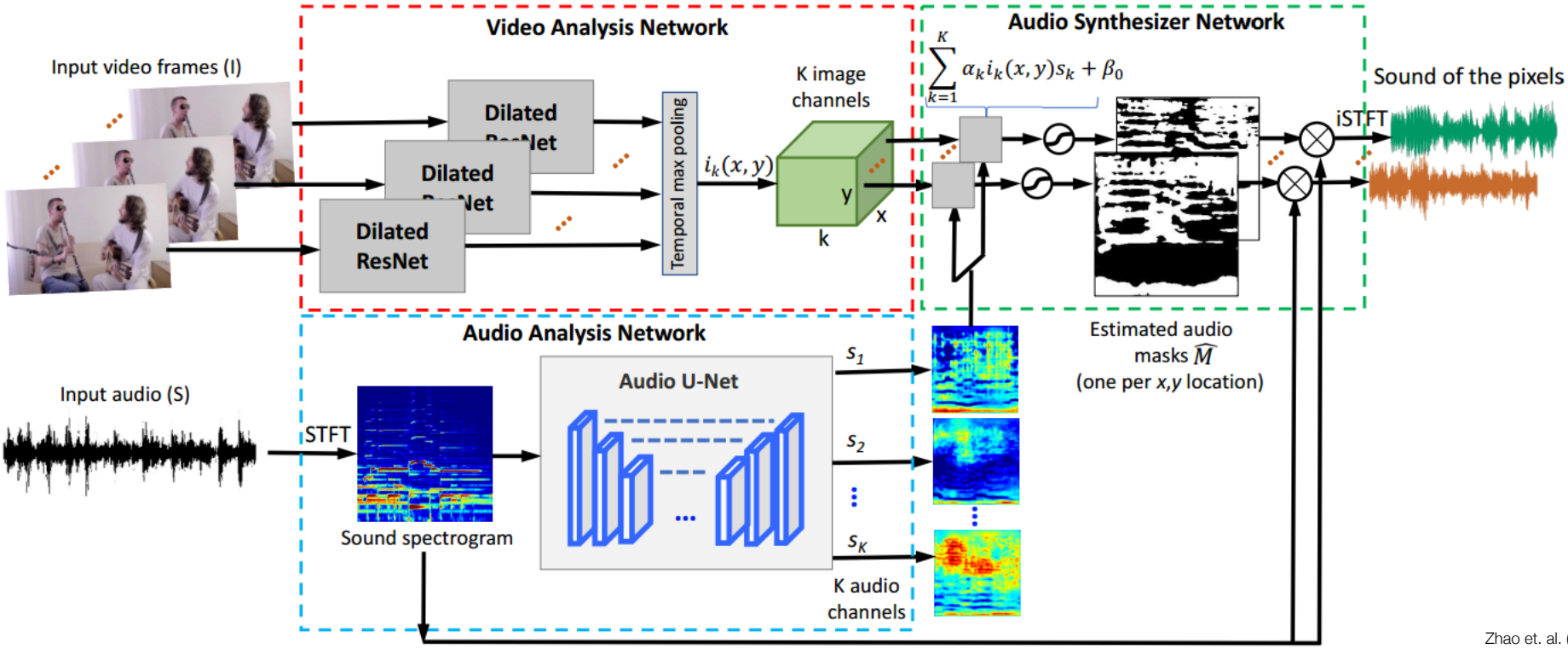
Discussion

- Pros
  - Multisensory feature
  - Three potential applications
    - Sound localization
    - Action recognition
    - Audio-visual source separation
- Cons
  - Action recognition unclear
  - Issue with shot cuts
  - Sound localization, ventriloquist
    - Not for multiple on-screen sound sources
    - Sound localization + source separation? => Sound of Pixels

Discussion

- Pros
  - Multisensory feature
  - Three potential applications
    - Sound localization
    - Action recognition
    - Audio-visual source separation
- Cons
  - Action recognition unclear
  - Issue with shot cuts
  - Sound localization, ventriloquist
    - Not for multiple on-screen sound sources
    - Sound localization + source separation? => Sound of Pixels

Sound of Pixels

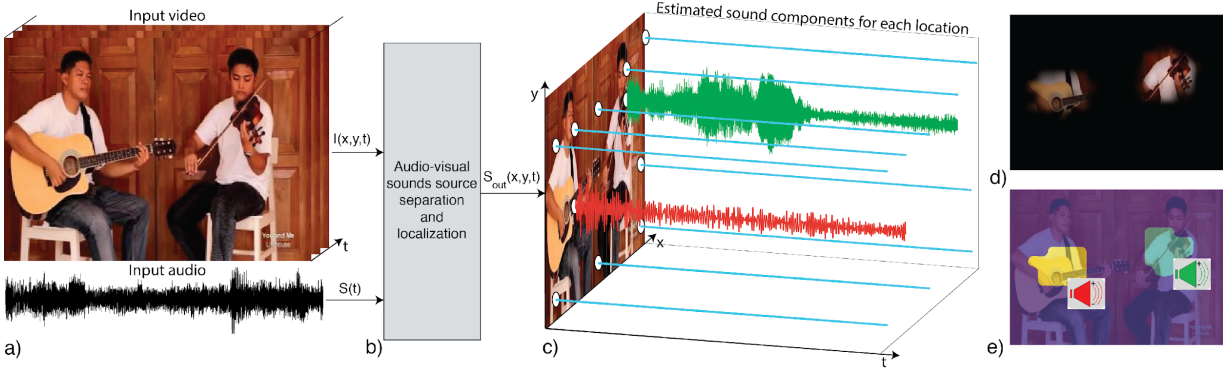


Zhao et. al. (2018)

Concurrent work

The Sound of Pixels

(Zhao, Gan, et al.)



Learning to Separate Object Sounds by Watching Unlabeled Video

(Gao, Feris, Grauman)