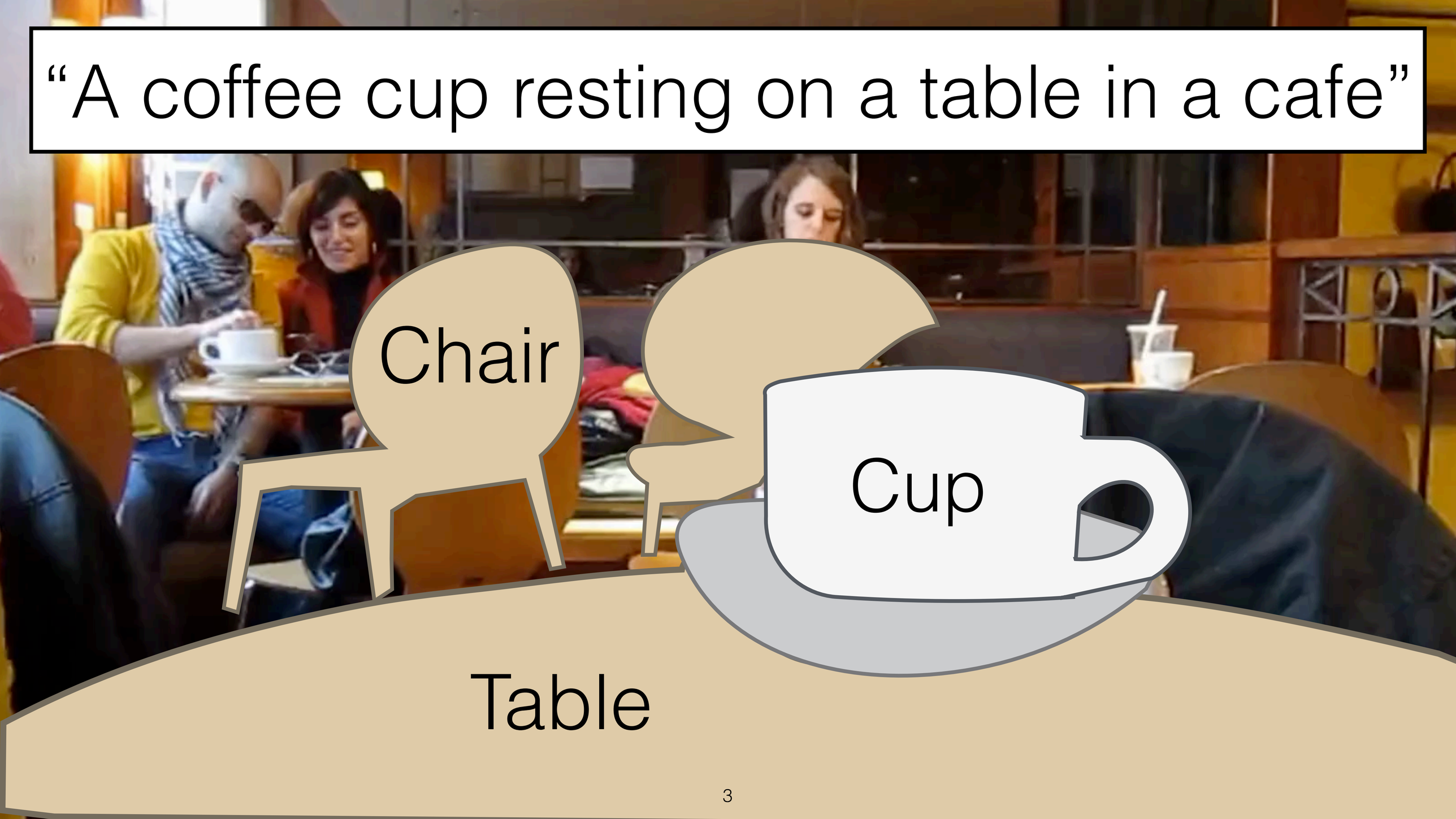
Lecture 15: Sound and touch

Announcements

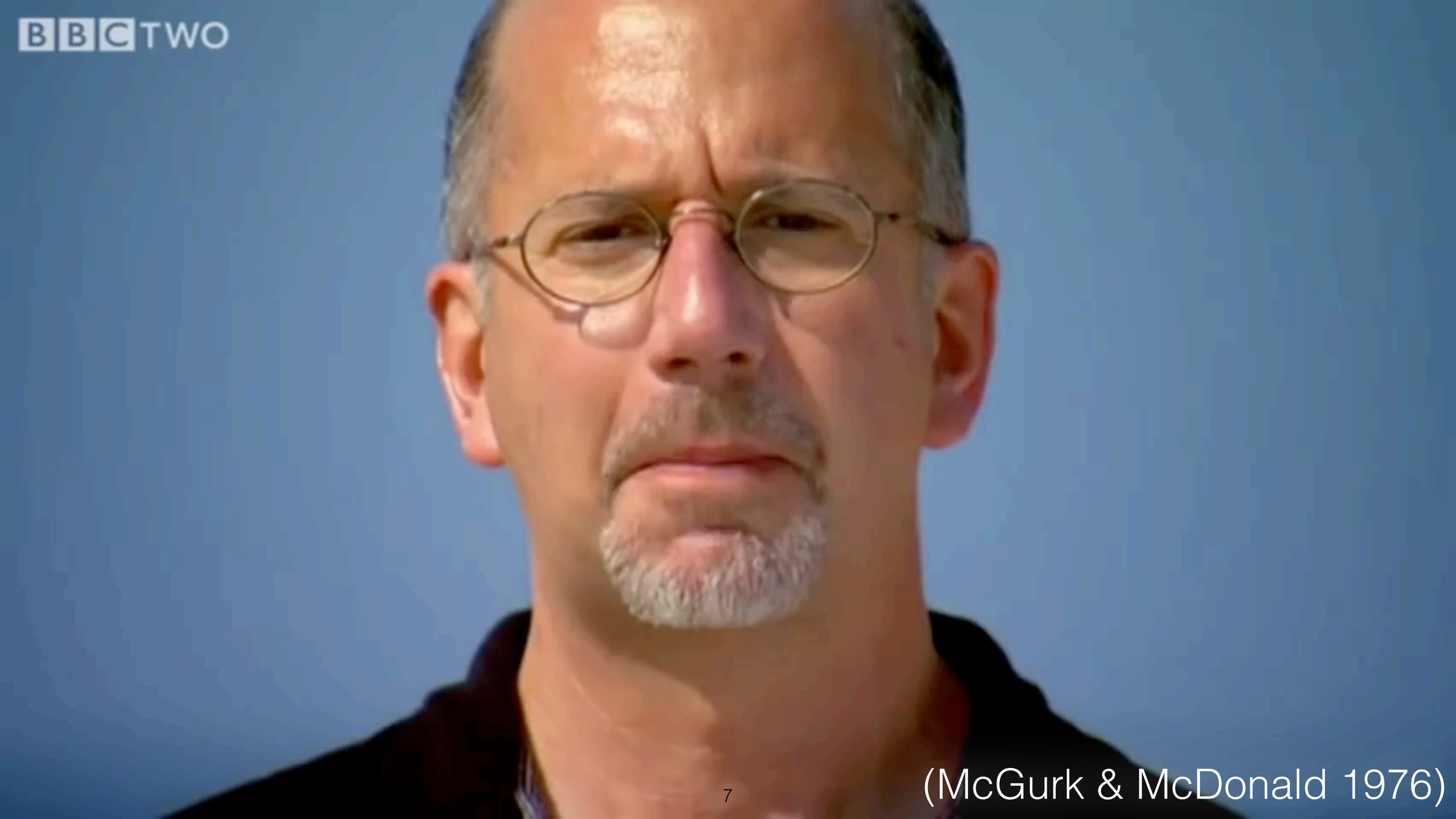
- PS2 grades out
- Remember: strict deadline on regrade requests
  - 9 days from today
- PS7: vision + language

“A coffee cup resting on a table in a cafe”

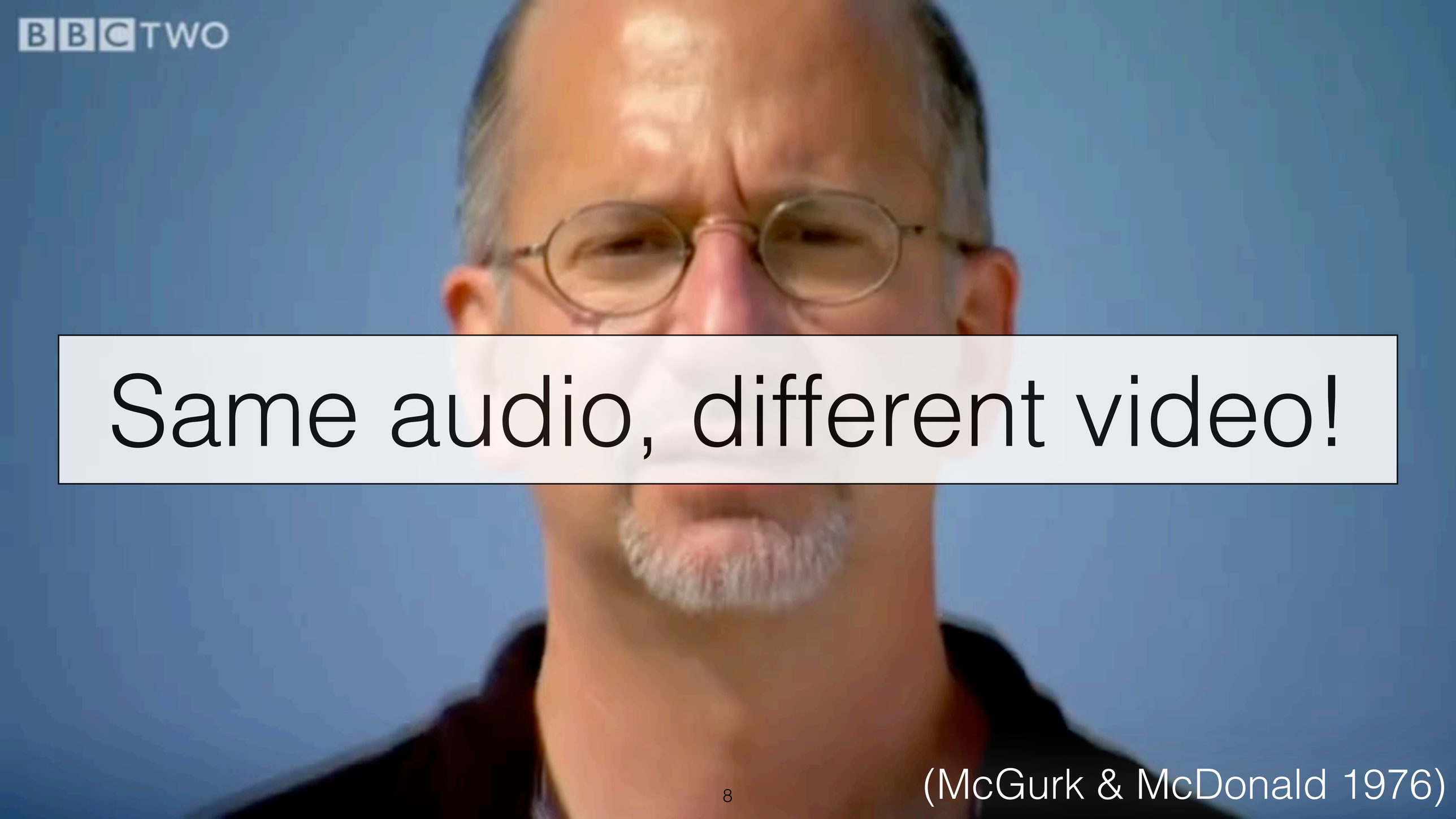


SHIPPING
RECEIVING

McGurk effect: https://www.youtube.com/watch?v=2k8fHR9jKVM



(McGurk & McDonald 1976)

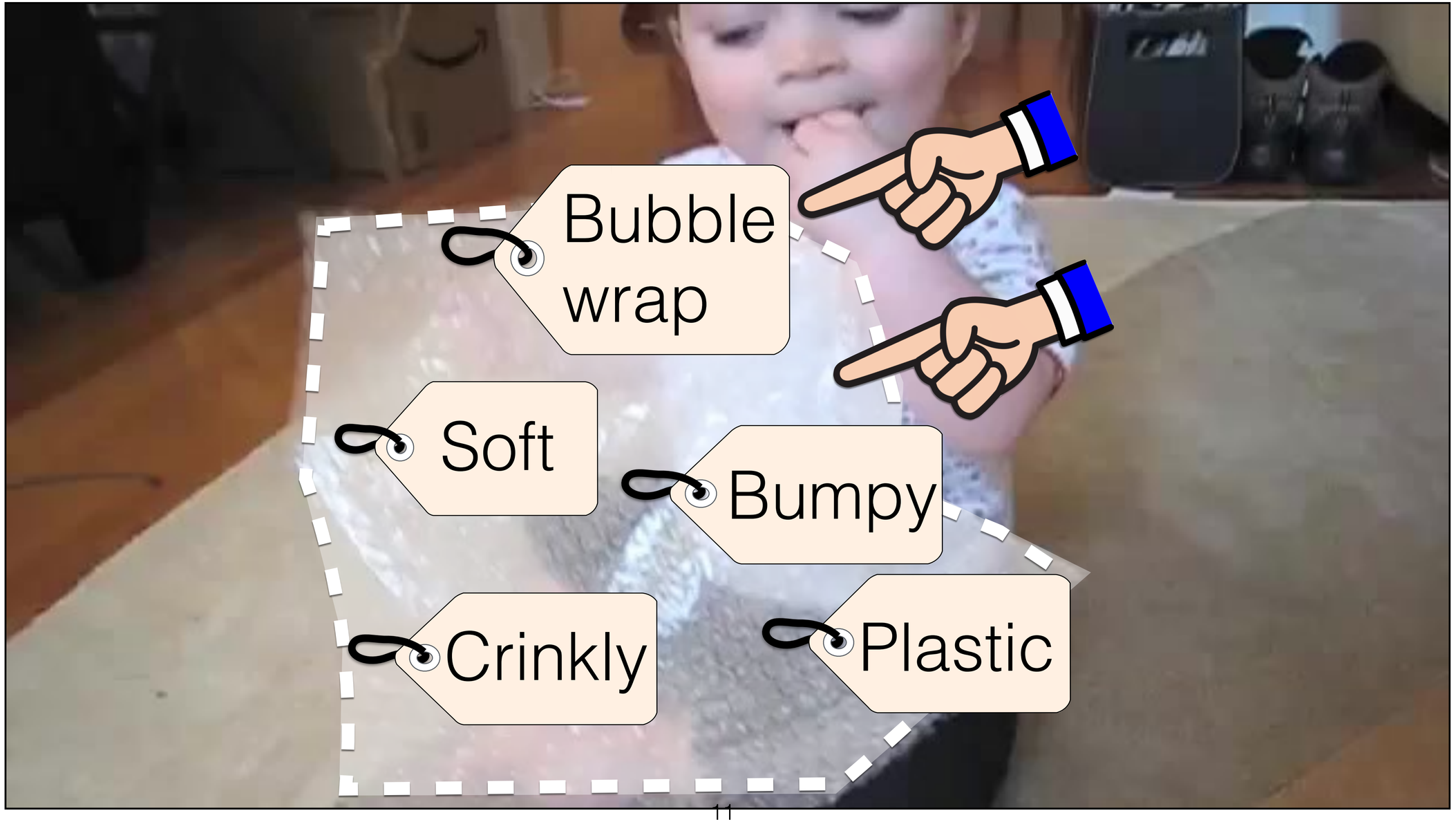


Same audio, different video!

(McGurk & McDonald 1976)

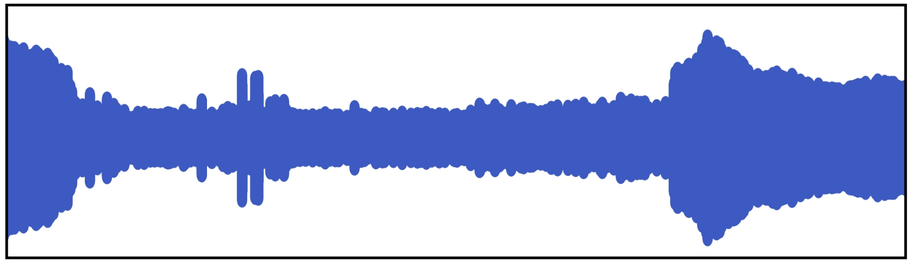
(McGurk & McDonald 1976)

Supervised learning



Human supervision

Self-supervision

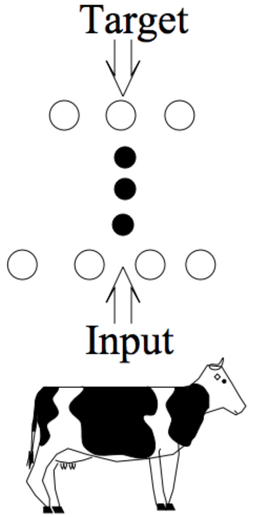


Multisensory self-supervision

Supervised

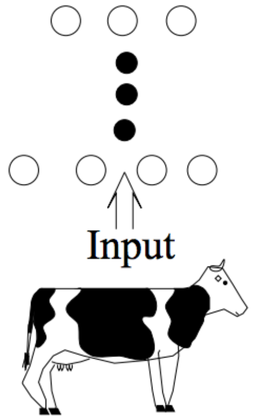
- implausible label

"COW"



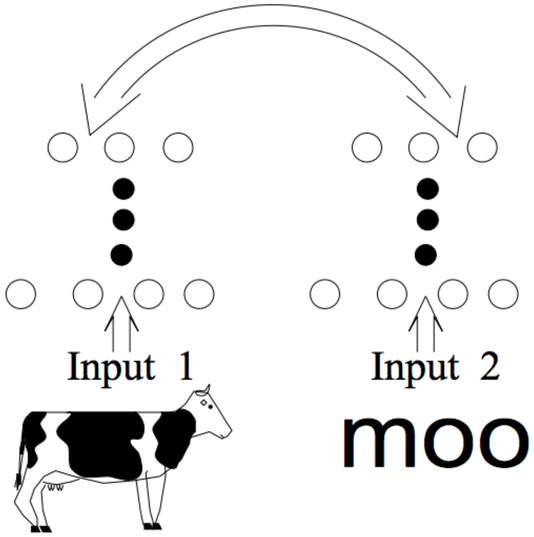
Unsupervised

- limited power



Self-Supervised

- derives label from a co-occurring input to another modality



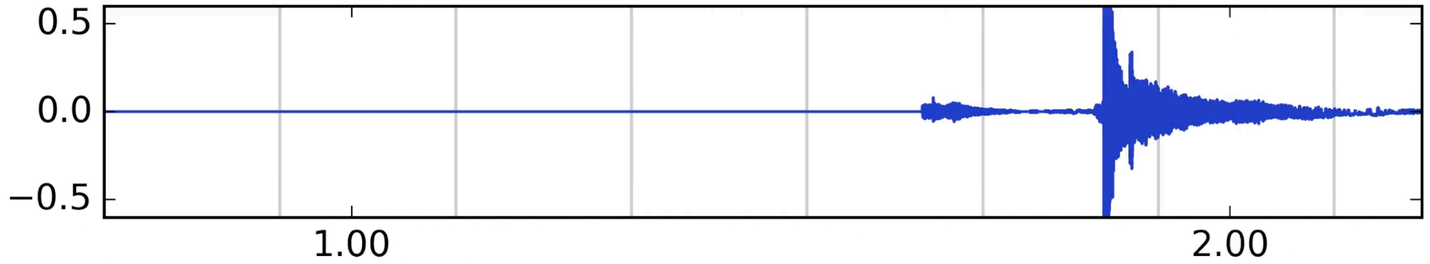
Virginia de Sa. Learning Classification with Unlabeled Data. NIPS 1994.

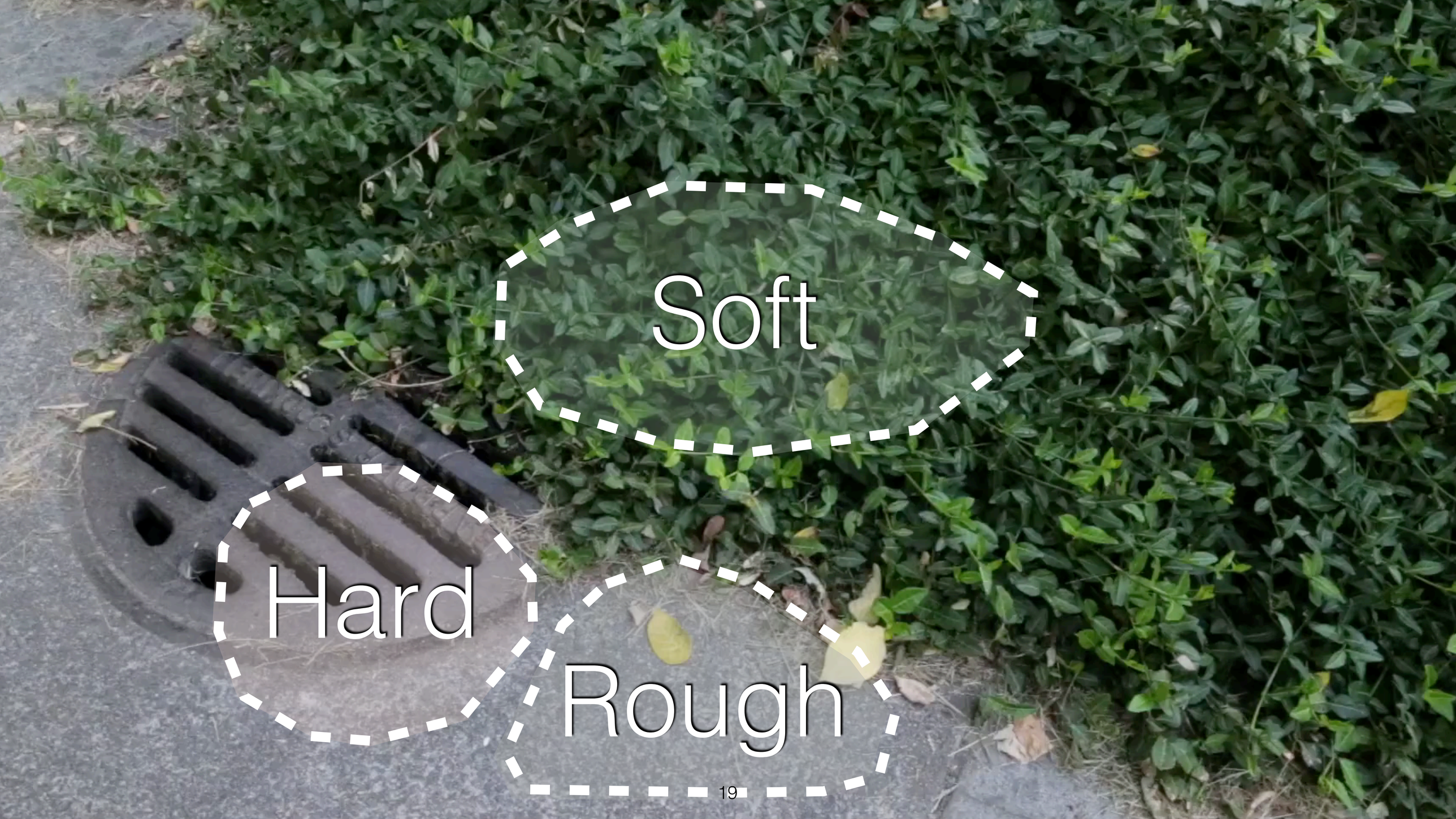
Multimodal models

- Have access to “orthogonal” sources of information.
- Can exploit cues in from multiple senses
- Can use one modality to train another, using self-supervision
- Can be more robust to sensor failures

Multimodal models

- Have access to “orthogonal” sources of information.
- Can exploit cues in from multiple senses
- **Can use one modality to train another, using self-supervision**
- Can be more robust to sensor failures


0.5
0.0
−0.5
1.00
2.00


Soft
Hard
Rough

Learning about physical interactions

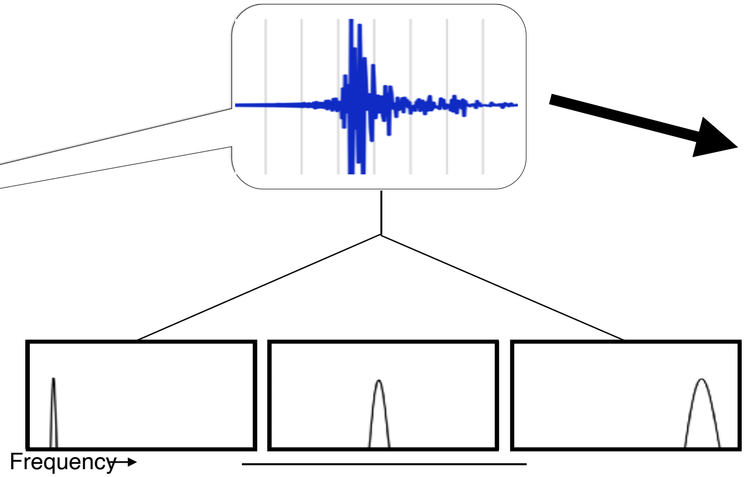
Predicted soundtrack

Silent video

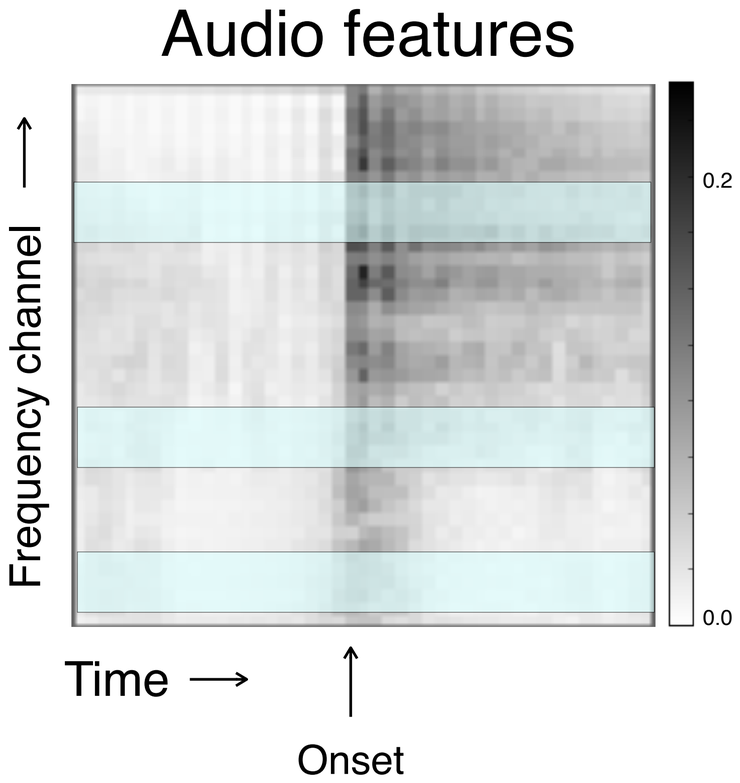
A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, W. T. Freeman. Visually Indicated Sounds. CVPR 2016.

The *Greatest Hits* Dataset: Volume 1

Can we predict physical properties from sound?



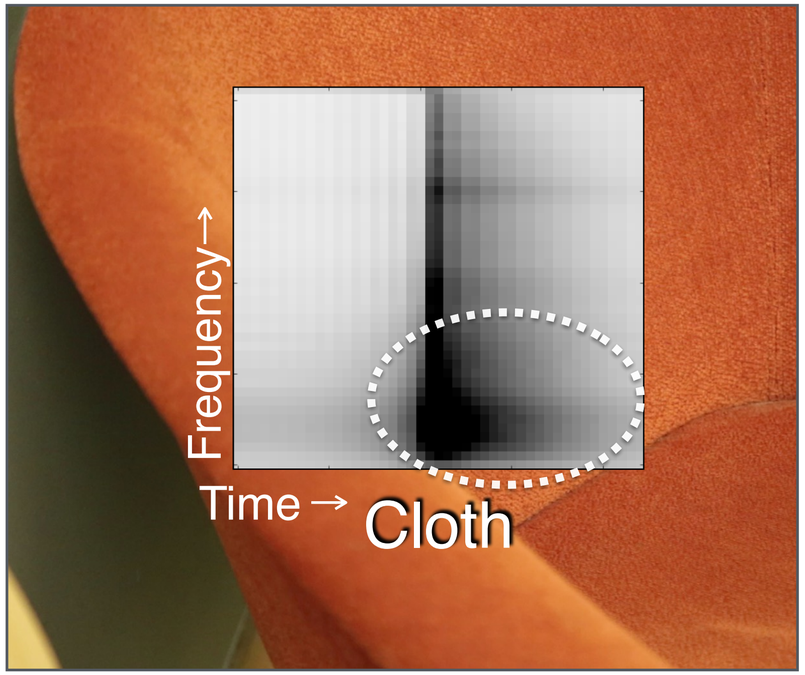
Filters tuned to different frequencies

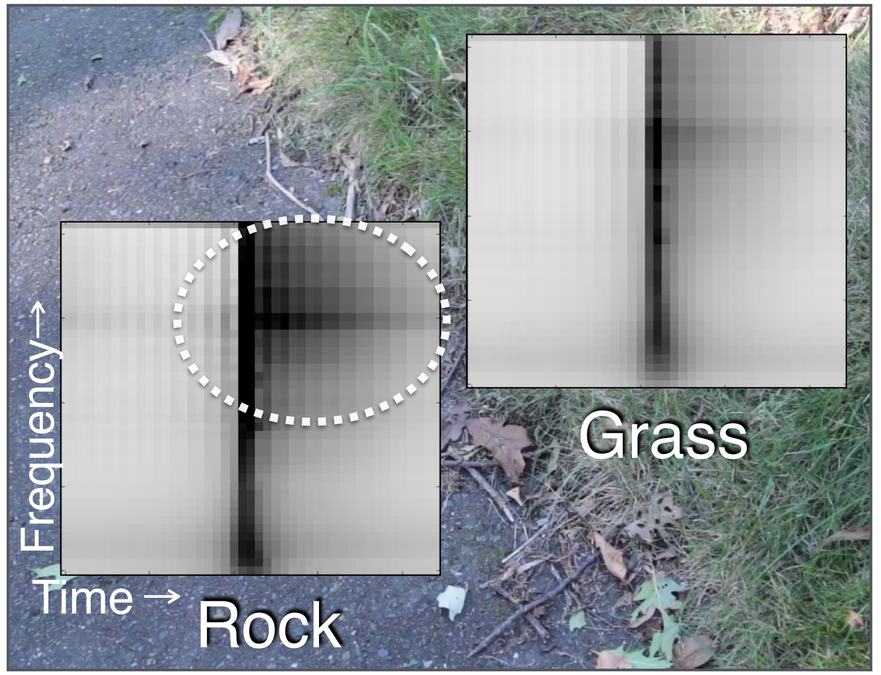


- 40 bandpass filters (+ high/low pass)

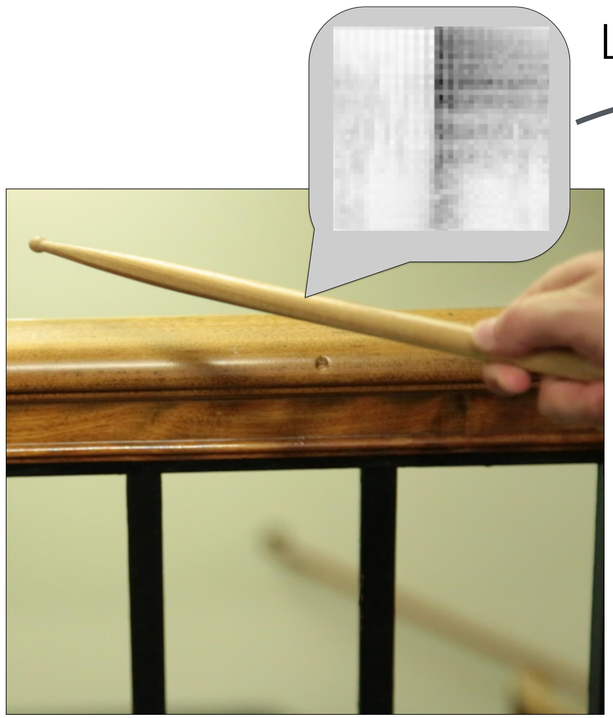
Can we predict physical properties from sound?

Mean sound features per category



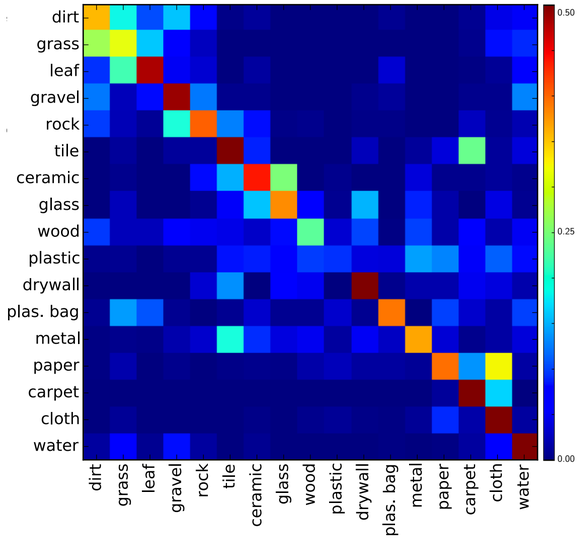


Can we predict physical properties from sound?



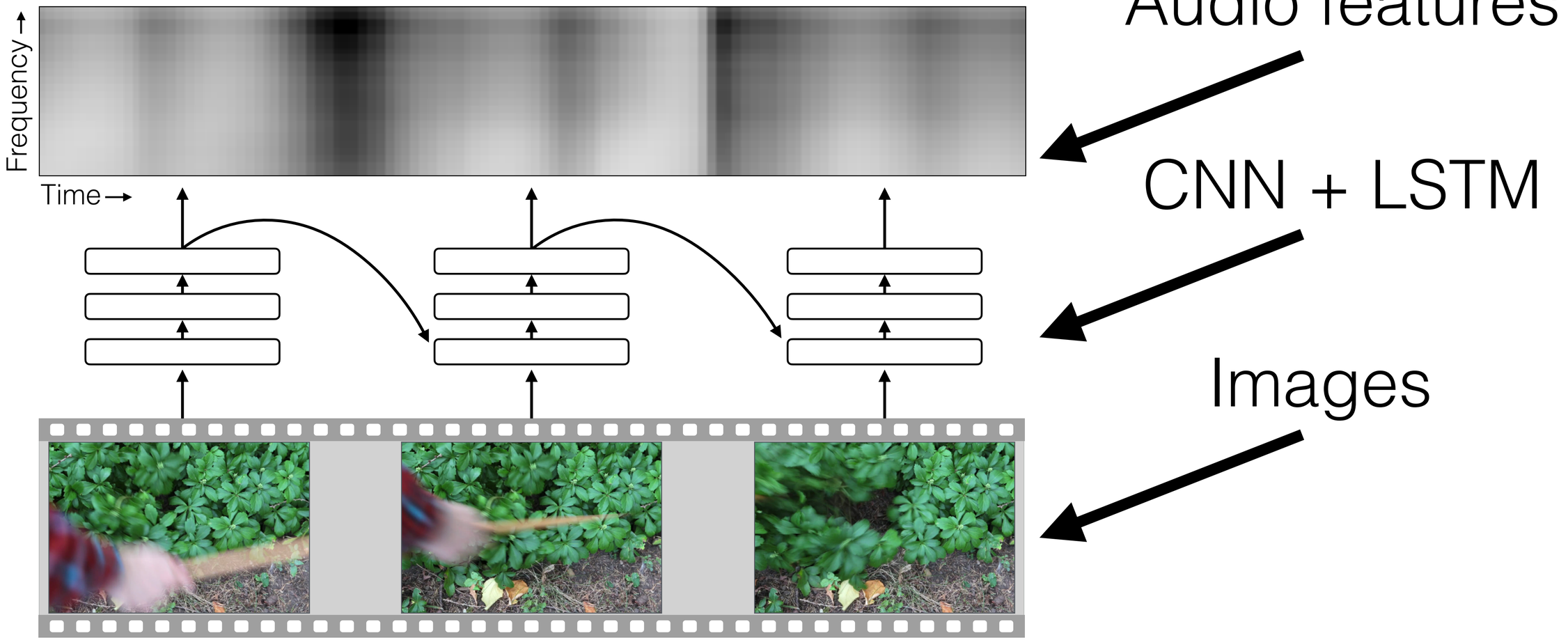
Linear classifier

“Wood”

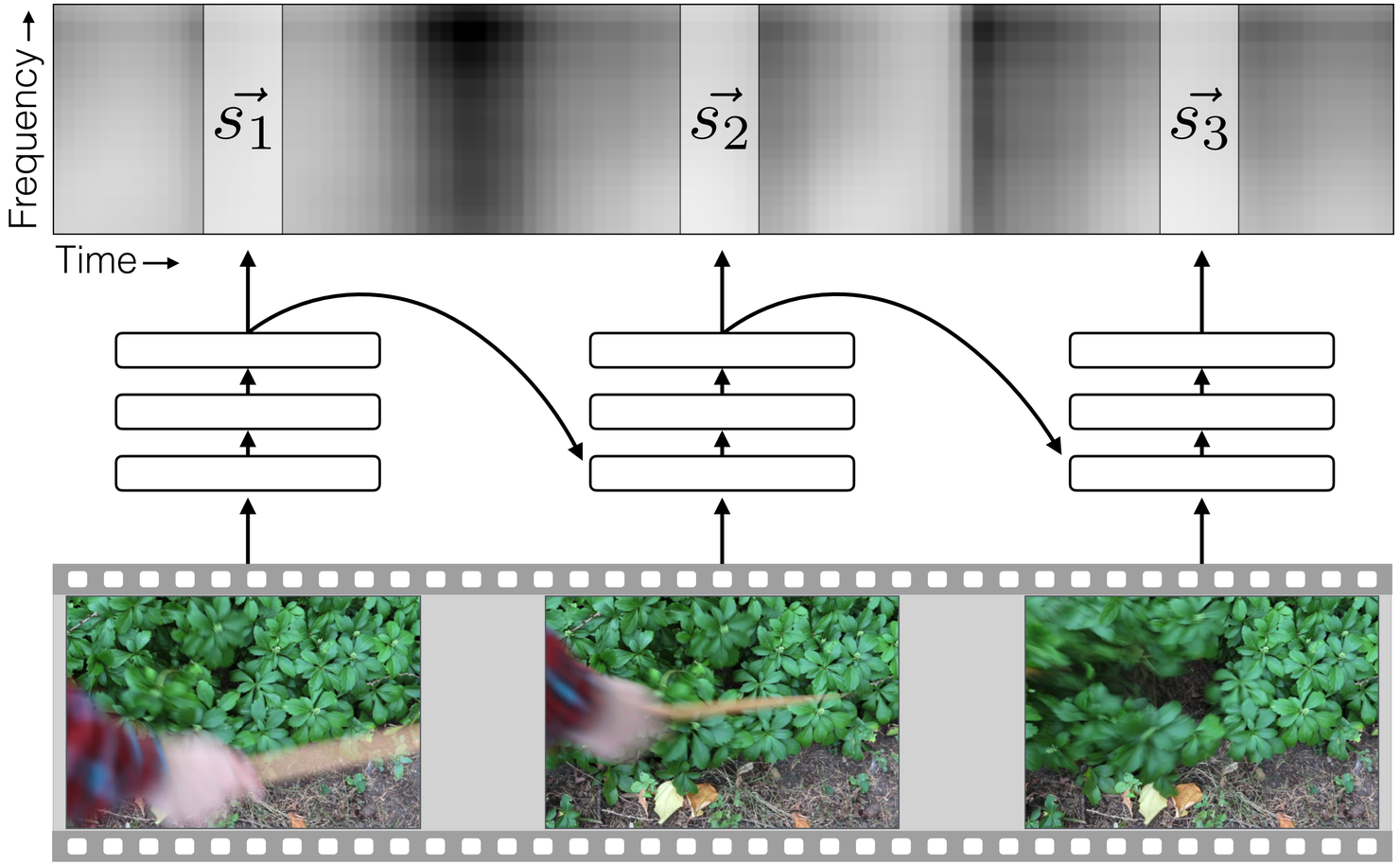


- 46% class-averaged accuracy (chance = 6%)
- Common audio confusions:
  - {cloth, paper}, {dirt, grass}, {rock, tile}

Predicting visually indicated sounds



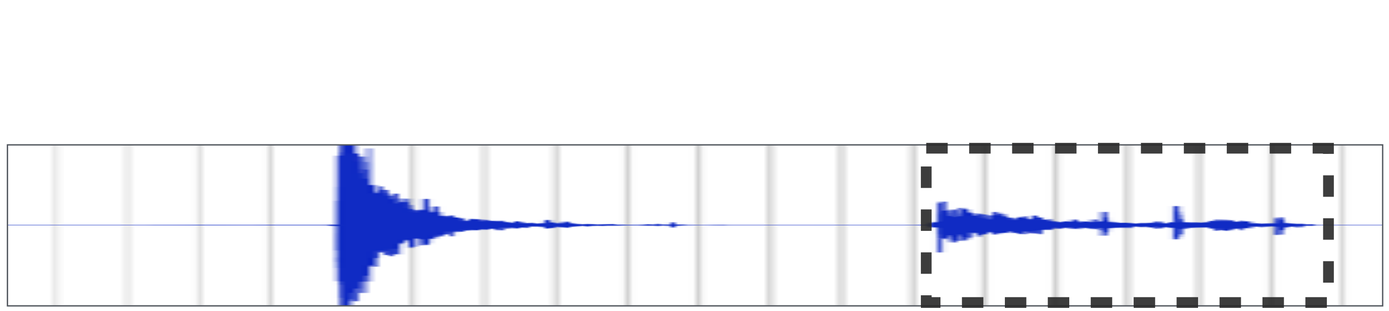
Predicting visually indicated sounds

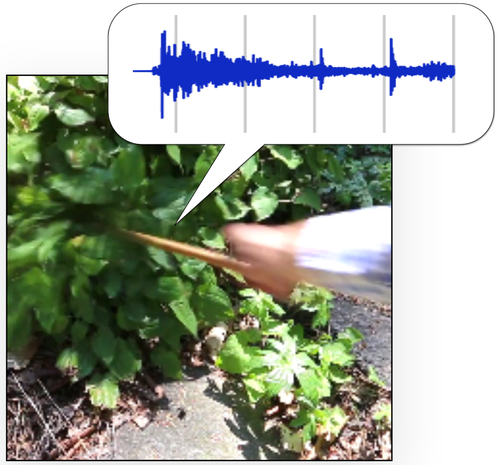


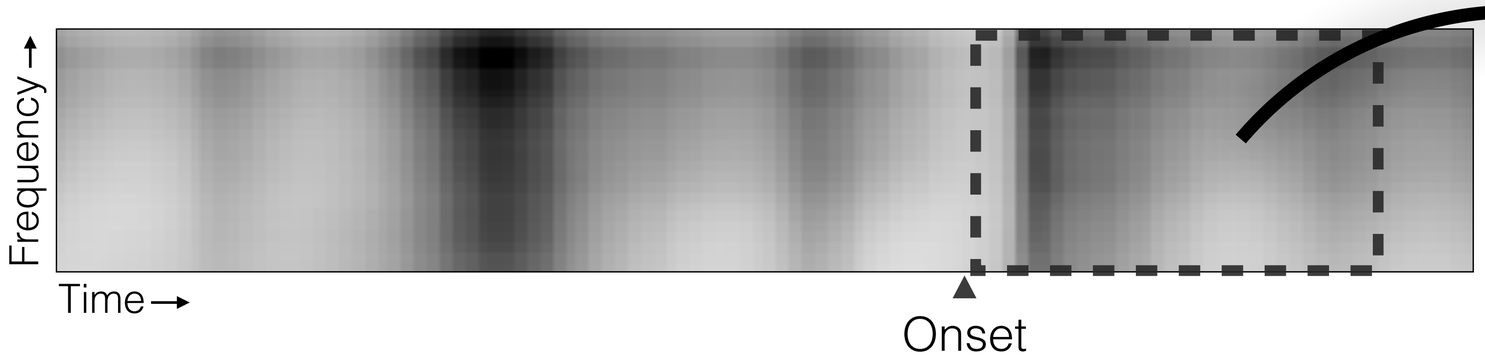
Ground truth

$$\sum_{t=1}^{T} \|\vec{s}_t - \tilde{\vec{s}}_t\|$$

Predicting visually indicated sounds







Sound database

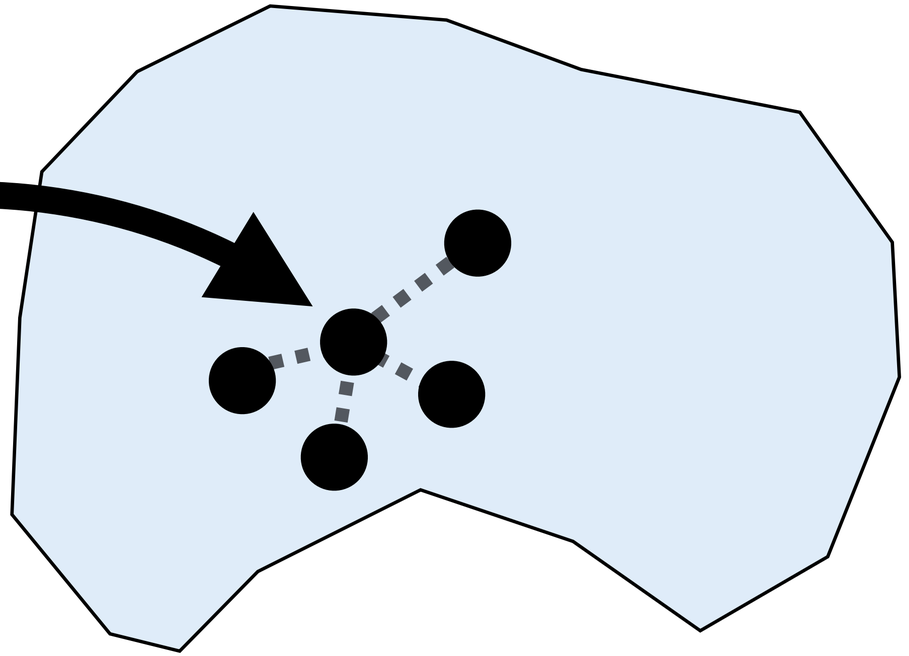
Predicted sound

Predicted sound

Predicted sound

Failure case

Translating between modalities



Audio space

Translating between modalities

Audio nearest neighbors have similar material properties

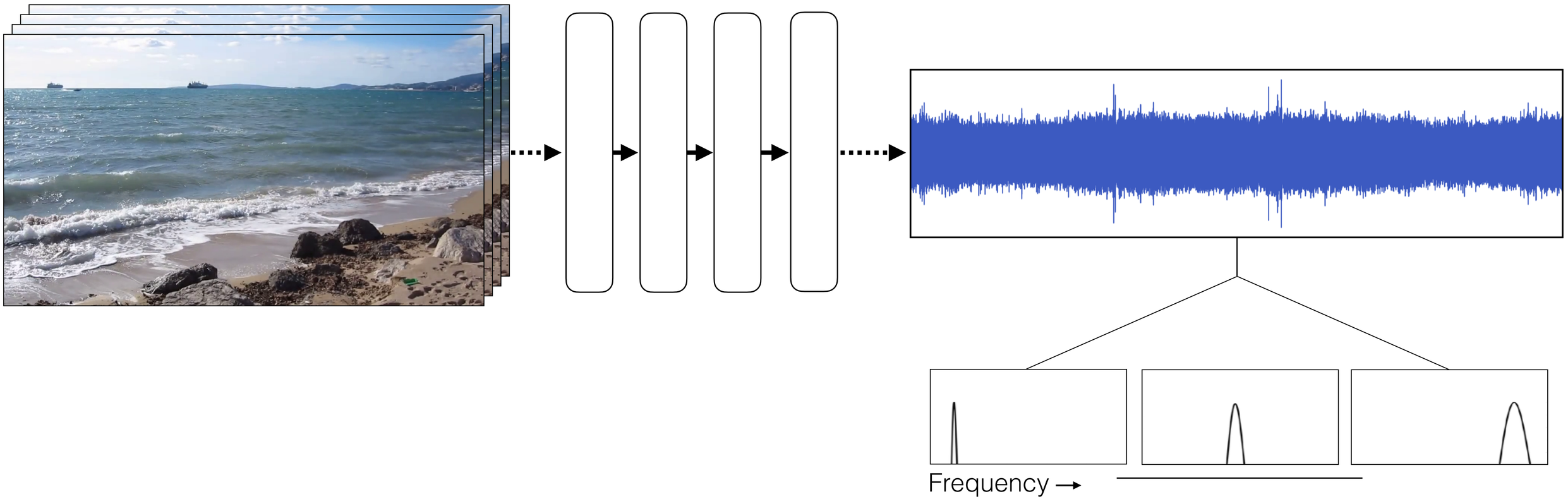
Active vs. passive perception

A. Owens, J. Wu, J. H. McDermott, A. Torralba, W. T. Freeman.
Ambient Sound Provides Supervision for Visual Learning. ECCV 2016.

Active vs. passive perception

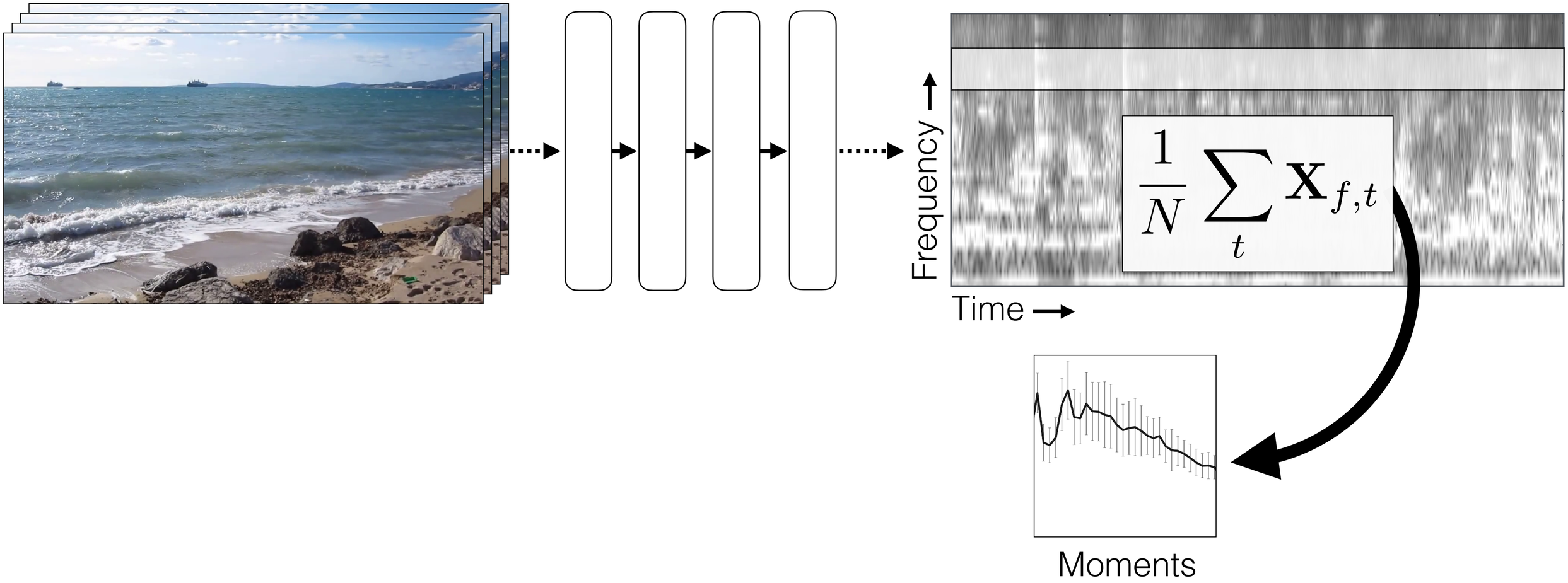
A. Owens, J. Wu, J. H. McDermott, A. Torralba, W. T. Freeman.
Ambient Sound Provides Supervision for Visual Learning. ECCV 2016.

Predicting ambient sound

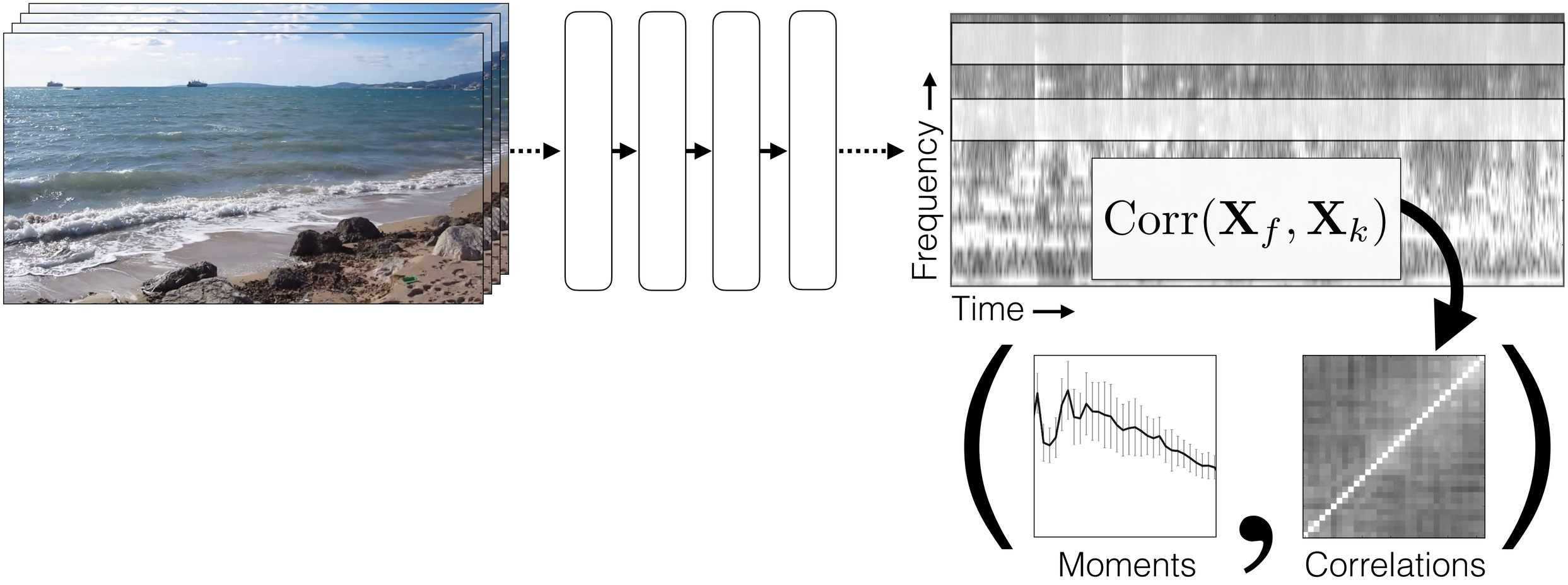


Filters tuned to different frequencies

Predicting ambient sound

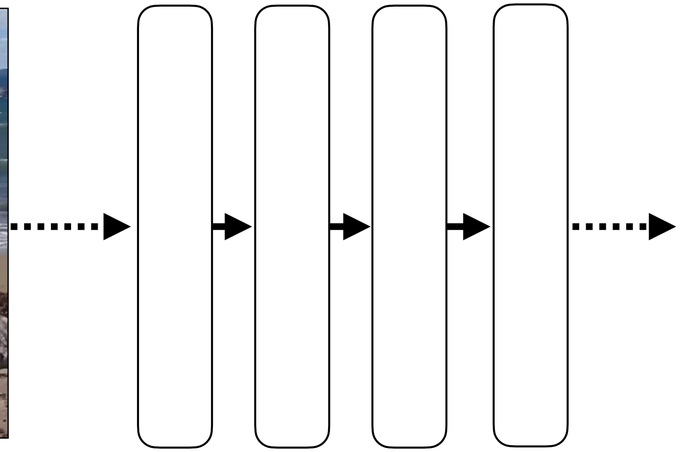


Predicting ambient sound

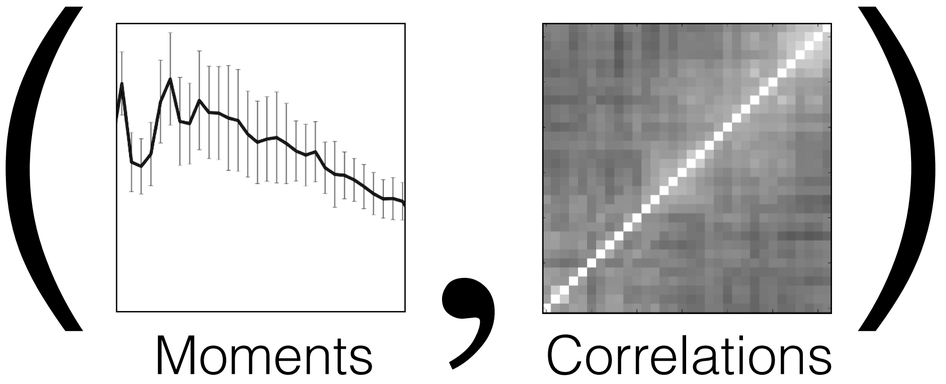


Predicting ambient sound

Video frame



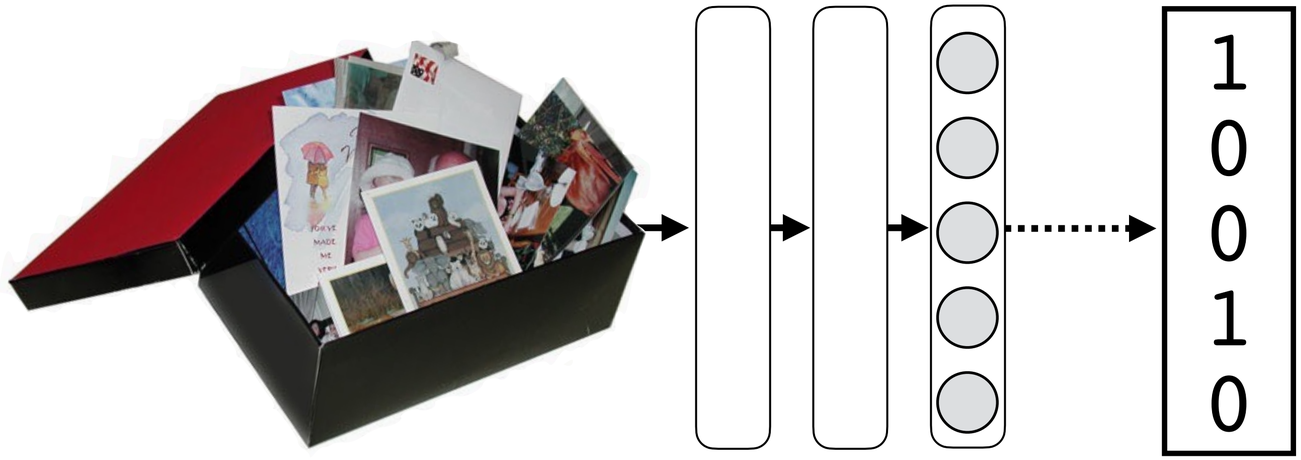
Convolutional network



Sound texture

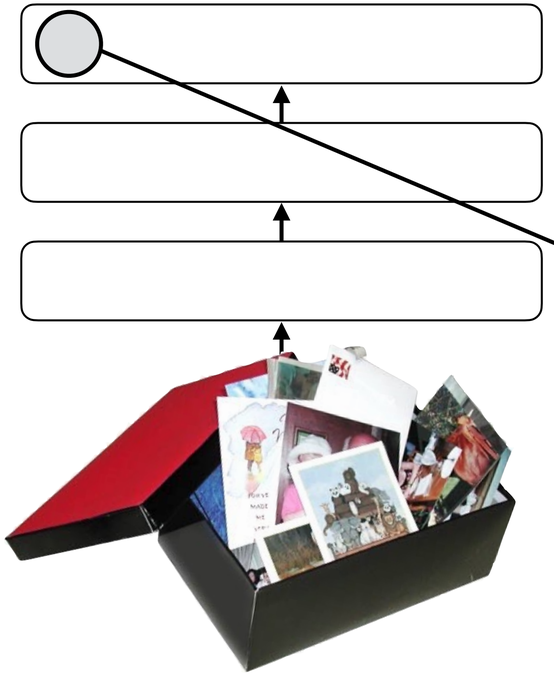
(McDermott & Simoncelli 2011)

What did the model learn?



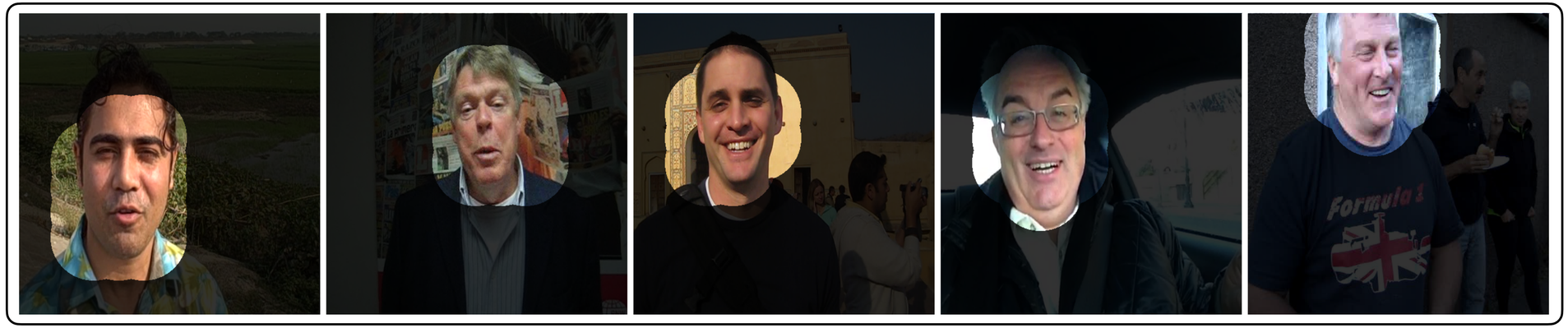
What did the model learn?

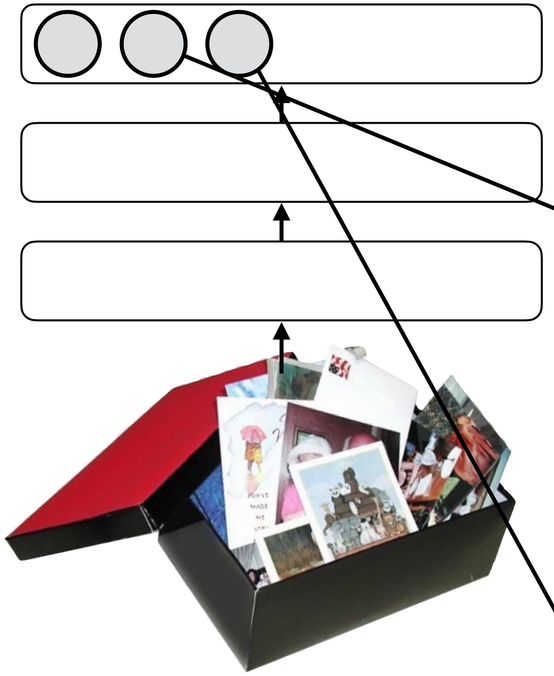
Unit #90 of 256

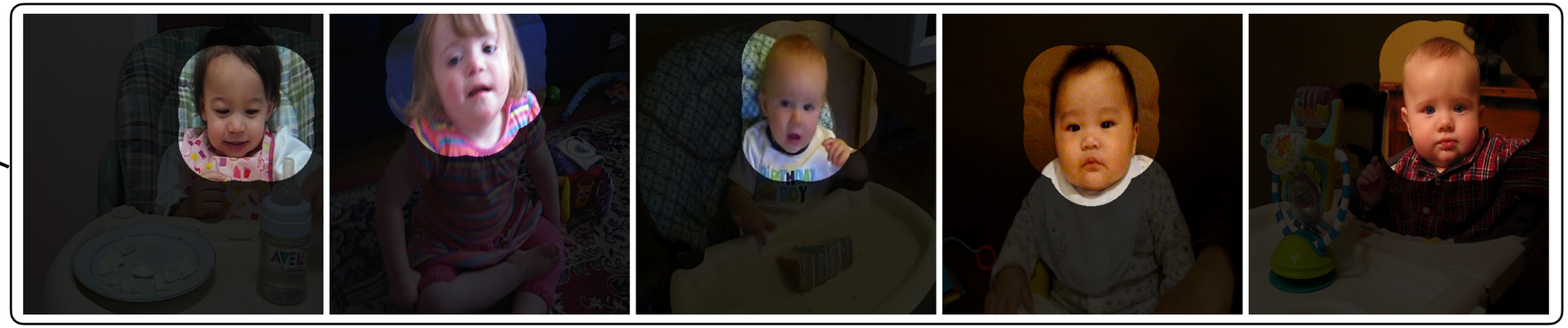


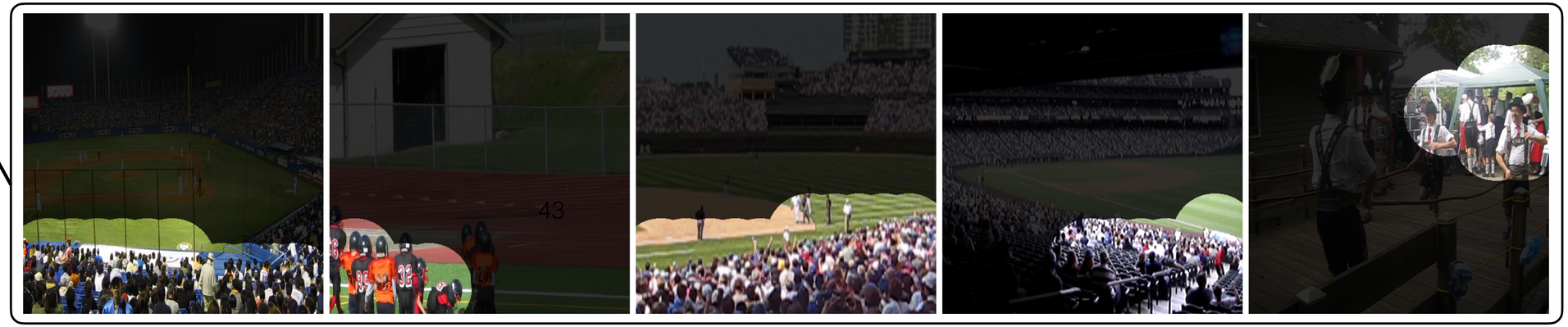
Strongest responses in dataset

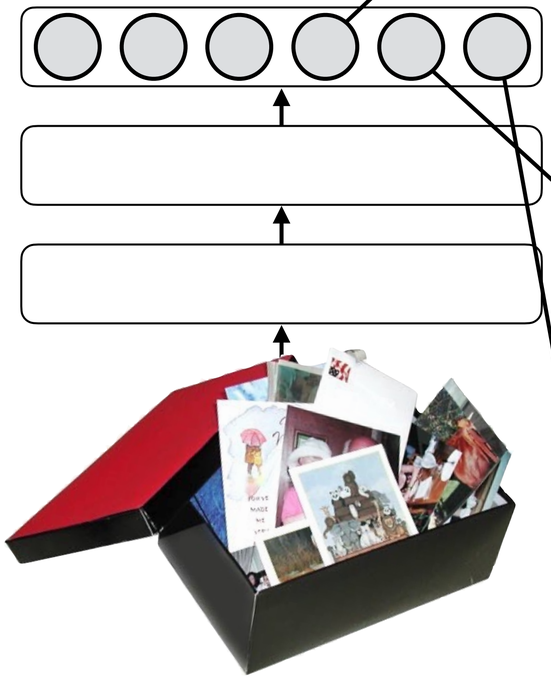
Visualization method from (Zhou 2015)

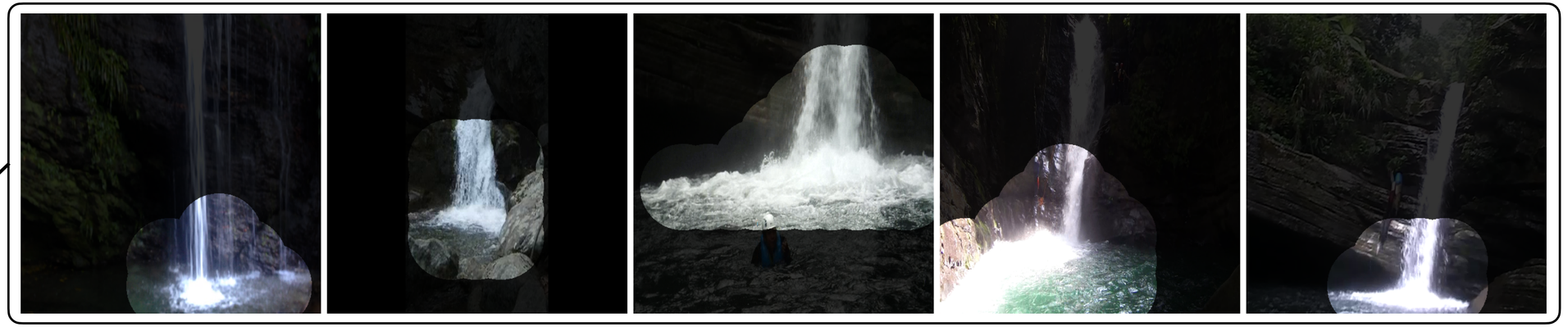




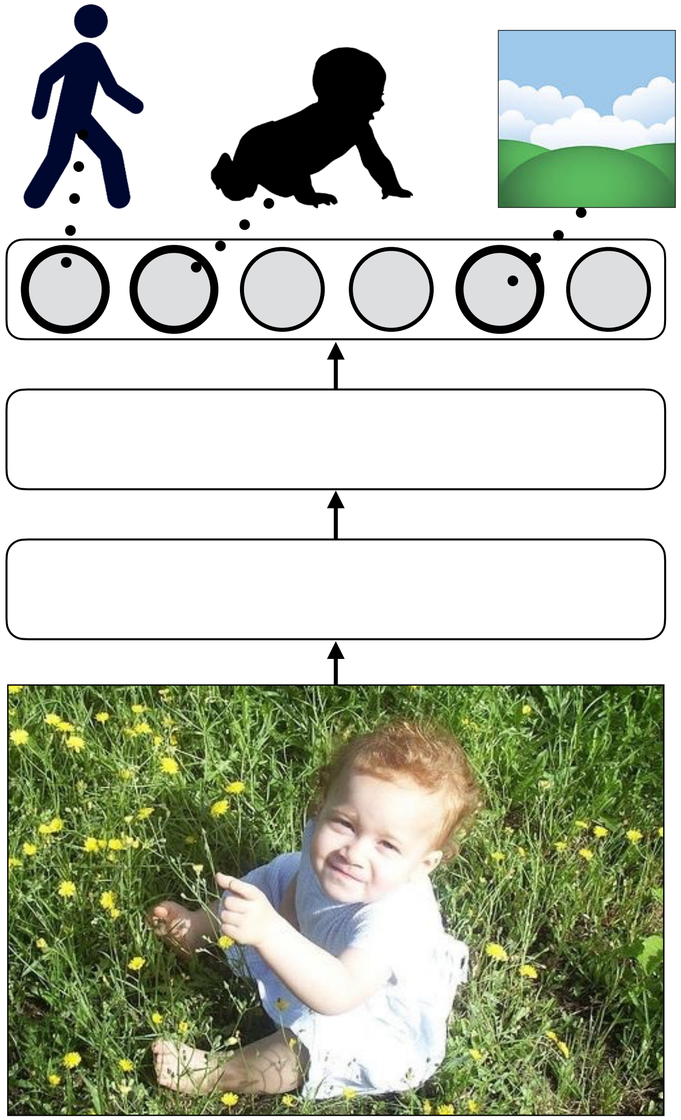




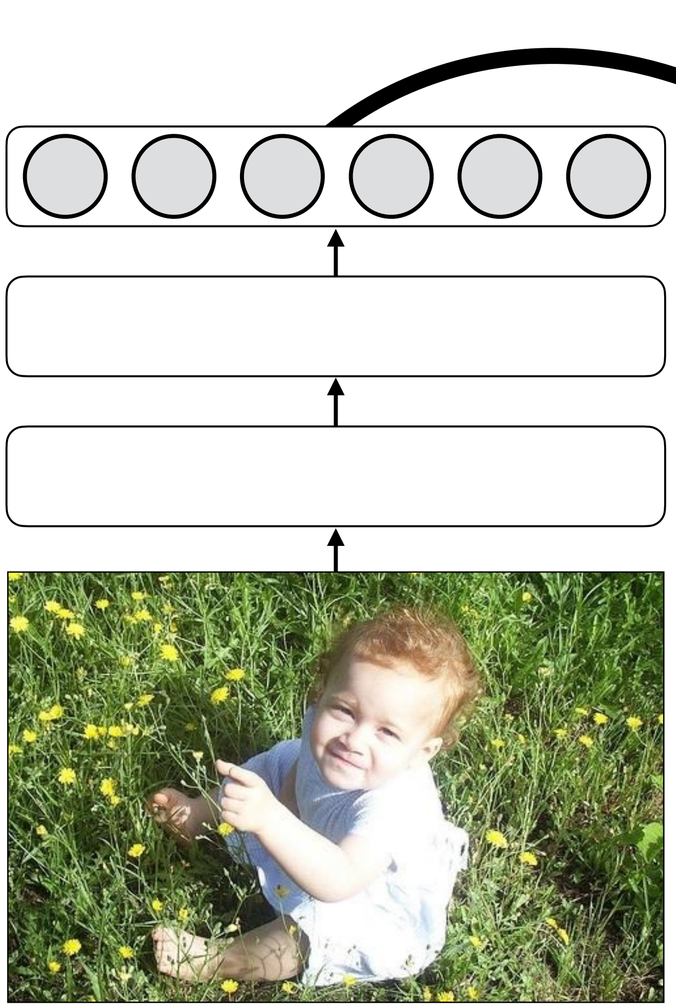




What did the model learn?



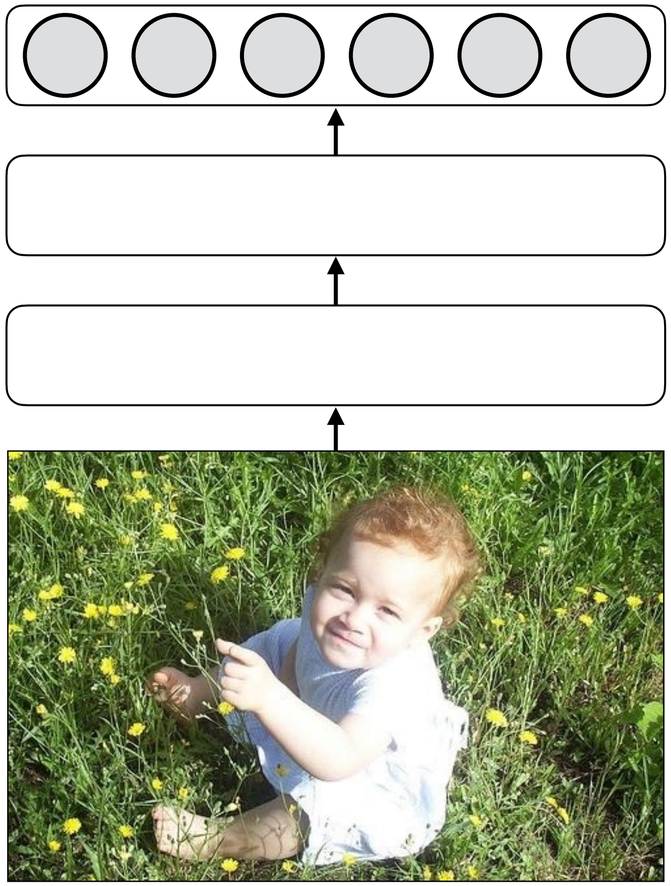
What did the model learn?

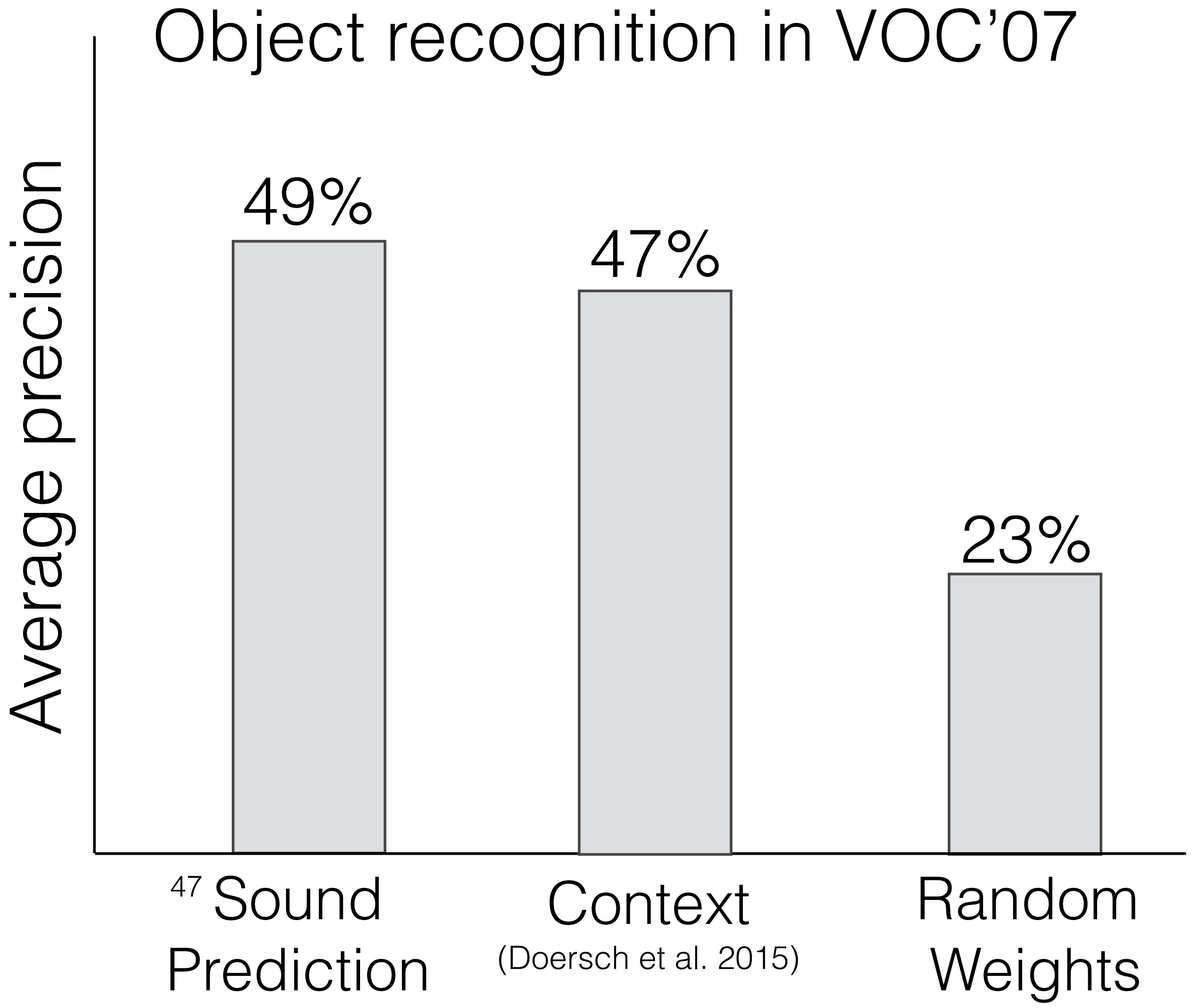


Linear Classifier

“Person”

What did the model learn?

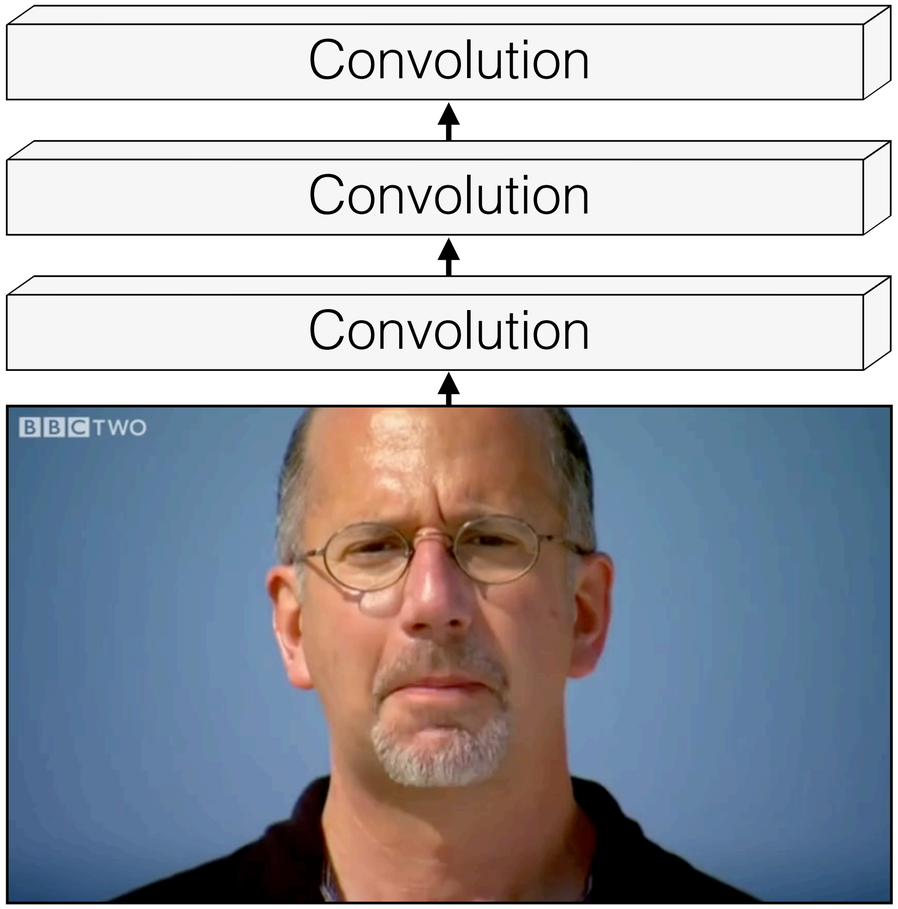




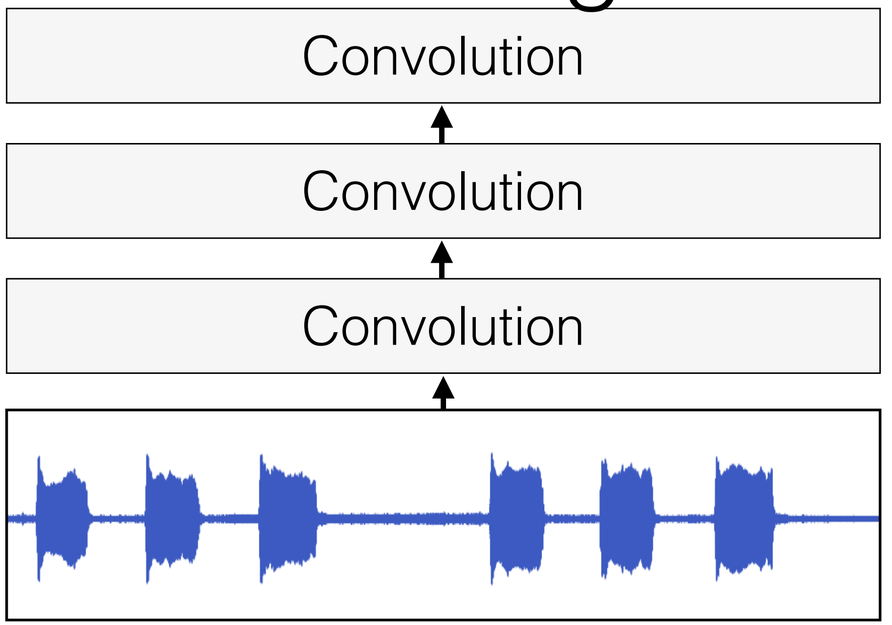
Learning multimodal representations

Single-modality representations

Vision



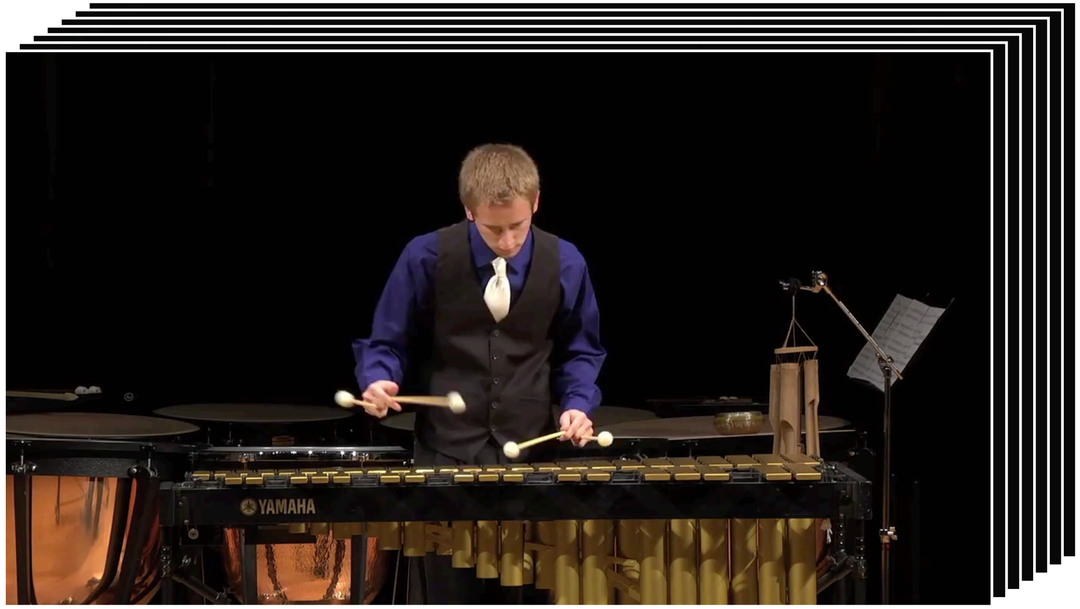
Hearing

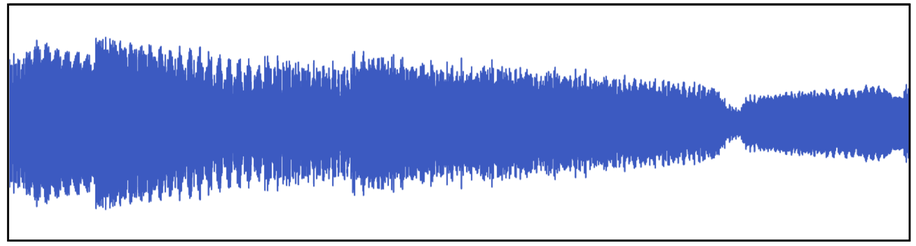


Idea: train a model to find audio-visual correspondences in video.

How do we get ground-truth correspondences?

Learning audio-visual correspondences

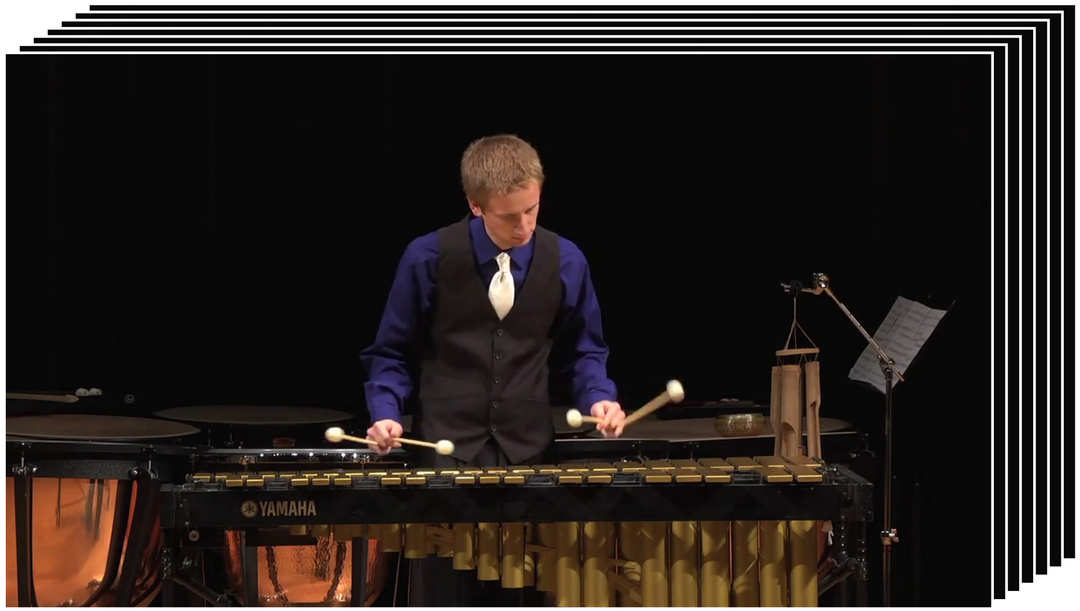
(

, )

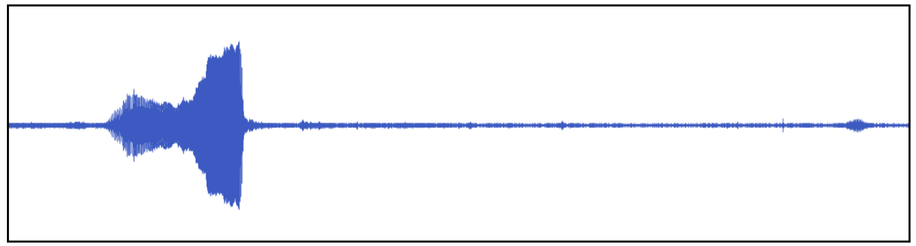
→ **real** or fake ?

Related work: L^3-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen 2010)

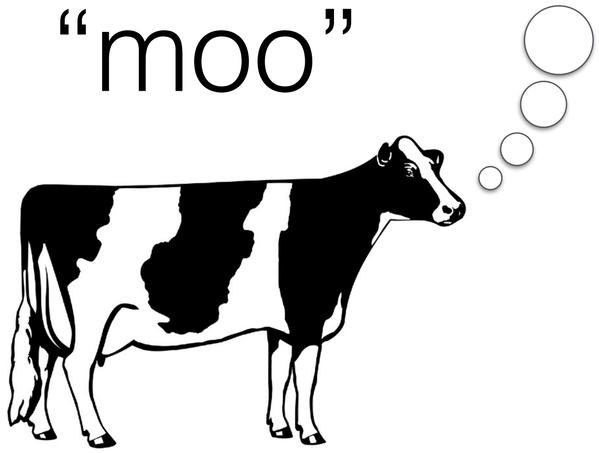
Learning audio-visual correspondences



,



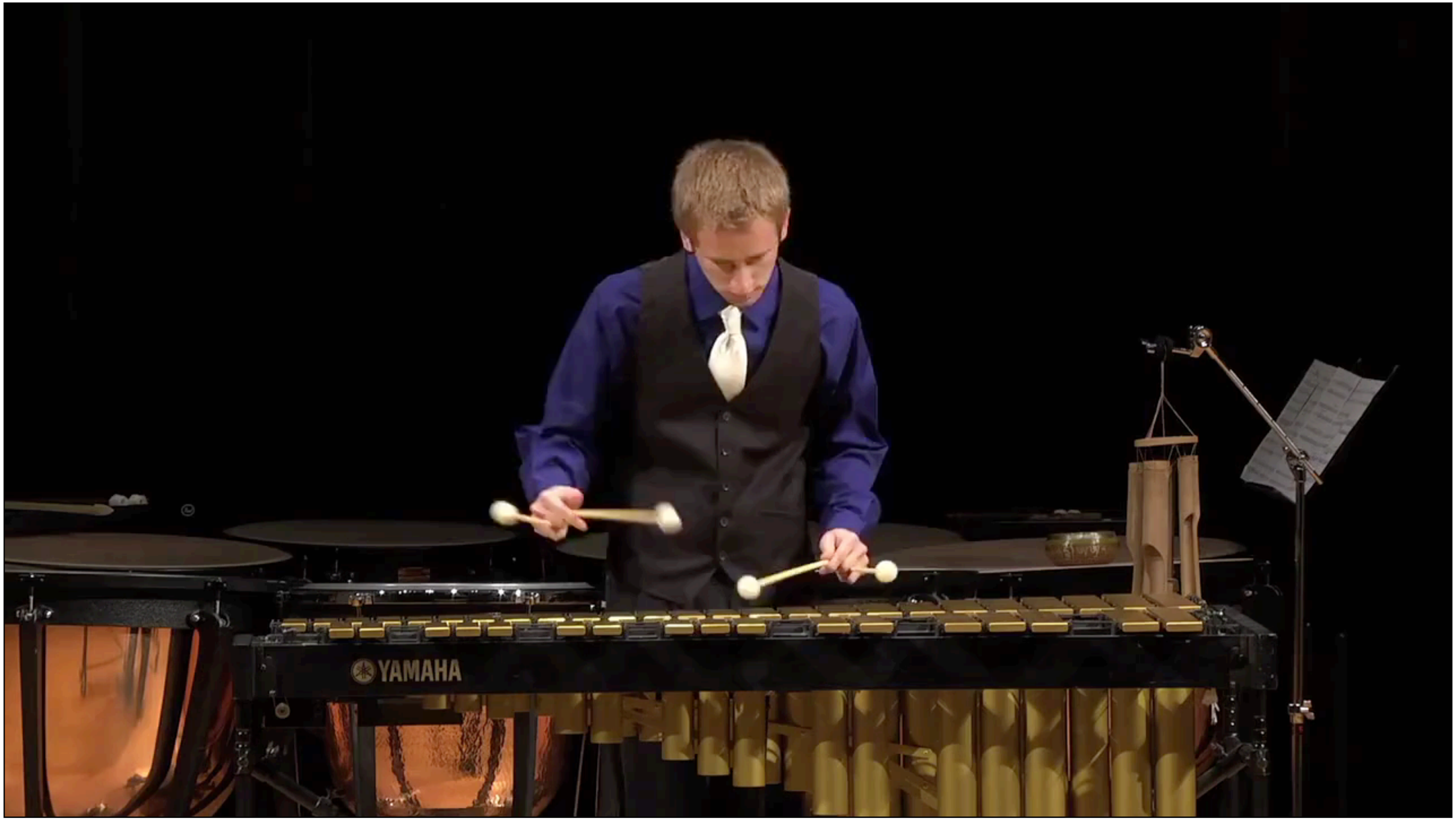
“moo”



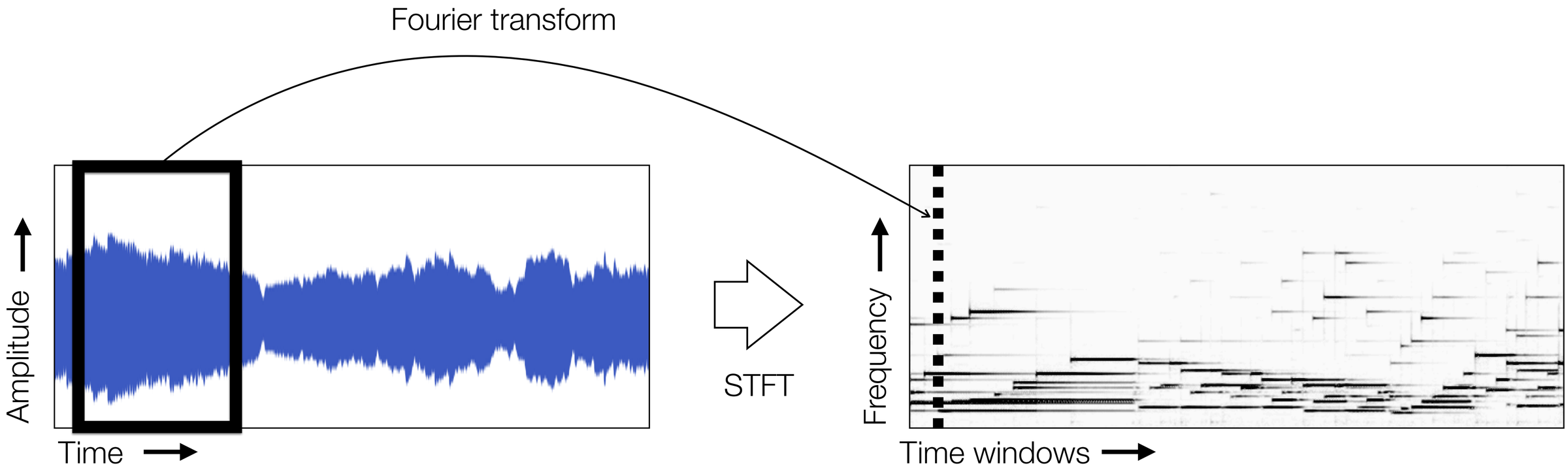
→ real or **fake**?

Related work: L^3-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen 2010)

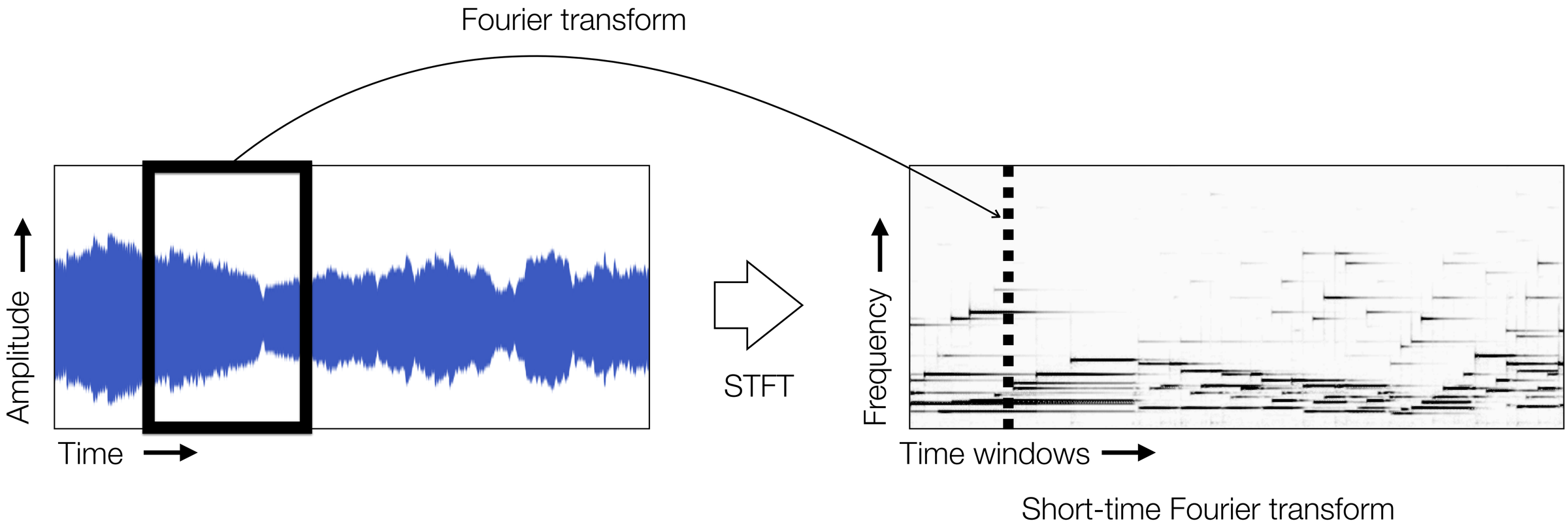
Idea #1: use random audio as negatives



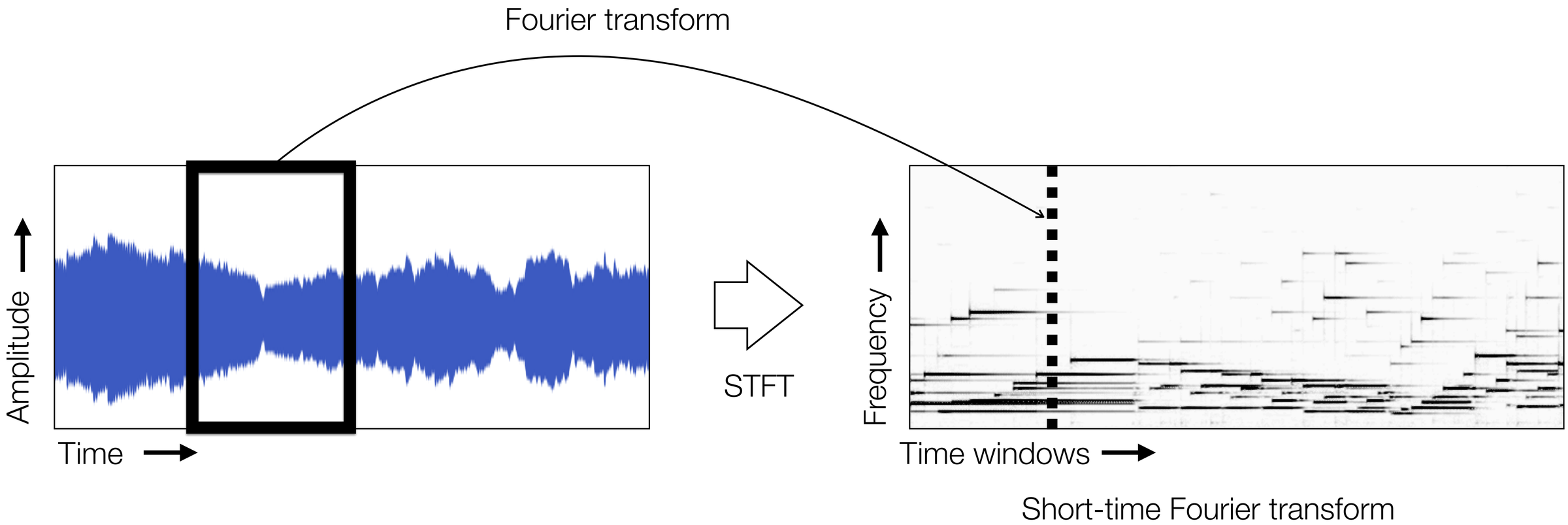
How do we represent the sound?



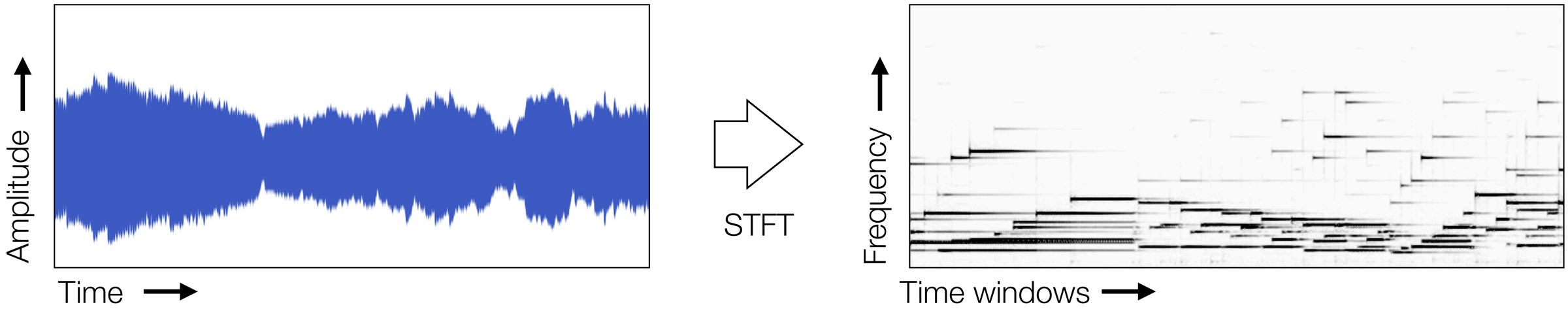
Spectrogram



Spectrogram



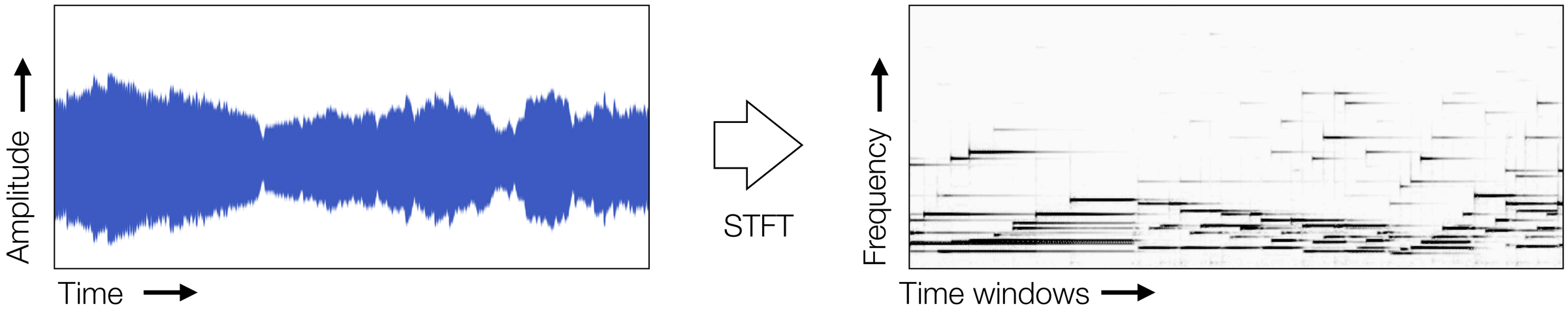
Spectrogram



Short-time Fourier transform

Spectrogram

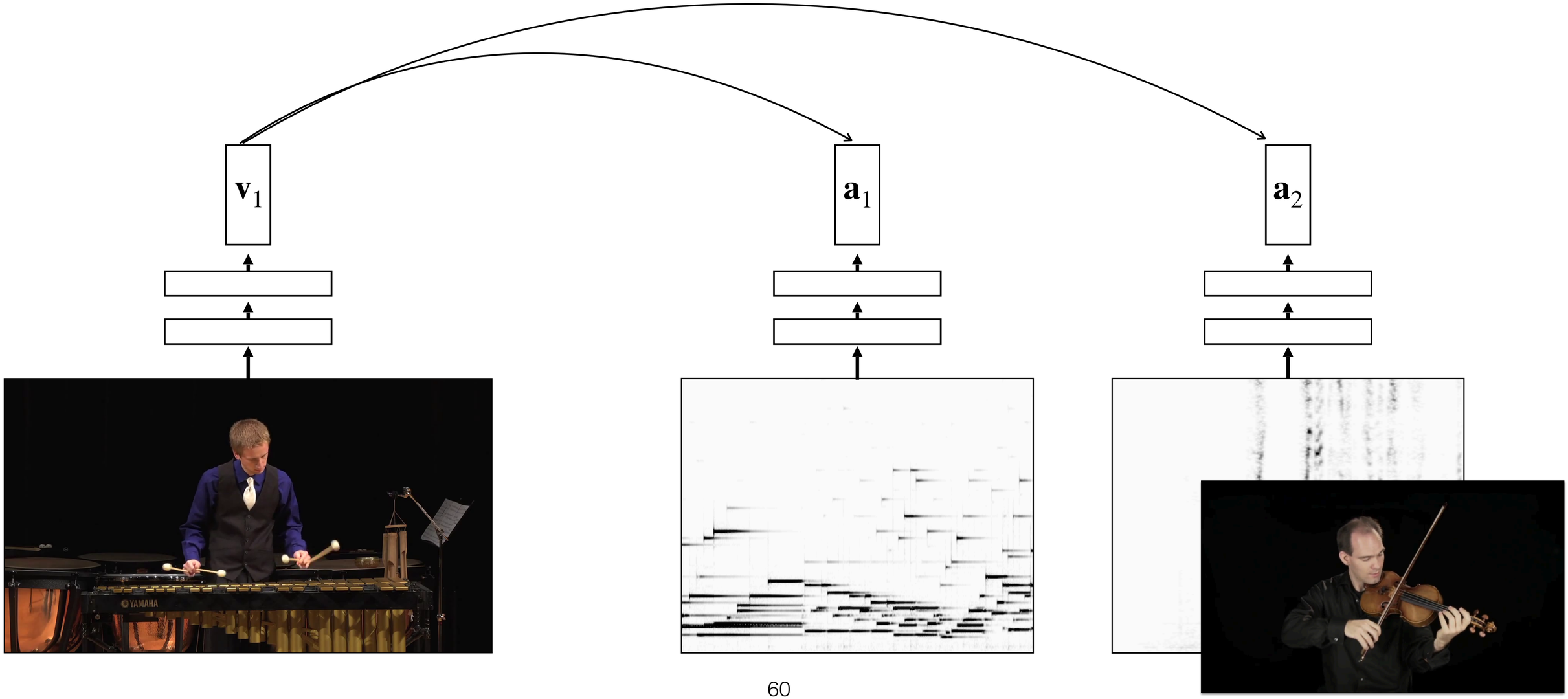
Treat it like an image and process with a CNN!



Short-time Fourier transform

Usually lower dimensional: e.g. 16,000 samples becomes 256 x 256 spectrogram

Contrastive audio-visual learning

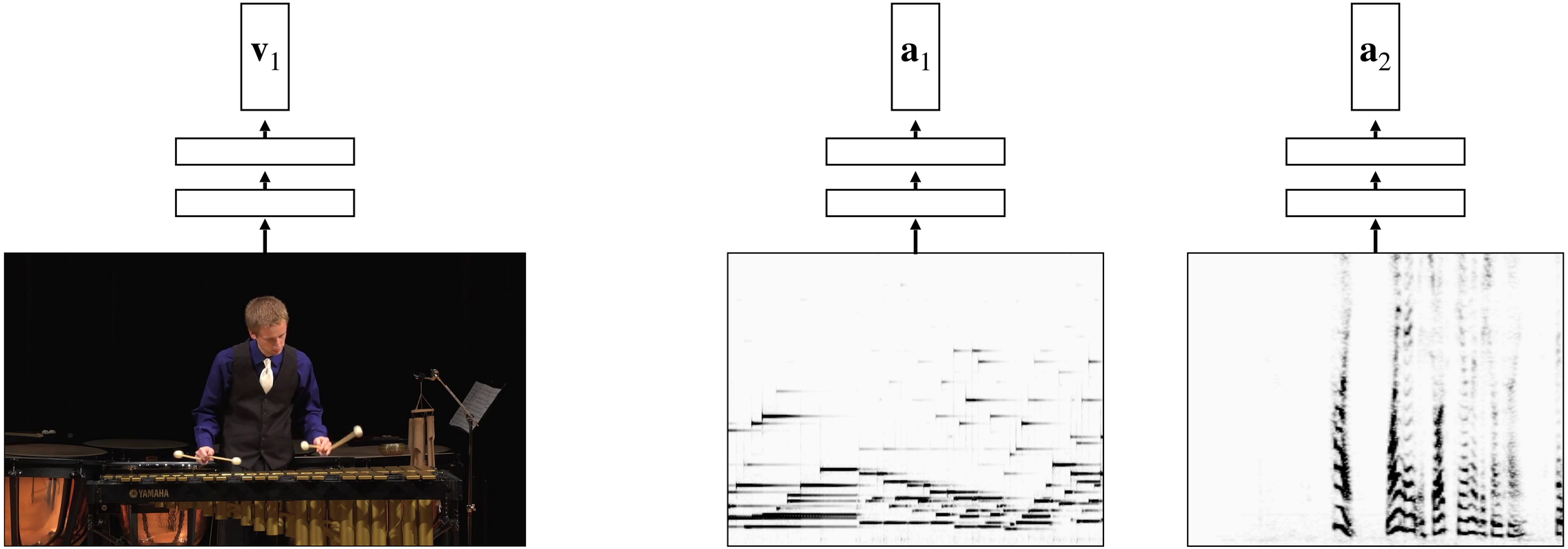


Contrastive audio-visual learning

Learn networks where:

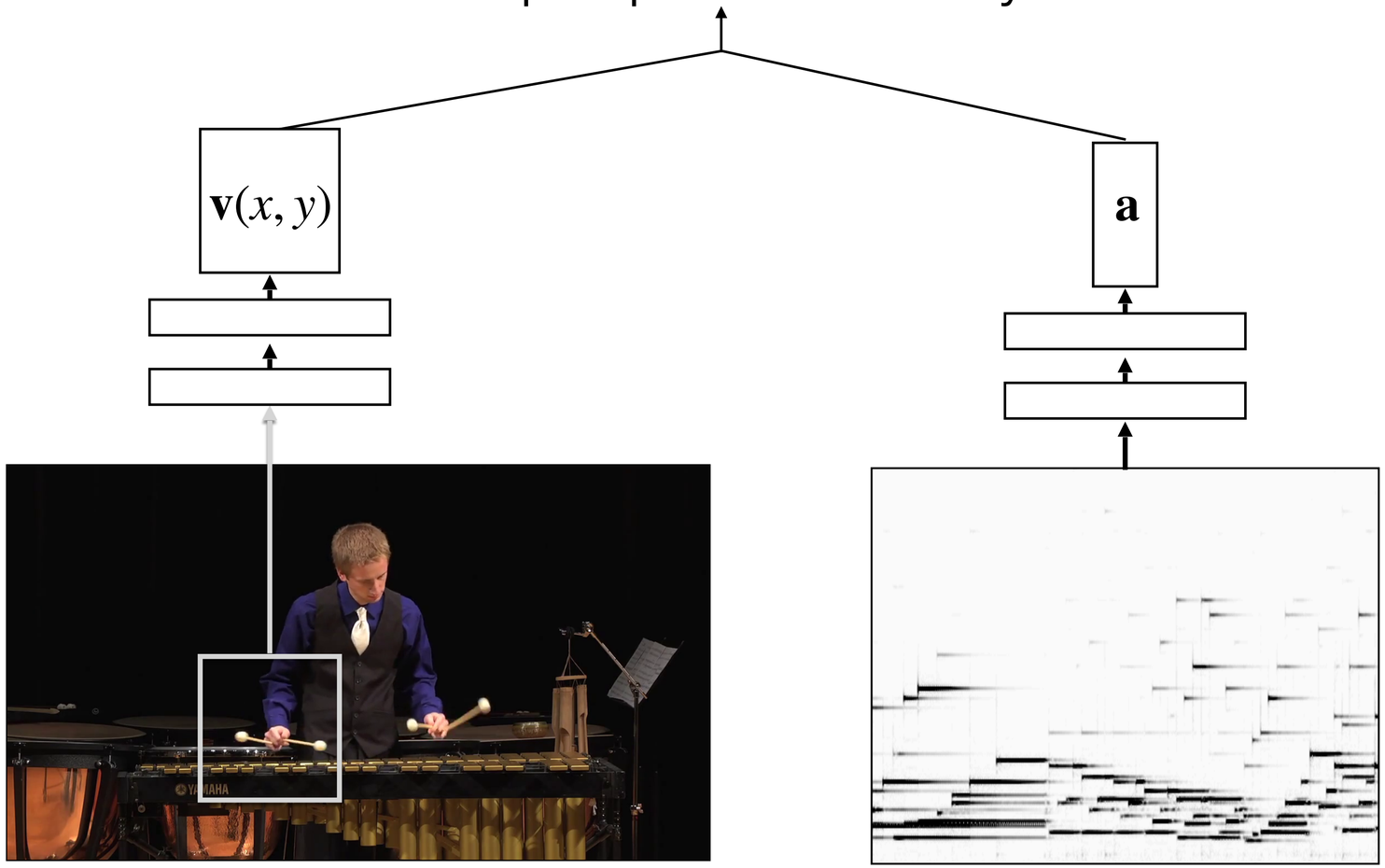
$\mathbf{v}_1^\top \mathbf{a}_1 \longrightarrow$ High dot product with self

$\mathbf{v}_1^\top \mathbf{a}_2 \longrightarrow$ Low dot product with others



Application: sound source localization

Combine per-patch similarity scores



During training, maximize:

$$\max_{x,y} \mathbf{v}(x, y)^\top \mathbf{a}$$

Underlying assumption: there is one sound source. You can ignore the rest of the image.

[R. Arandjelovic and A. Zisserman, 2018]

Application: sound source localization

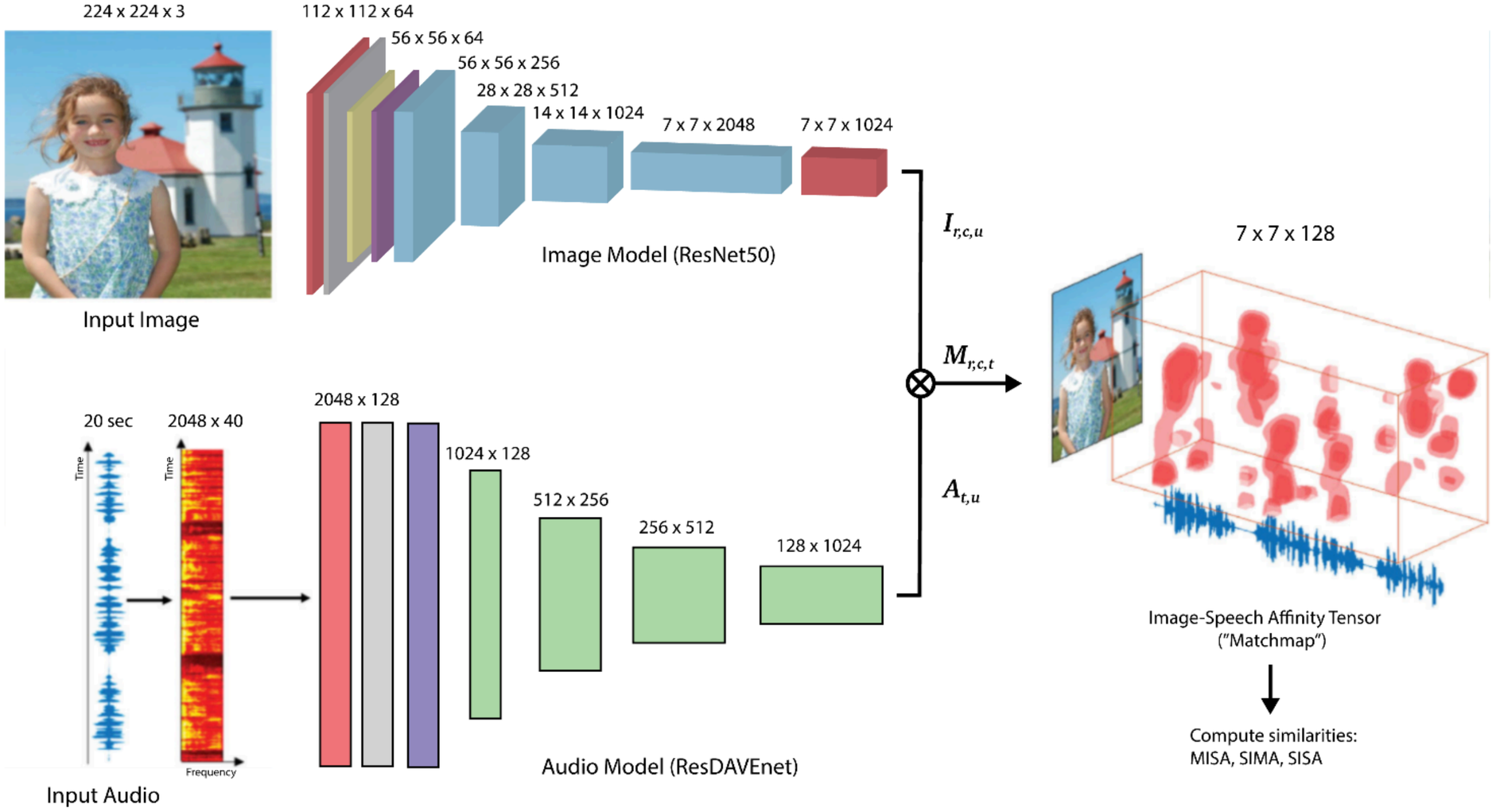
Objects that Sound

Relja Arandjelović[1] , Andrew Zisserman[1,2]
[1]DeepMind [2]University of Oxford

Frames are processed completely independently, motion information is not used, and there is no temporal smoothing

[R. Arandjelovic and A. Zisserman, 2018]

Application: associating image content with speech



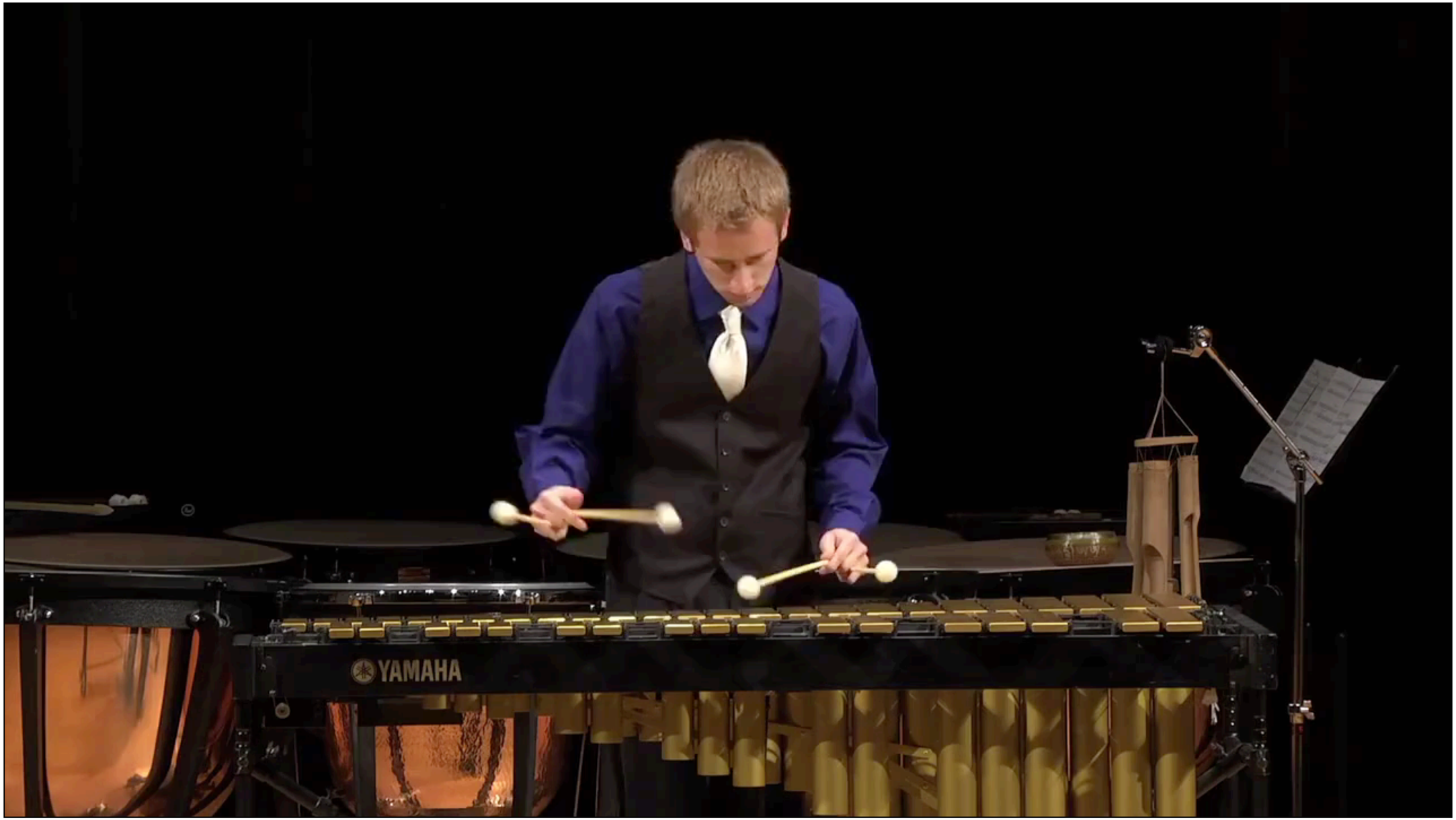
[Harwath et al. "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input", 2018]

Discovering words in raw audio-visual data

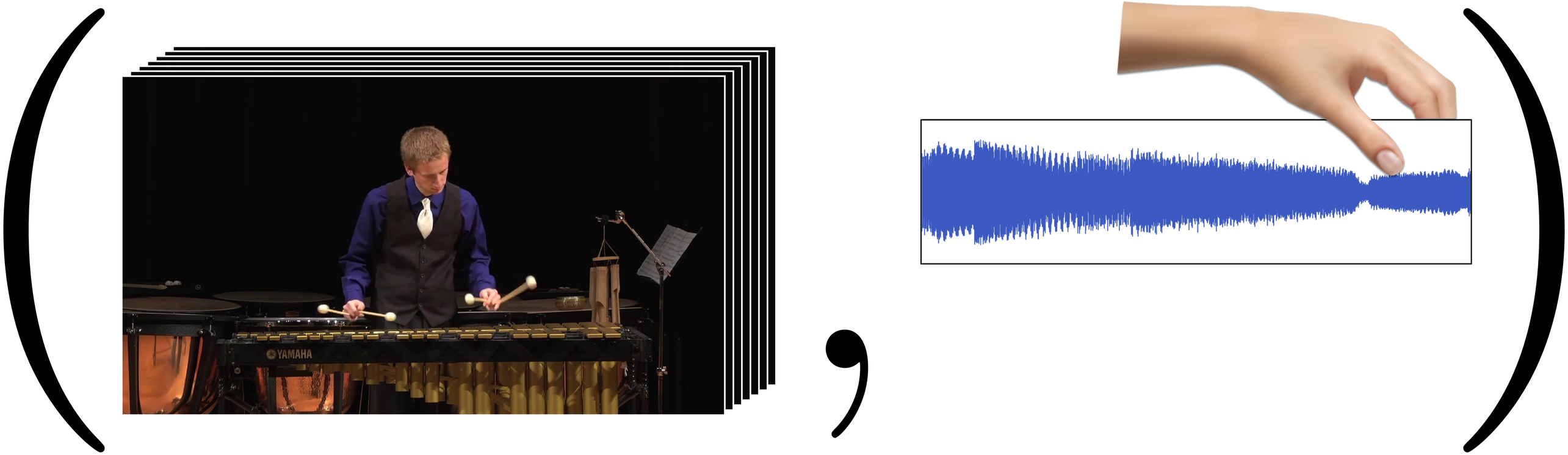
[Harwath et al. “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input”, 2018]

Video source

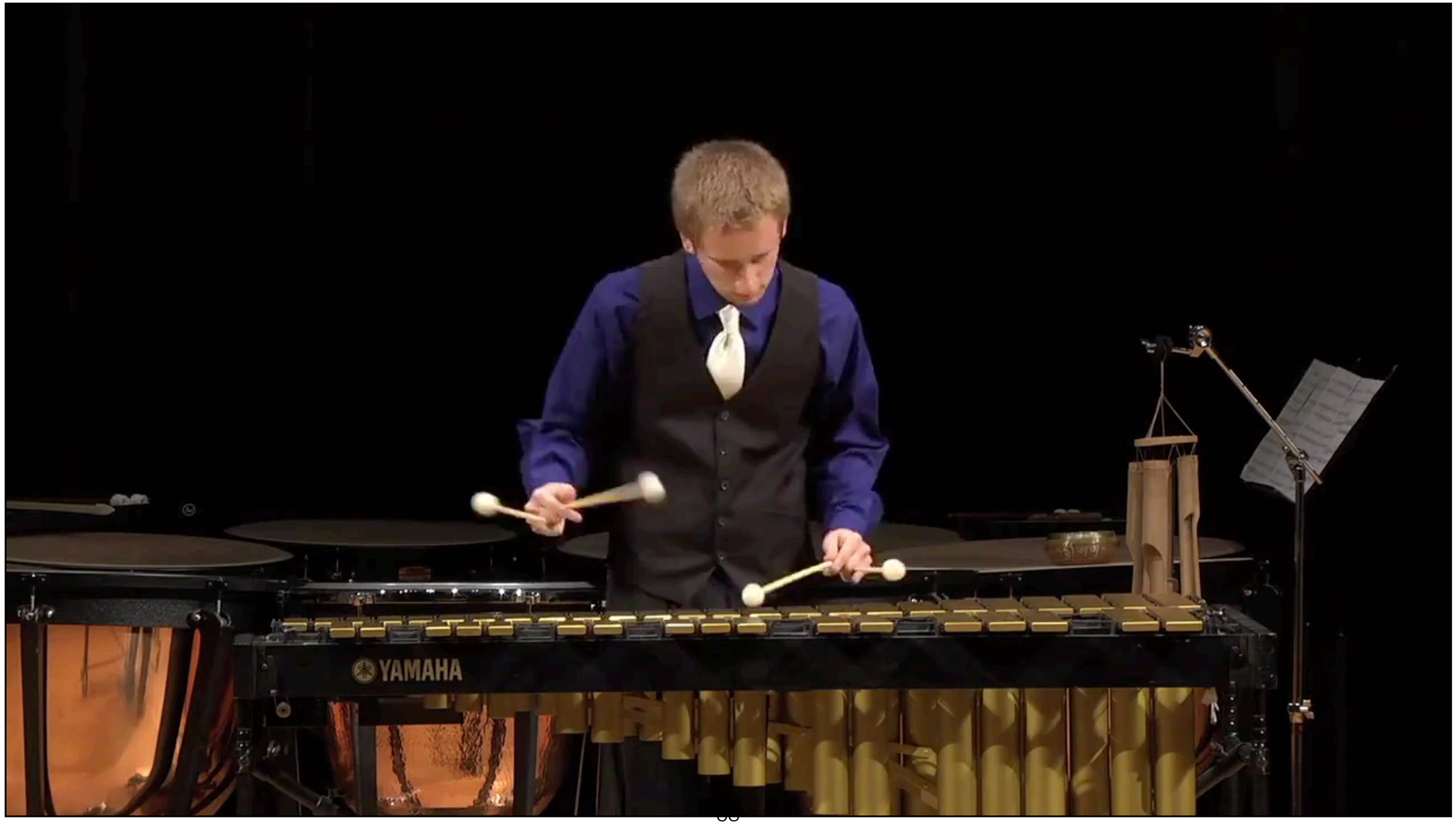
Is this task too easy?



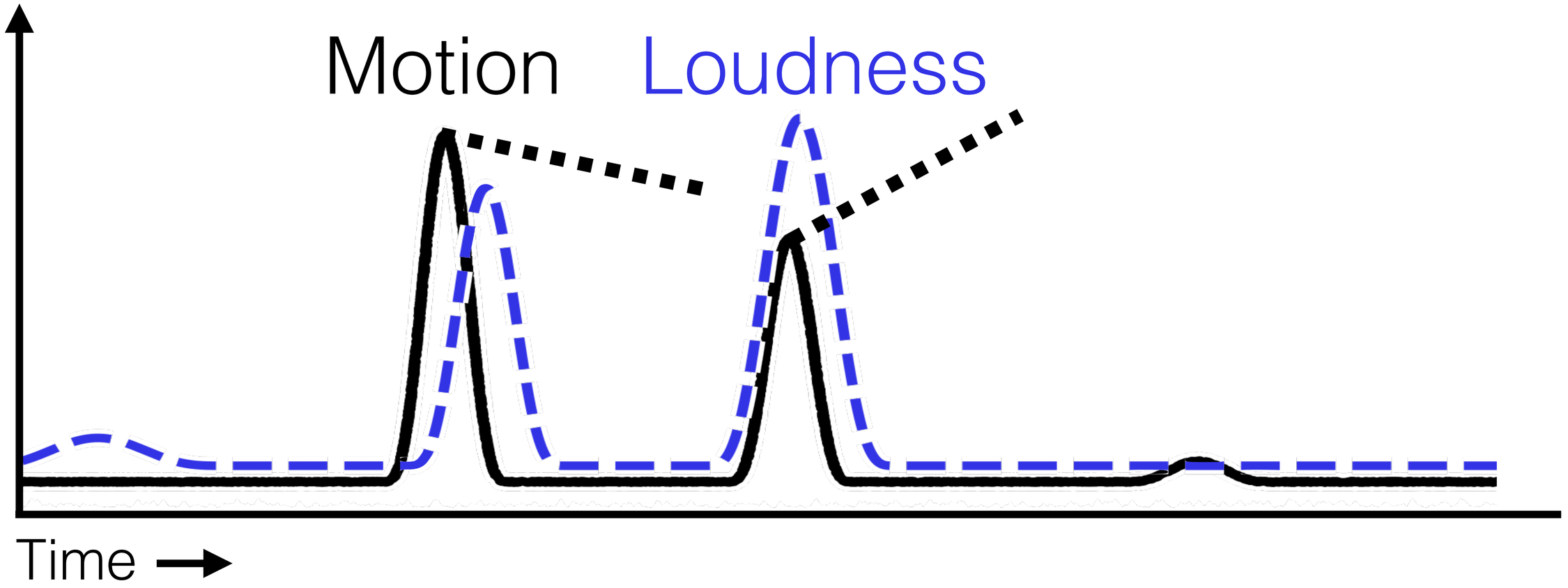
Idea #2: use time-shifted audio as negatives



Idea #2: use time-shifted audio as negatives



Idea #2: use time-shifted audio as negatives



Dribbling basketball

Dribbling basketball

Dribbling basketball

Playing organ

Playing organ

Playing organ

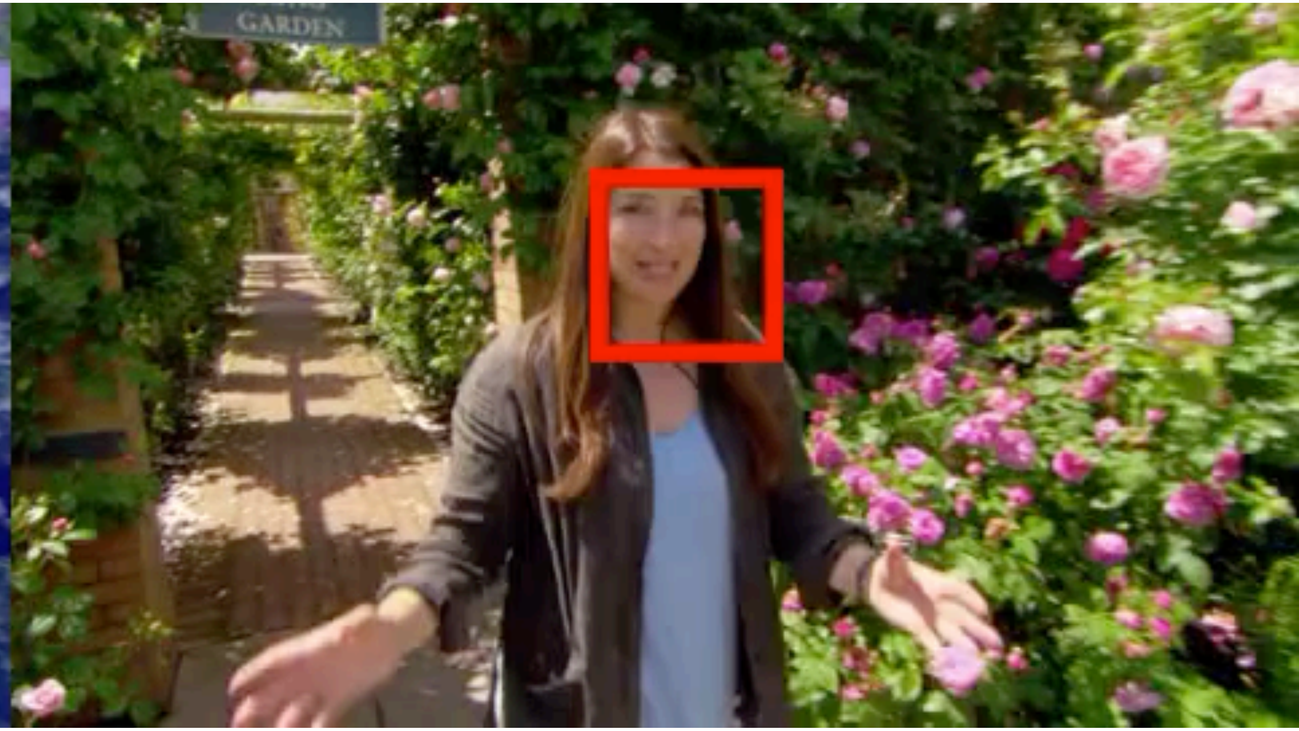
Chopping wood

Chopping wood

Chopping wood

Detecting sound sources

Attention map

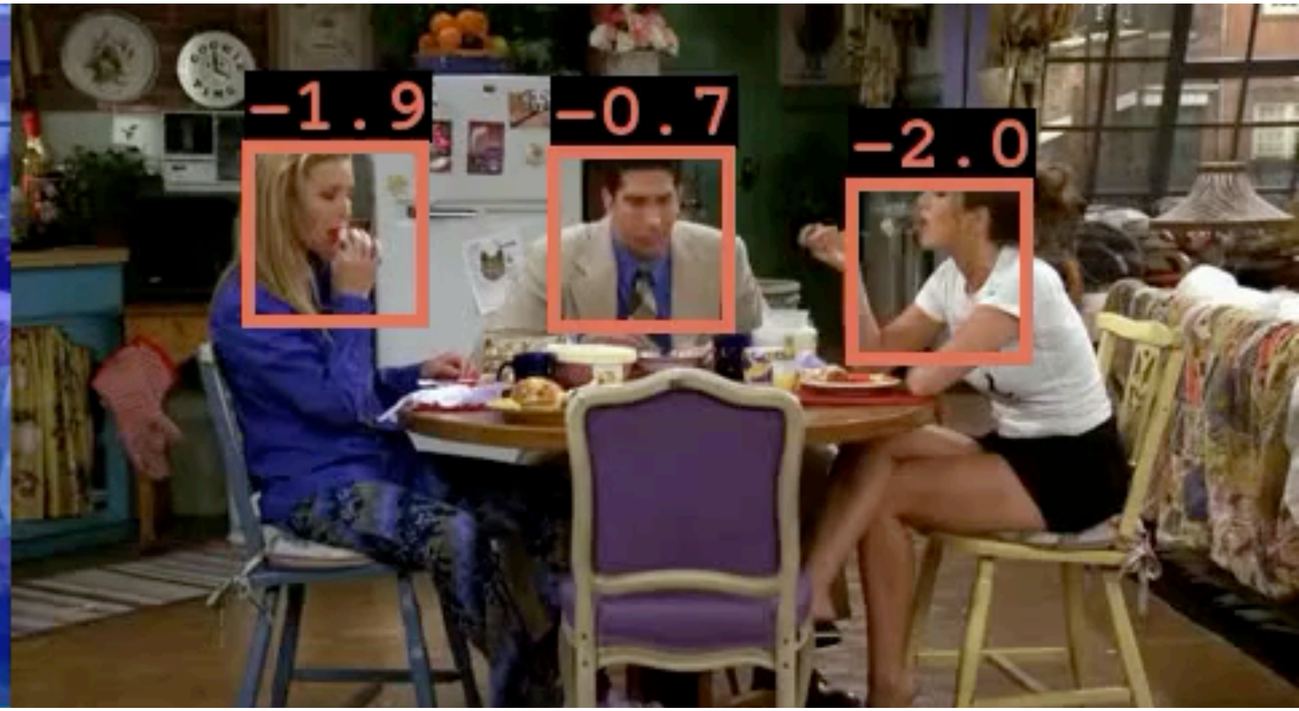


Audio-visual object

A. Afouras, A. Owens, J. S. Chung, A. Zisserman, ECCV 2020

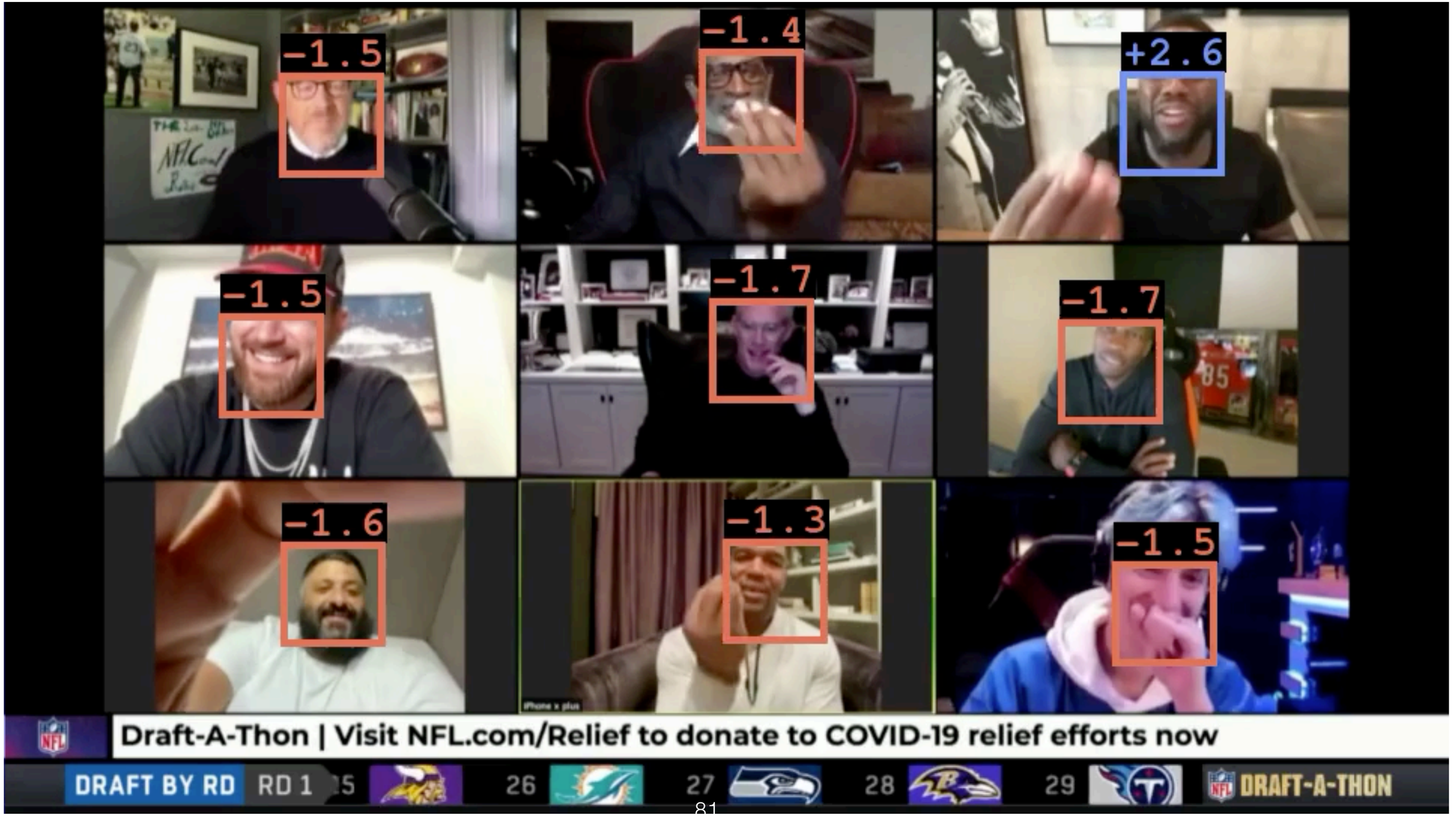
Active Speaker Detection

Examples from the *Friends* series



Blue = active speaker
Red = inactive speaker

Active Speaker Detection



Other applications of audio-visual models

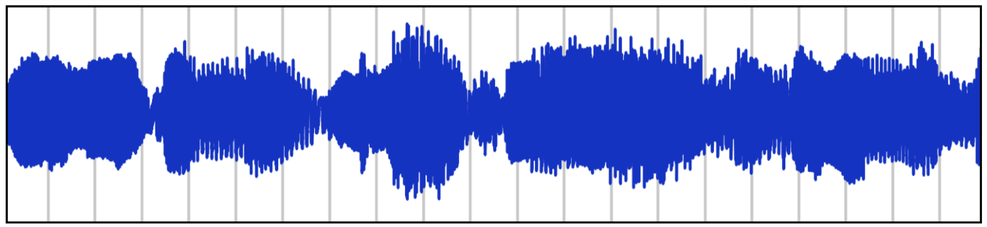
Application: on/off-screen source separation

Cocktail party problem

Underdetermined!

a	=	b	+	c
Sound mixture		On-screen		Off-screen

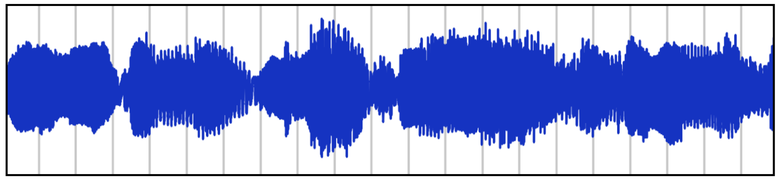
Learning from synthetic sound mixtures



Synthetic mixture

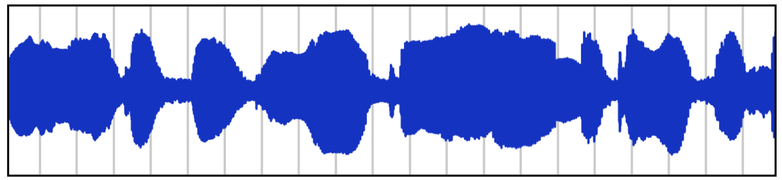
Offscreen

Learning from synthetic sound mixtures

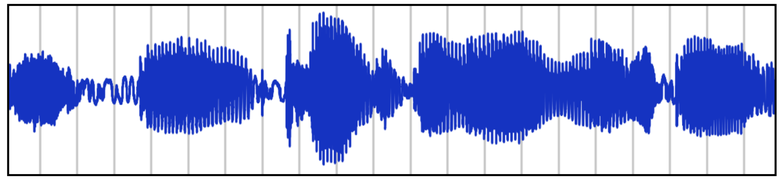


Synthetic mixture

Regression



On-screen



Off-screen

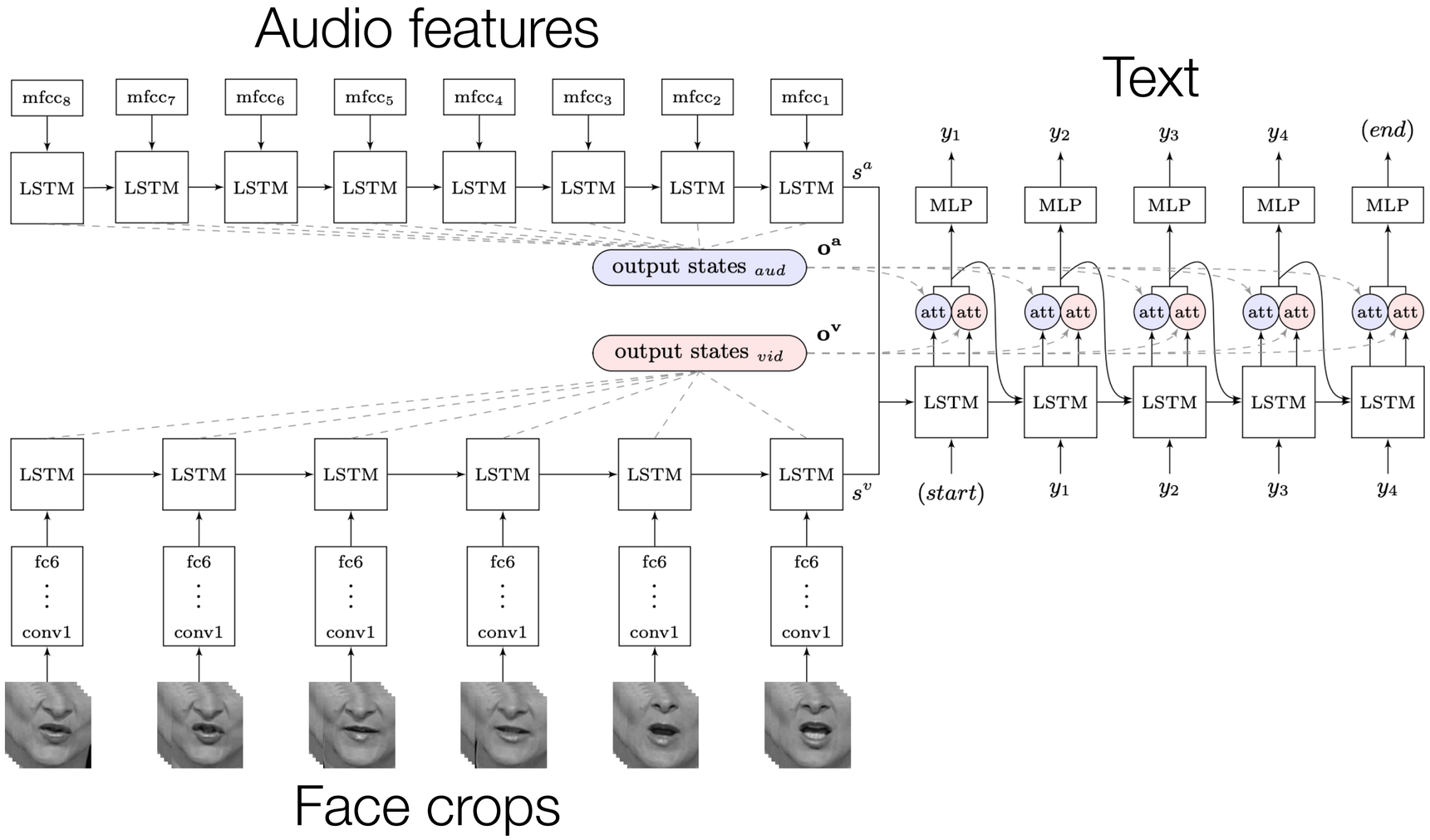
Related work: audio-only separation (Hershey 2016)

Looking to Listen at the Cocktail Party:
A Speaker-Independent Audio-Visual Model for Speech Separation

Ariel Ephrat Inbar Mosseri Oran Lang Tali Dekel Kevin Wilson
Avinatan Hassidim William T. Freeman Michael Rubinstein

Sports debate

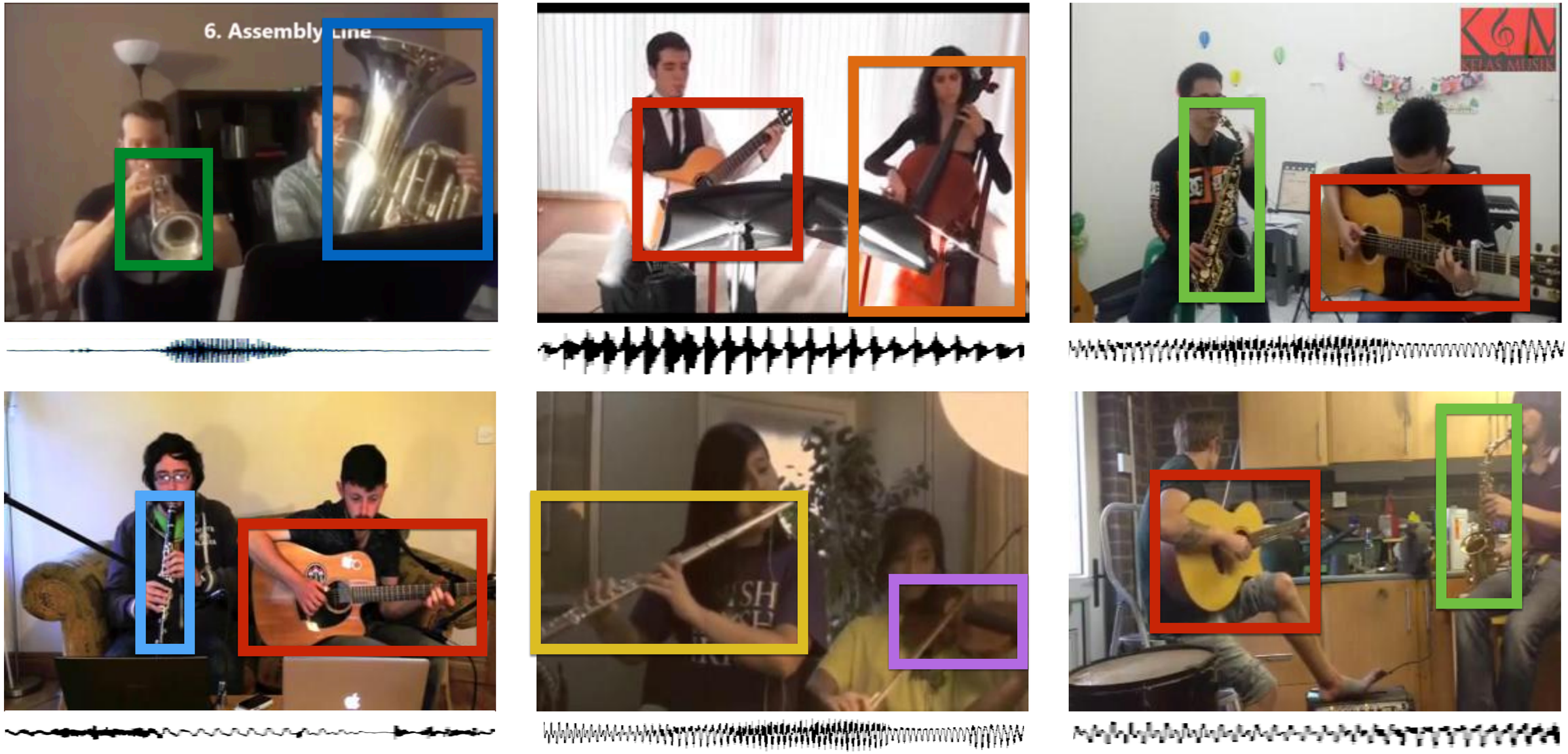
Lip reading



[Chung et al. "Lip reading sentences in the wild", 2017]

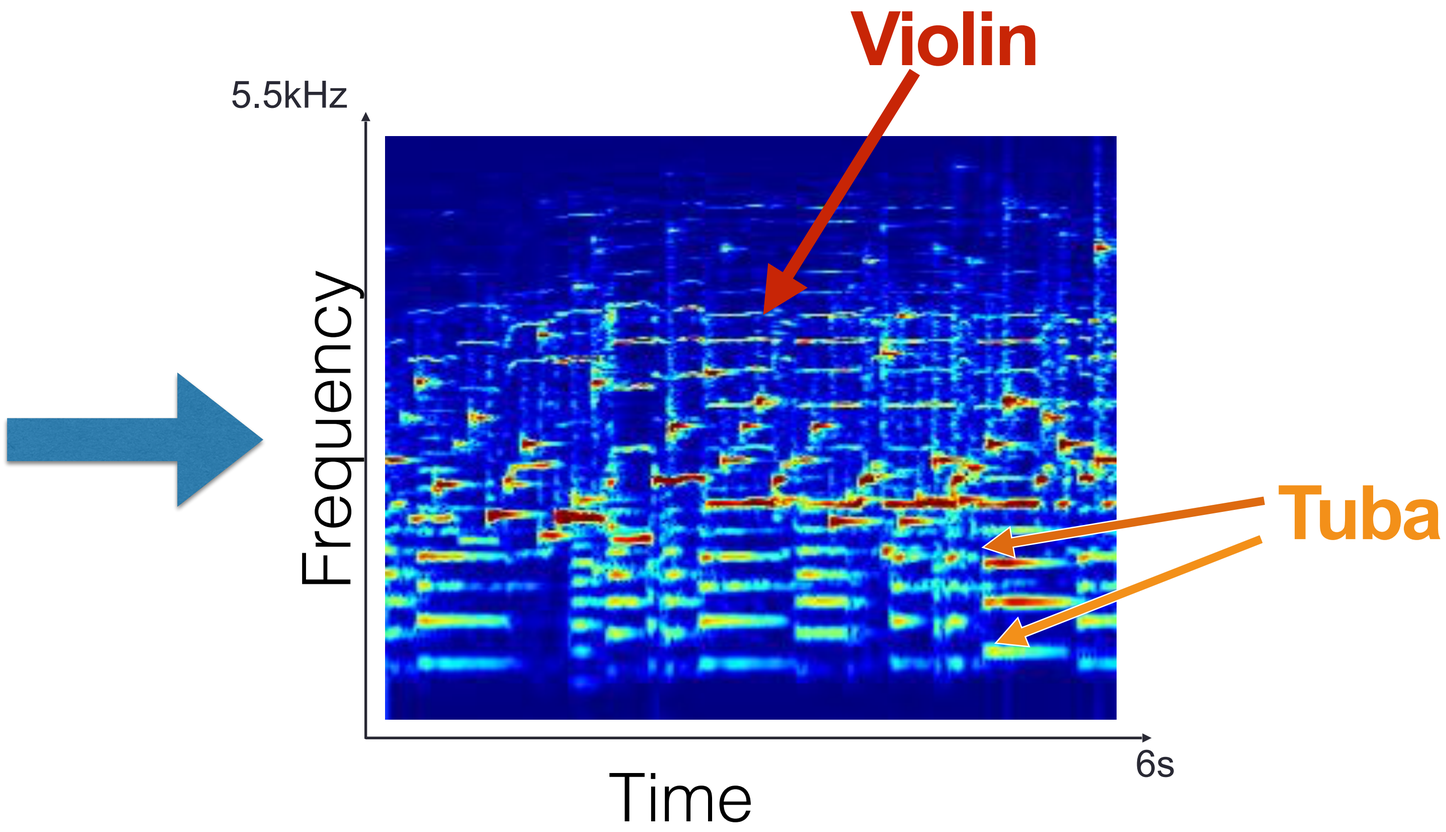
The video-to-text alignment is also generated by the model.

Discovering musical instruments

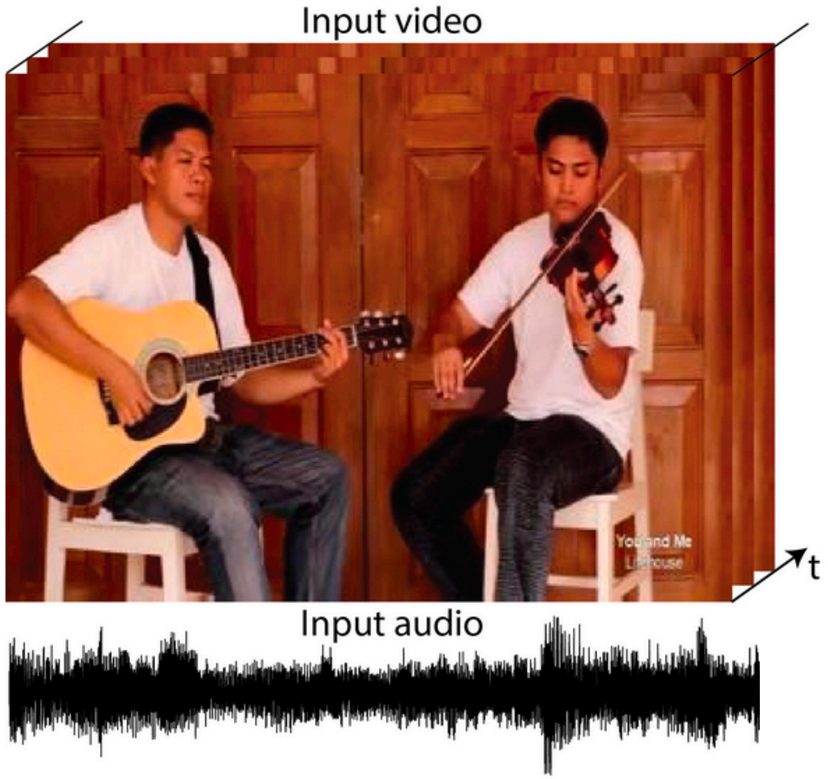


[Zhao et al., “The Sound of Pixels”, 2018]

Musical instrument separation

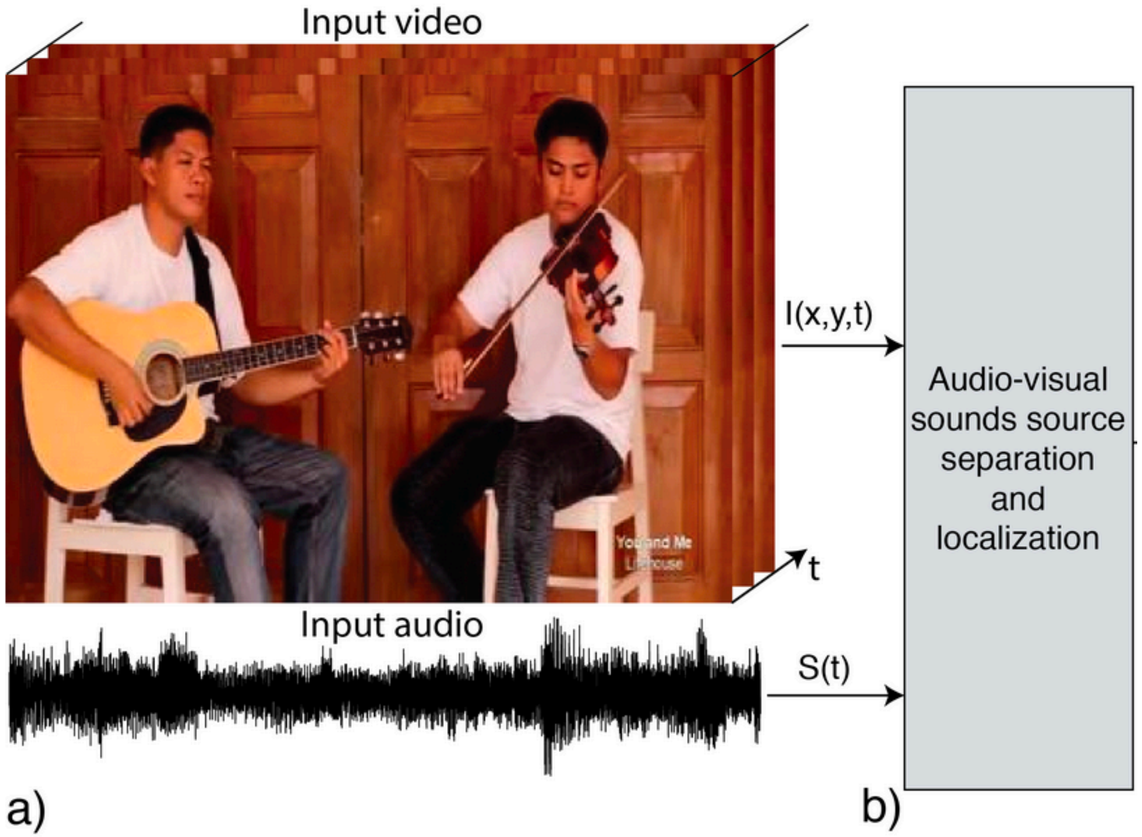


[Zhao et al., “The Sound of Pixels”, 2018]

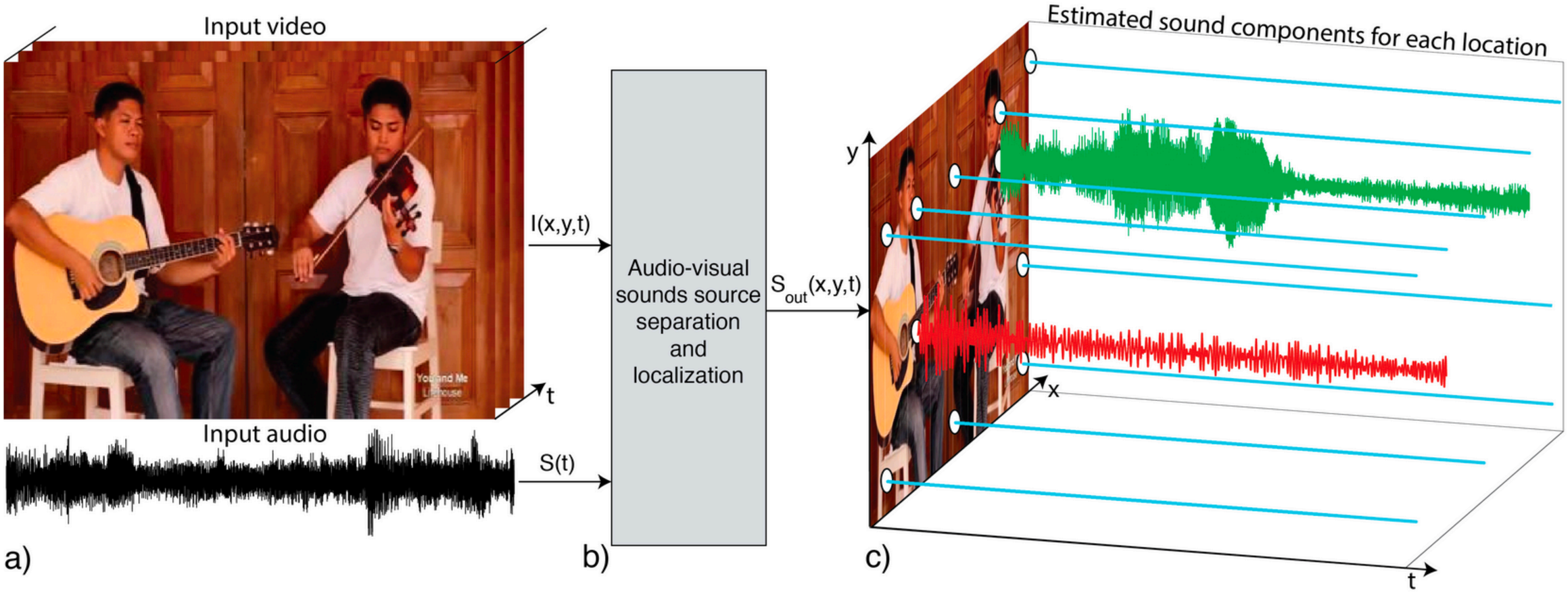


http://sound-of-pixels.csail.mit.edu/

[Zhao et al., “The Sound of Pixels”, 2018]



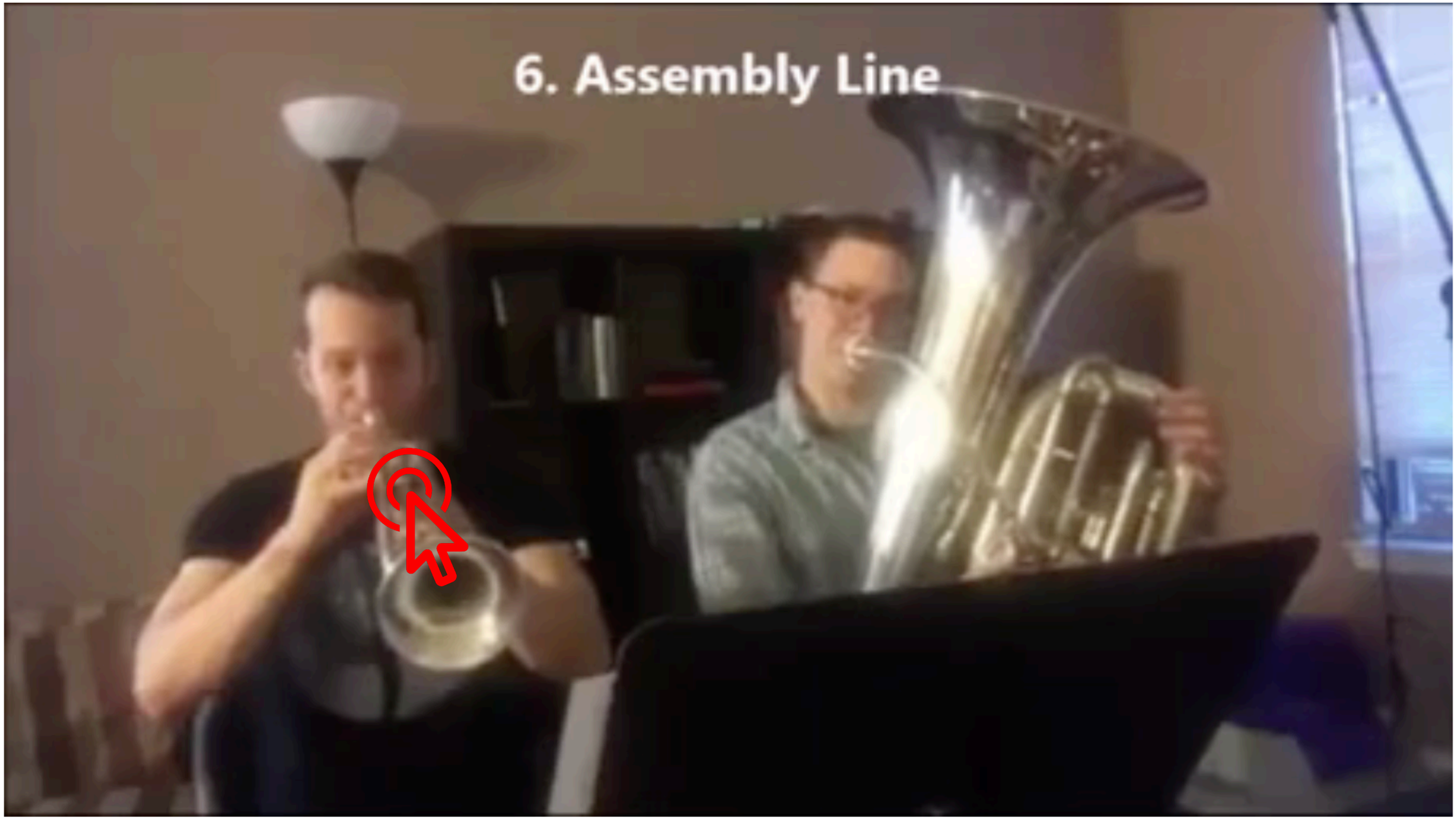
[Zhao et al., “The Sound of Pixels”, 2018]



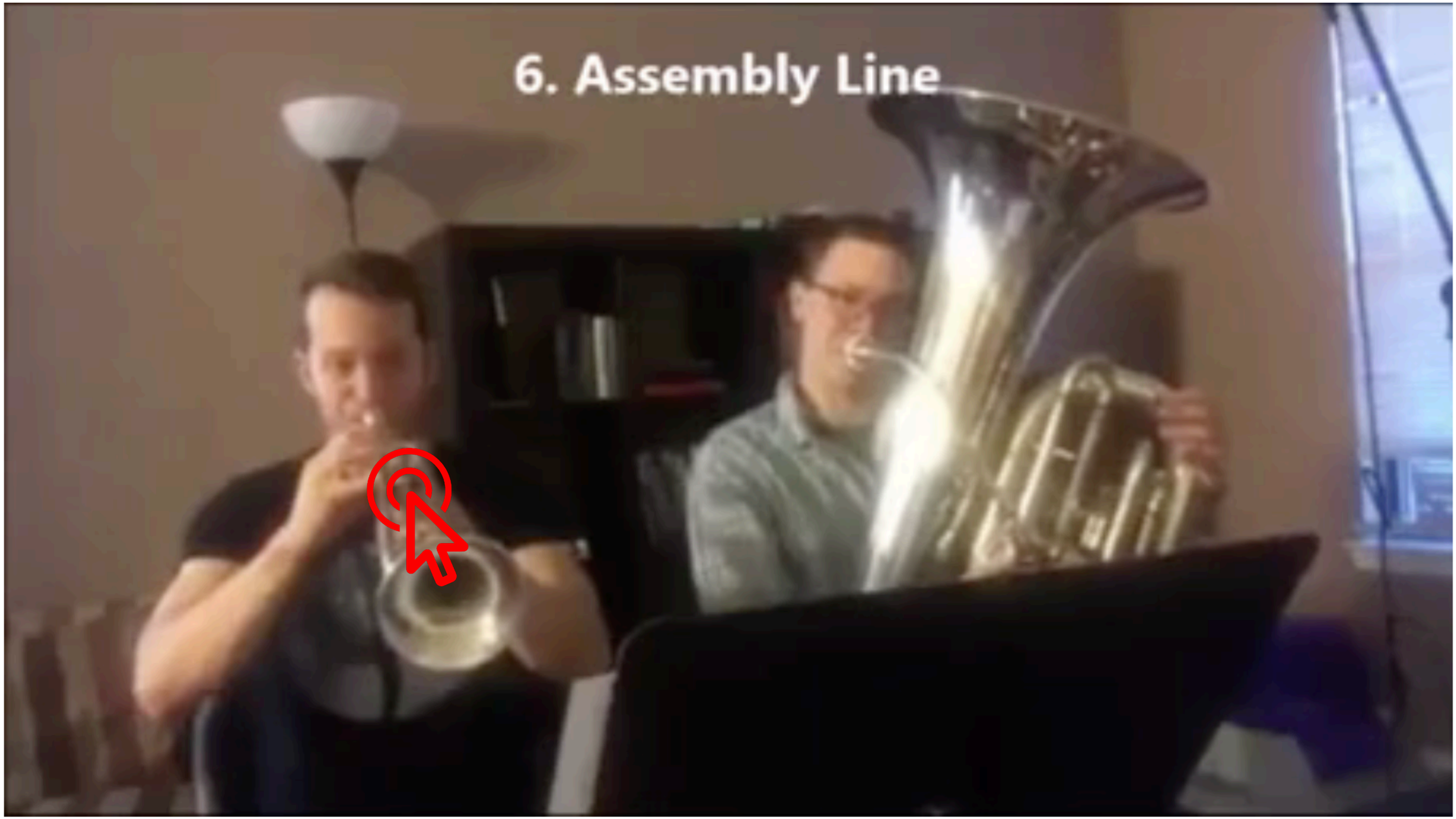
[Zhao et al., “The Sound of Pixels”, 2018]

http://sound-of-pixels.csail.mit.edu/

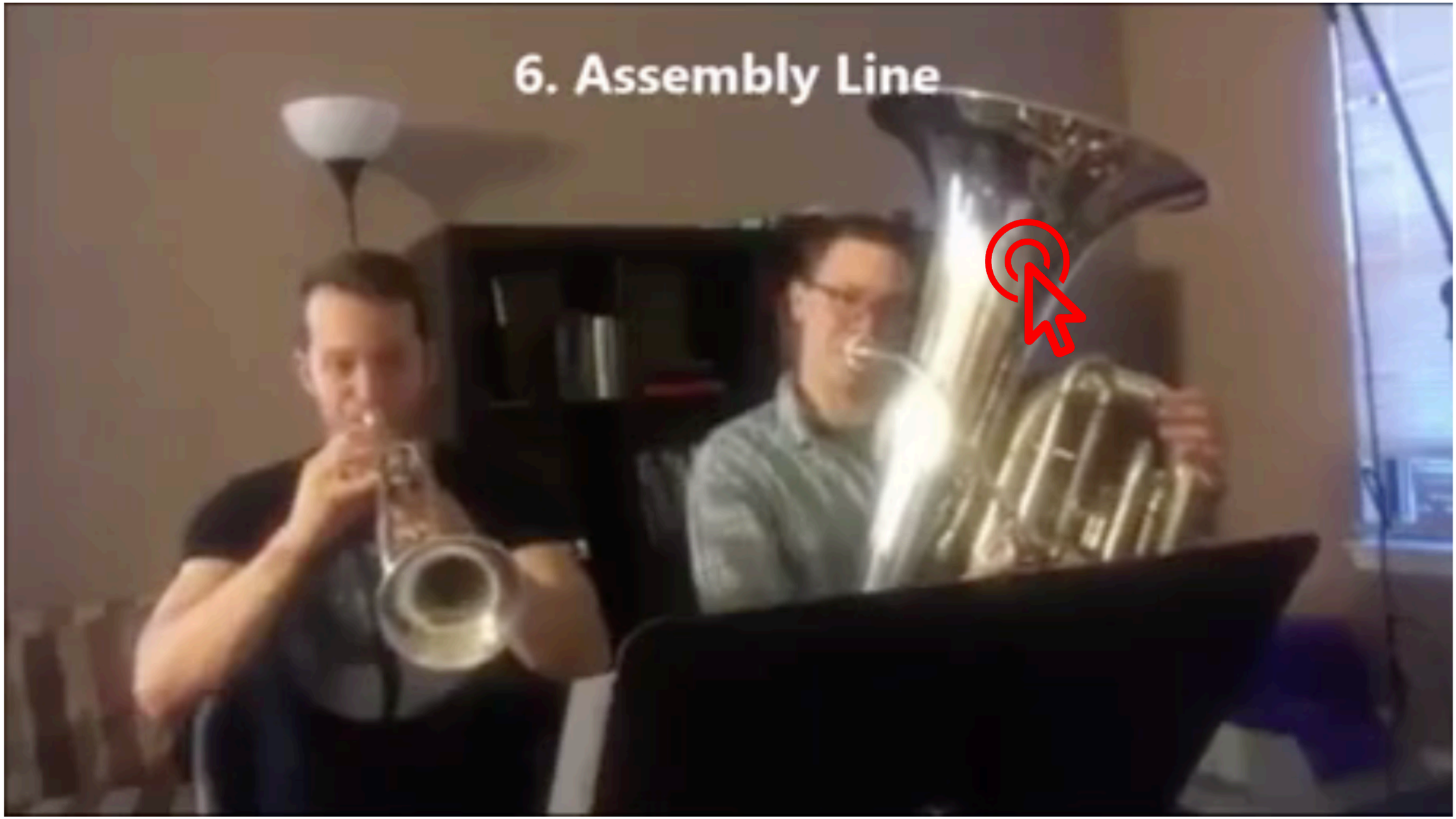
http://sound-of-pixels.csail.mit.edu/



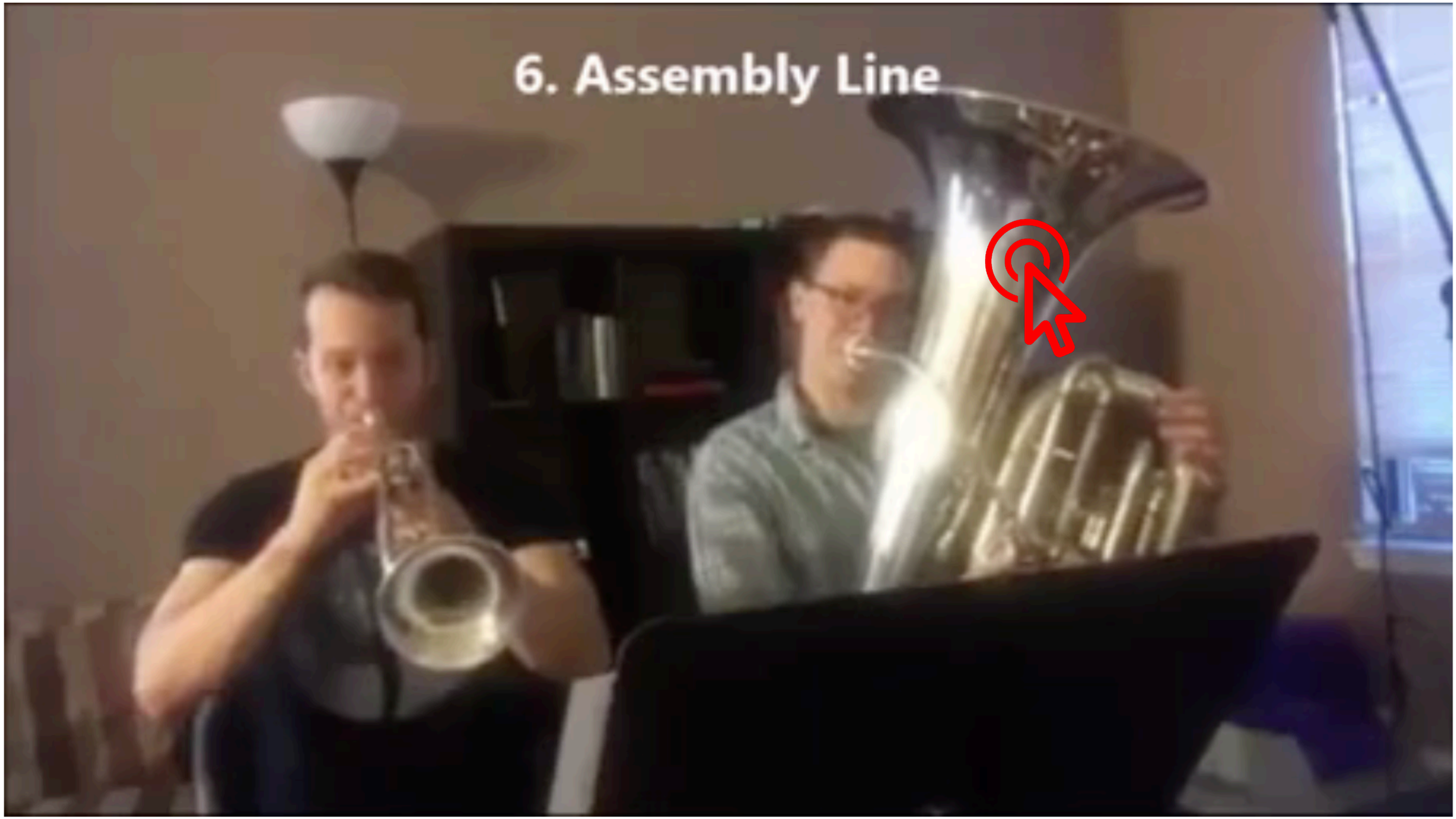
http://sound-of-pixels.csail.mit.edu/



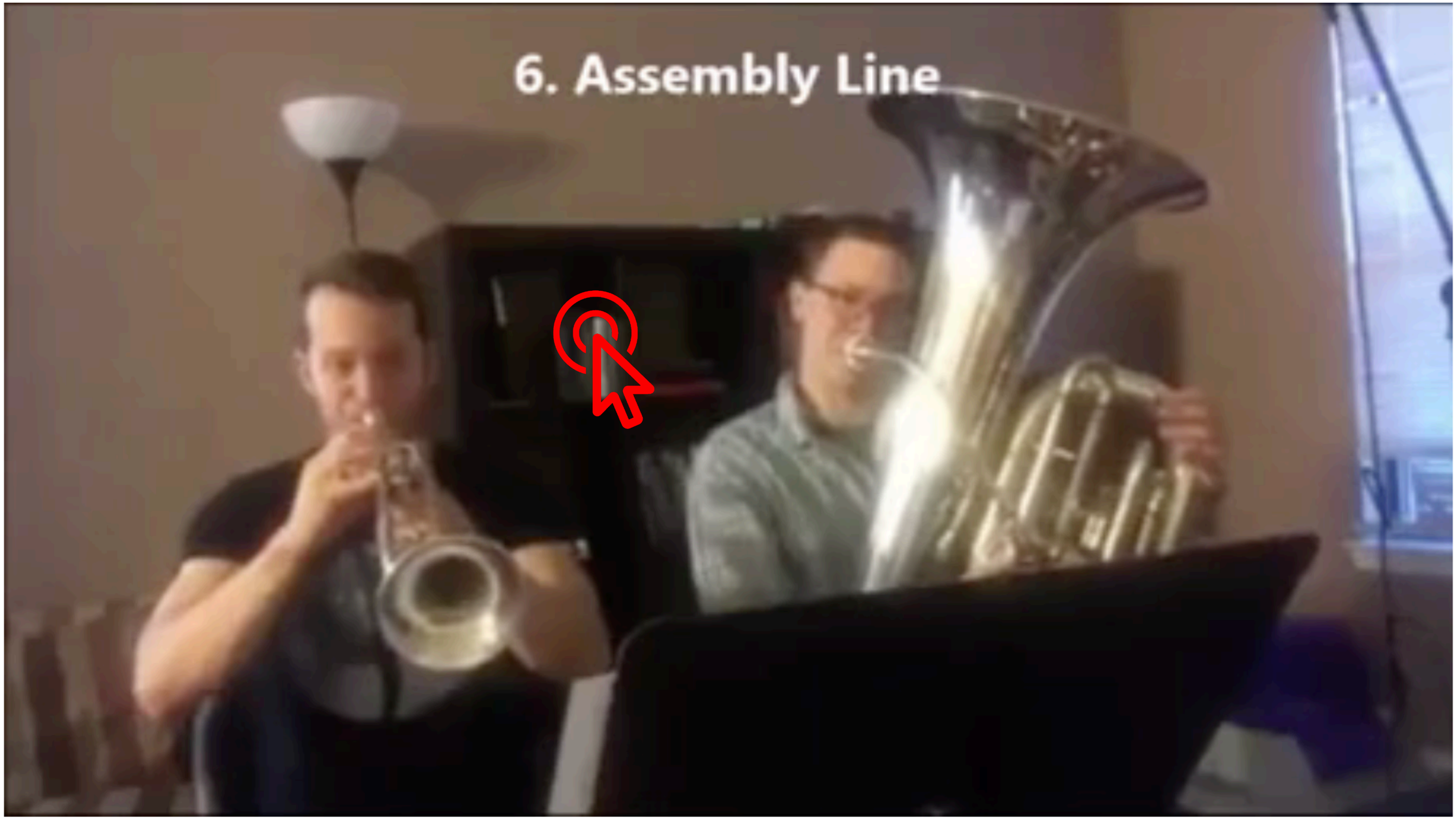
http://sound-of-pixels.csail.mit.edu/



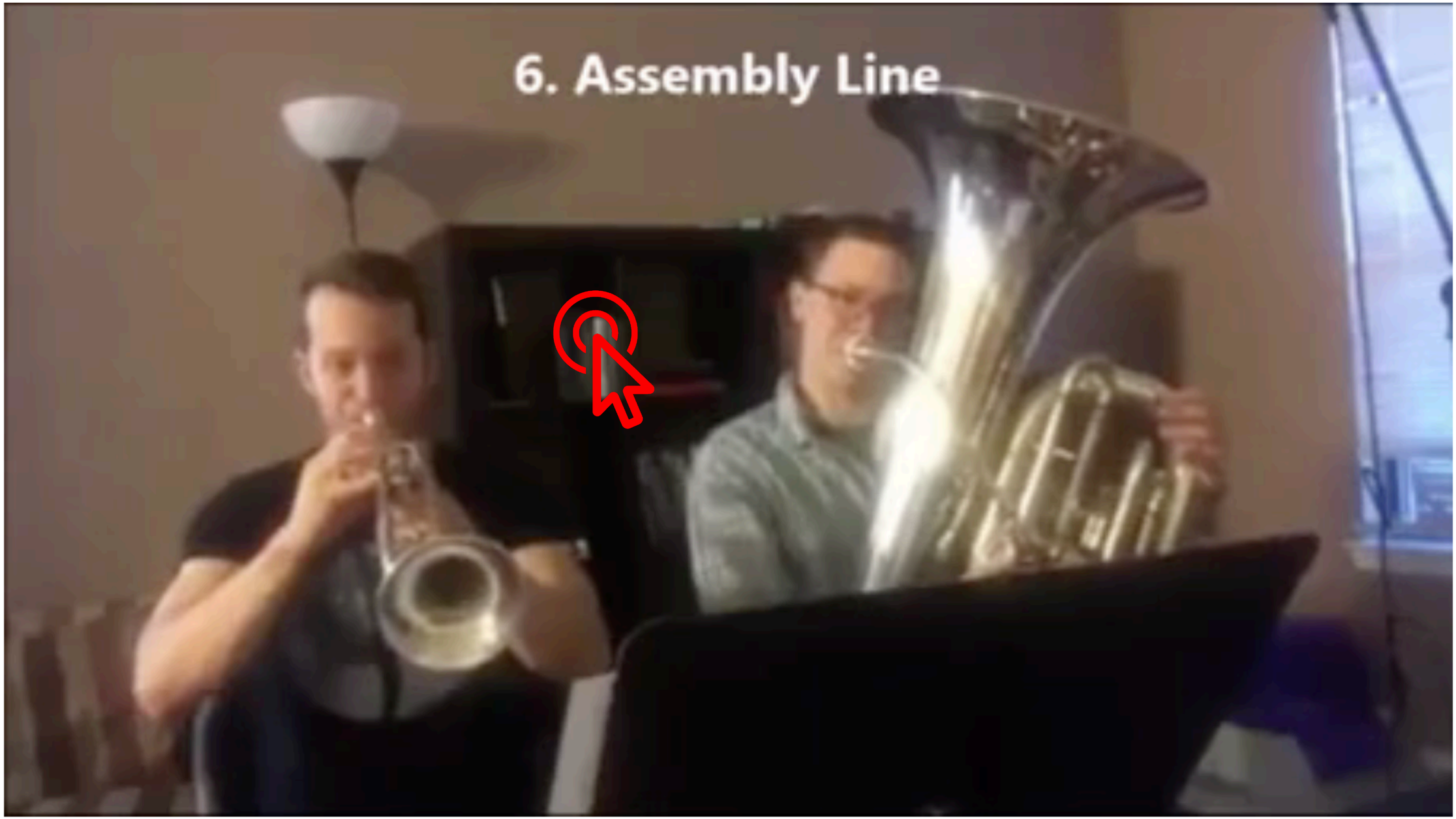
http://sound-of-pixels.csail.mit.edu/



http://sound-of-pixels.csail.mit.edu/



http://sound-of-pixels.csail.mit.edu/



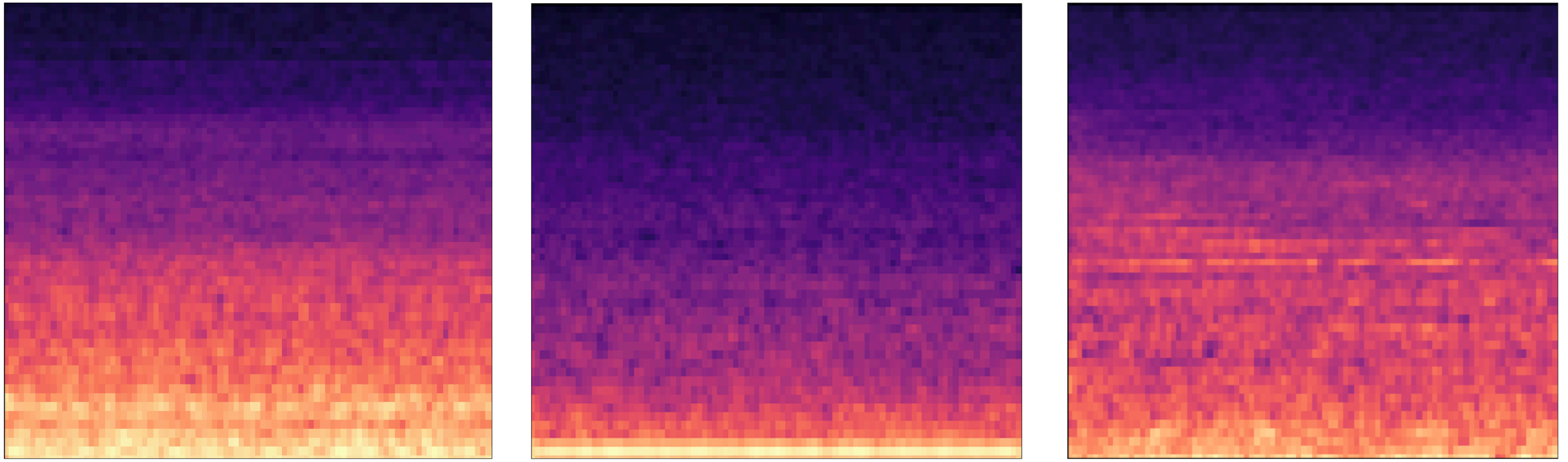
What does sound tell you about scenes?

Sound conveys scene structure

Kitchen

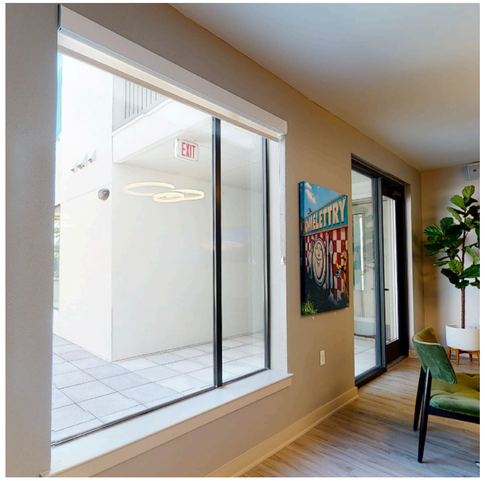
Vent

Window



[Hu, Chen, Owens, "Structure from Silence", 2021]

Ambient sounds convey 3D scene structure



Context

[Traer et al. 2016]

Reverberation

[Ashmead et al. 1998]

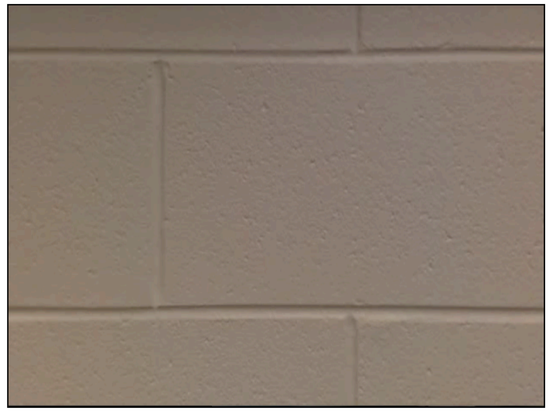
Interference

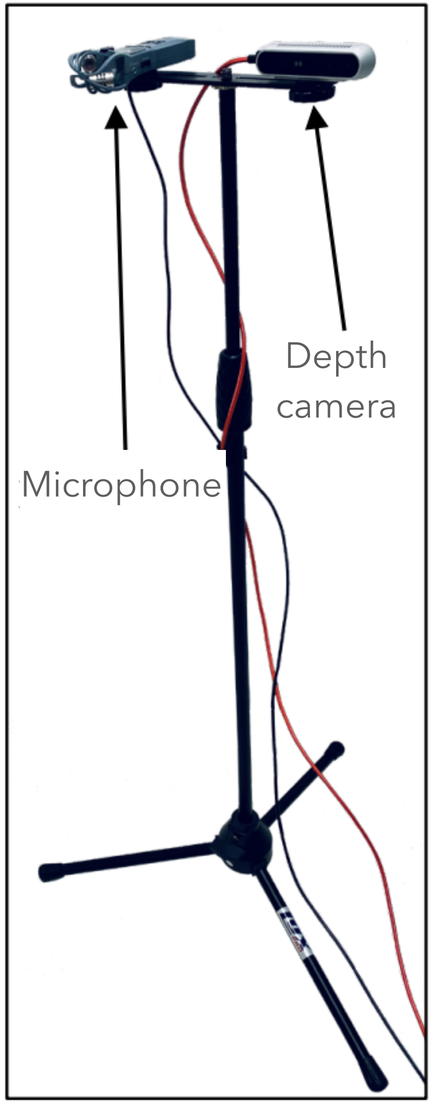
[Rosenblum et al. 2007]

Sound shadow

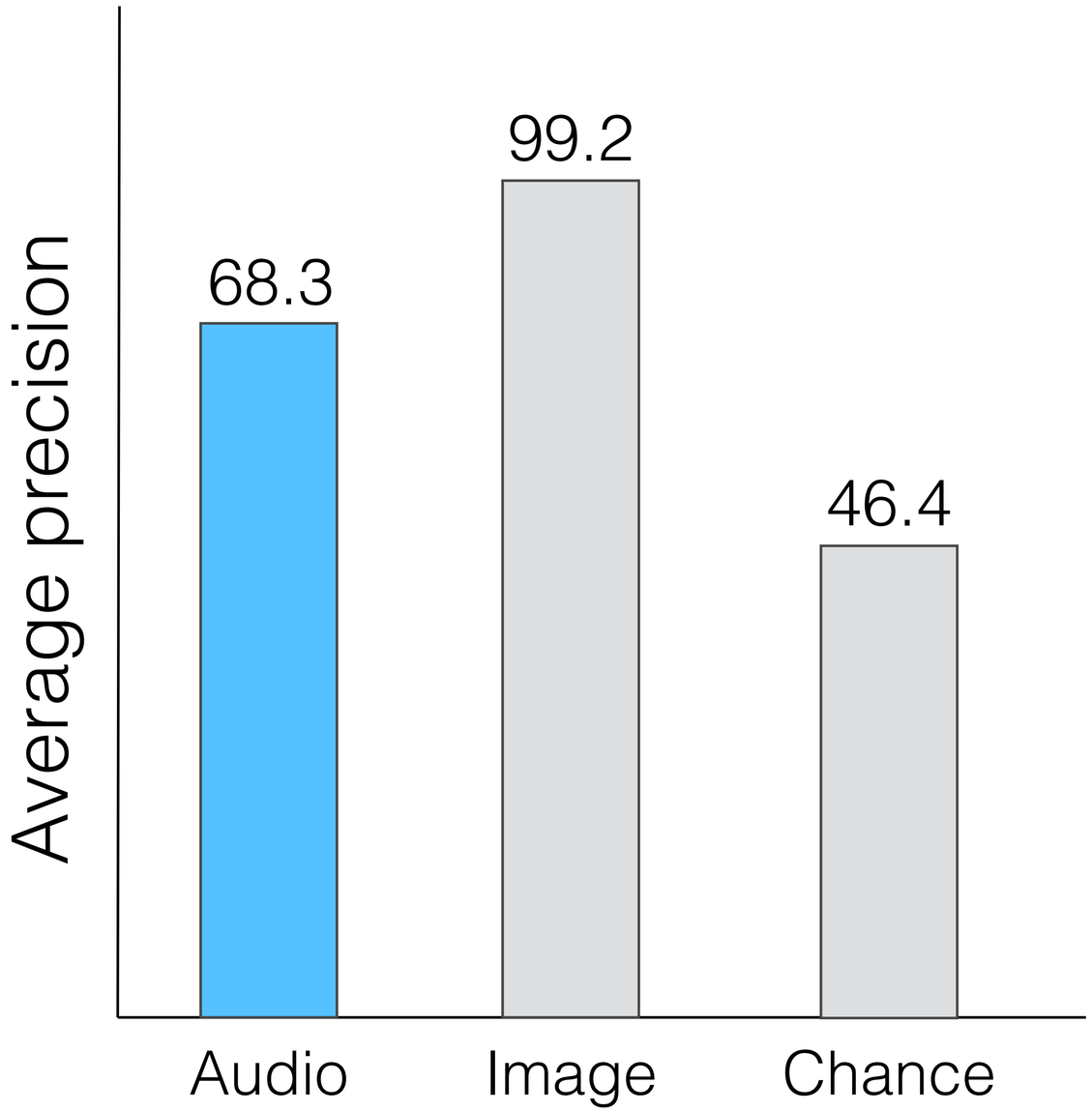
The *Quiet Campus* Dataset

- Sound + RGB-D
- 46 rooms, 12 buildings
- 14 hours

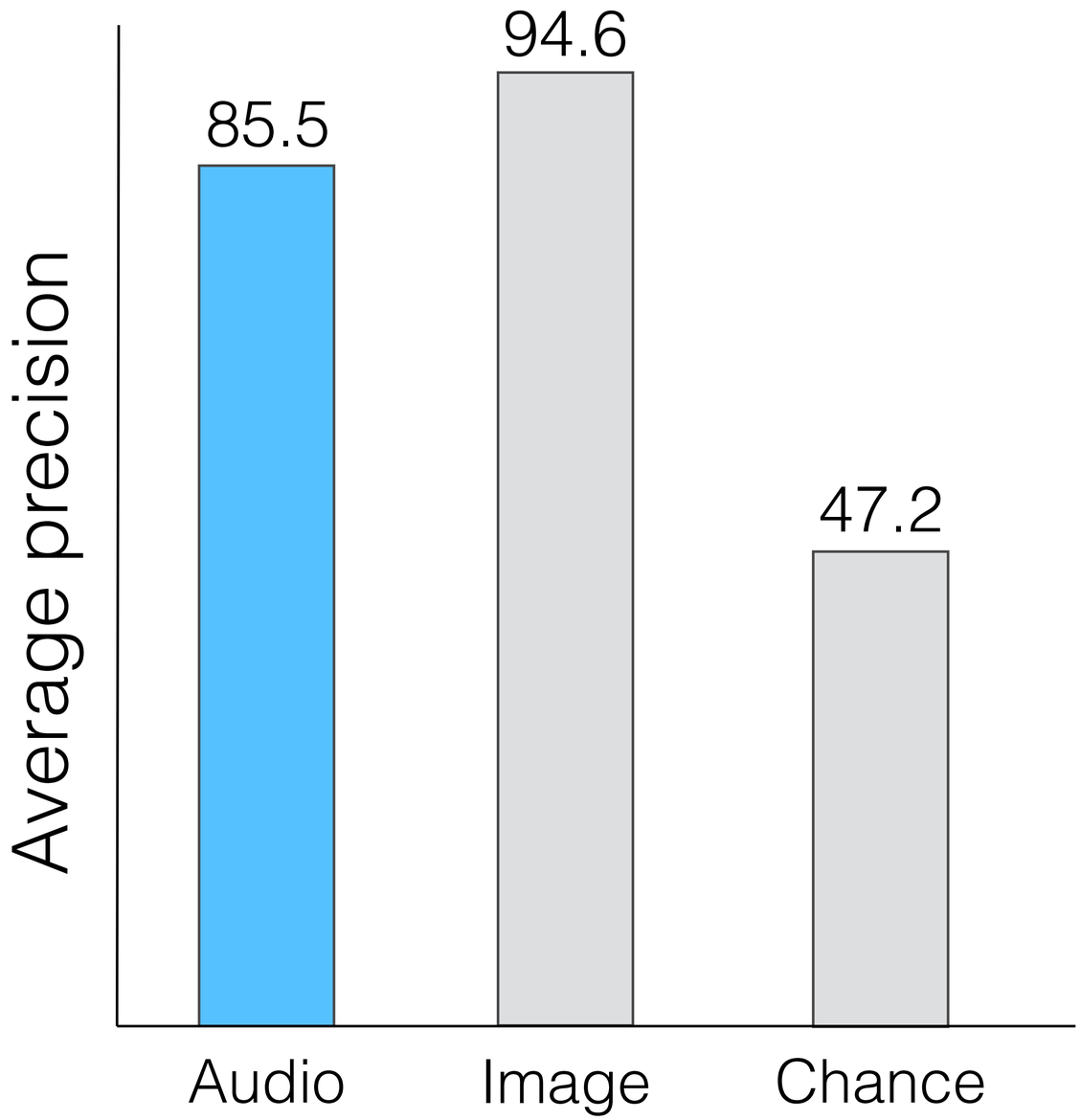




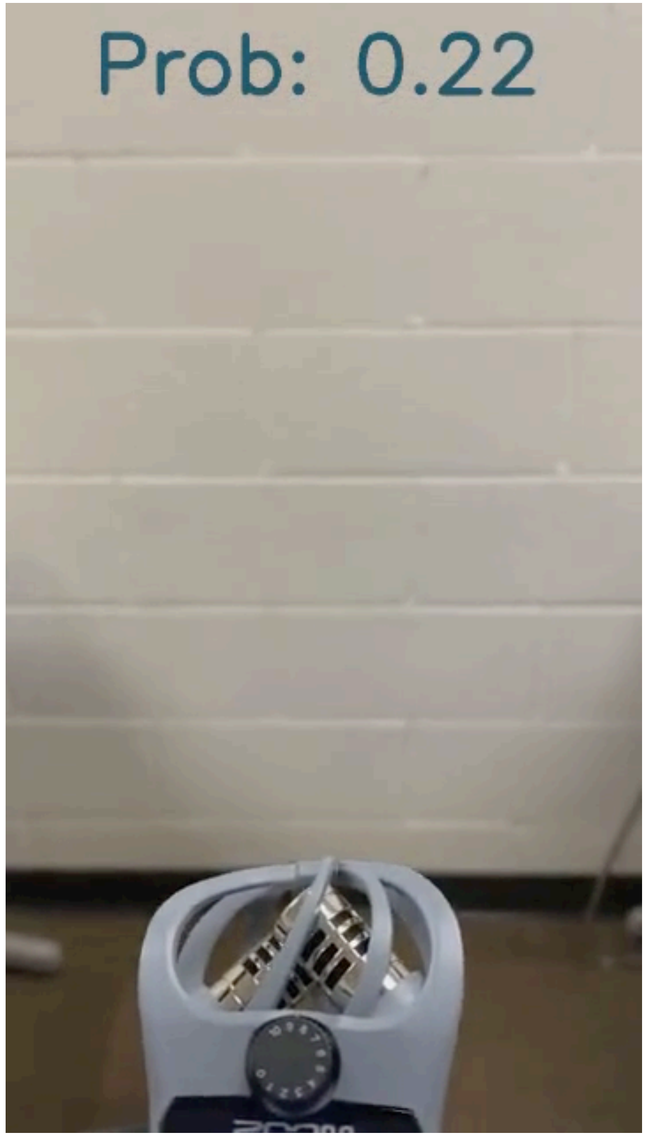
Depth estimation



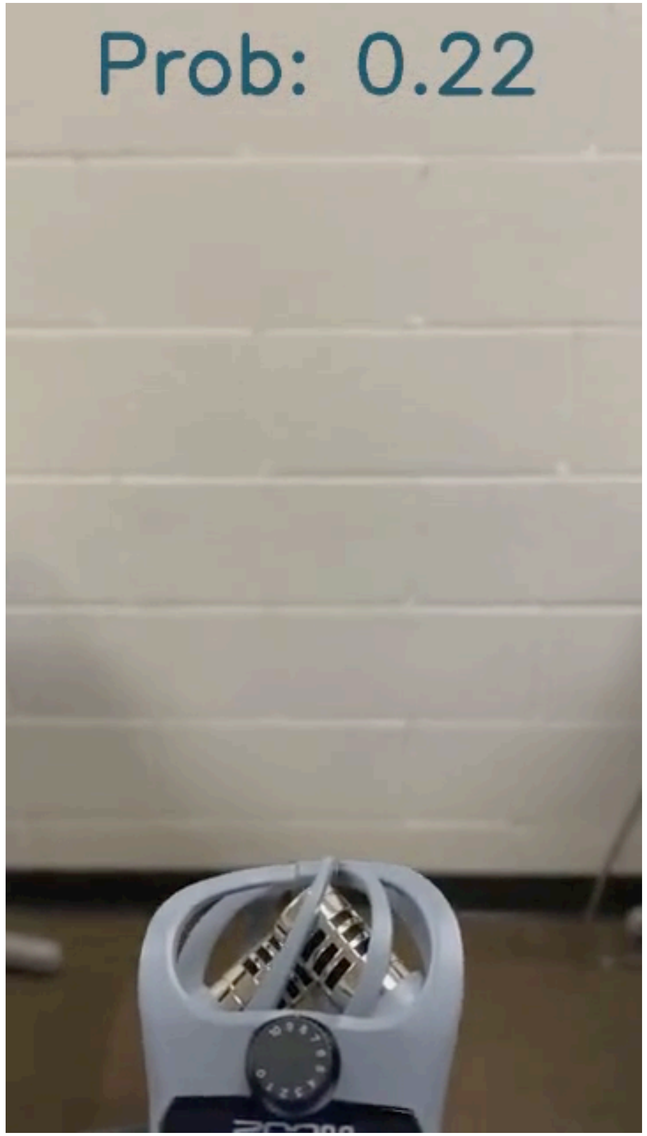
Relative depth estimation

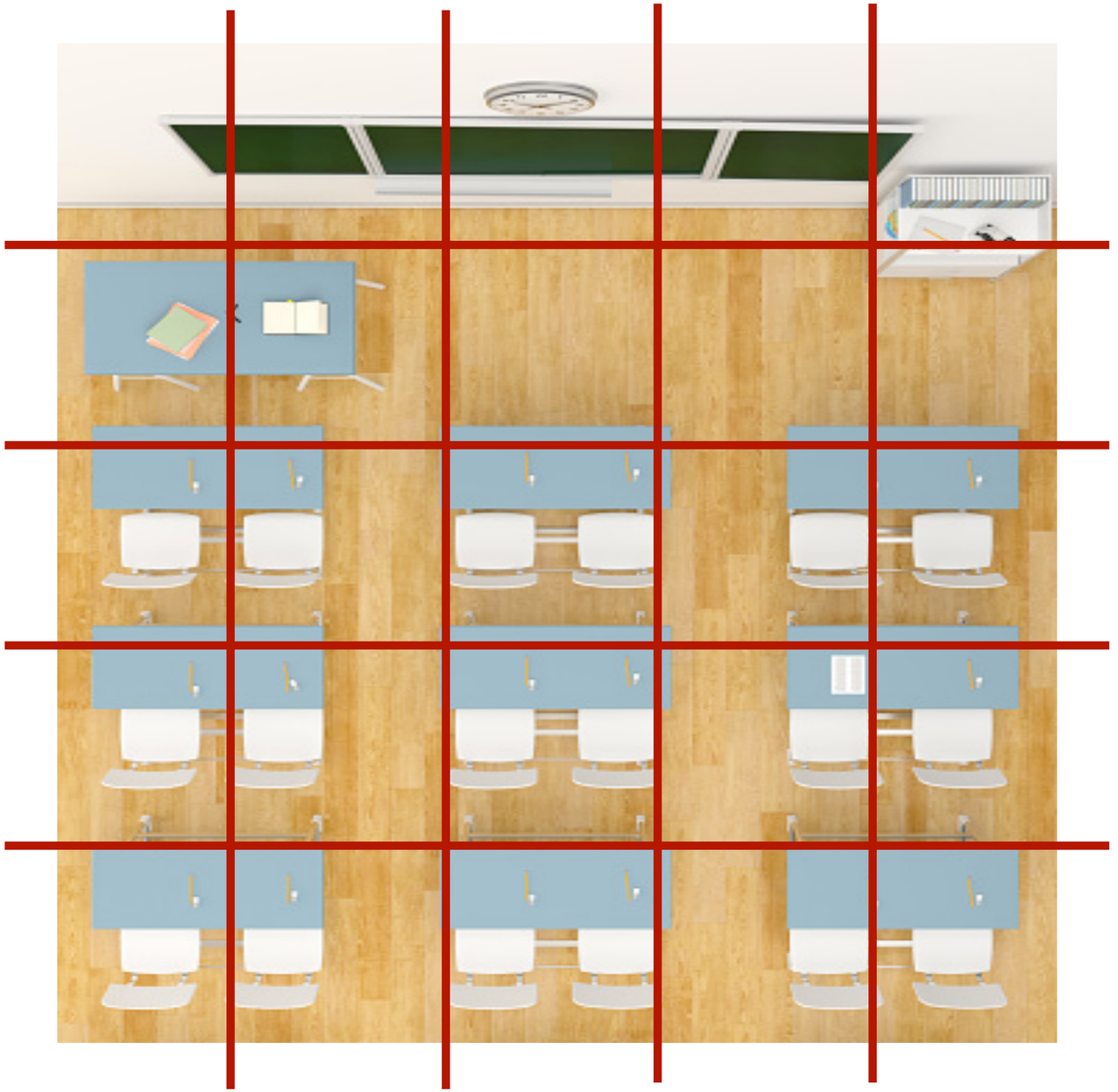


“Near a wall” prediction

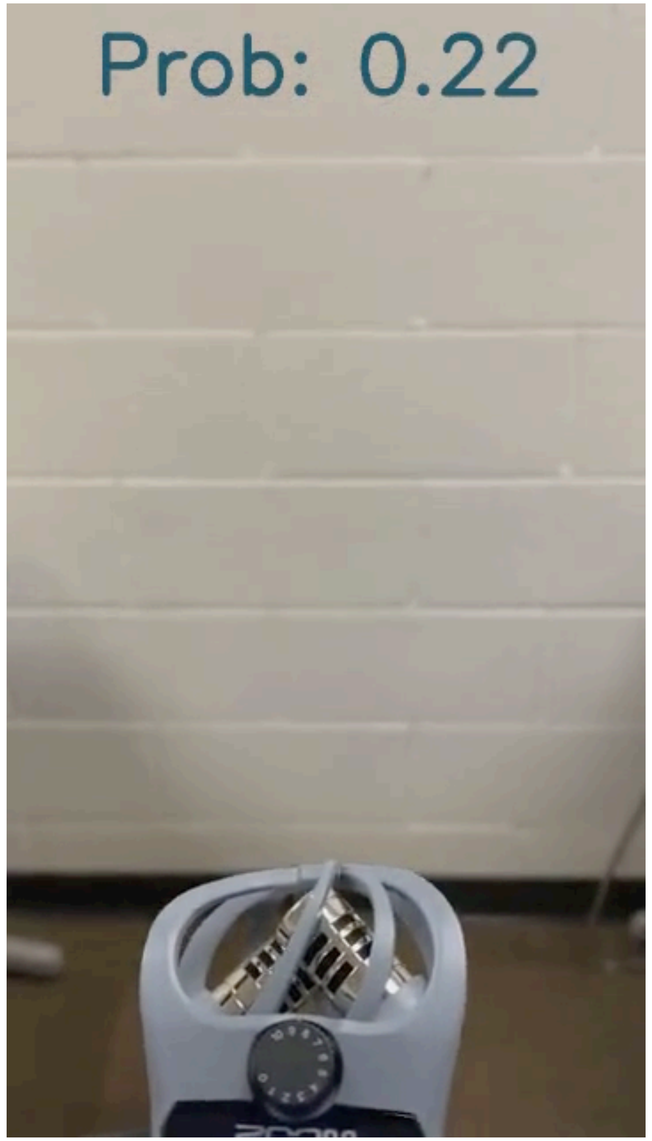


“Near a wall” prediction



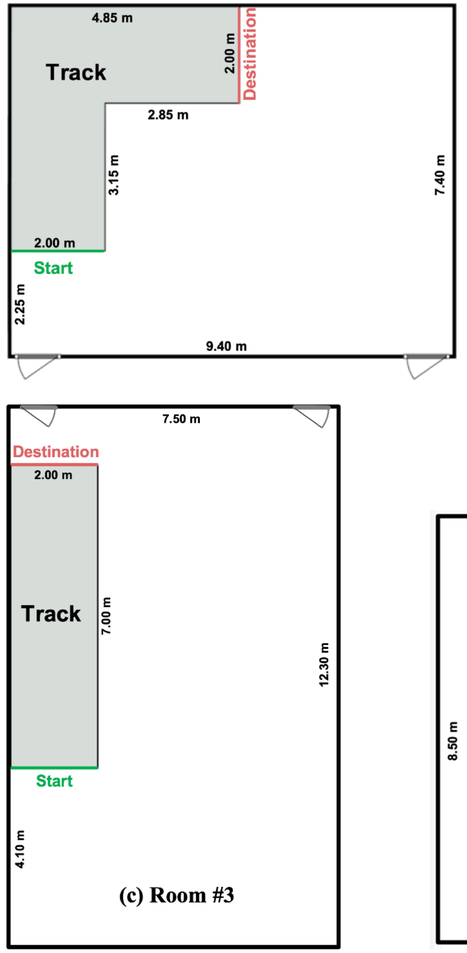


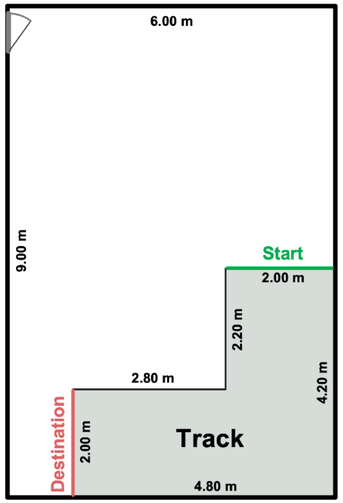
“Near a wall” prediction

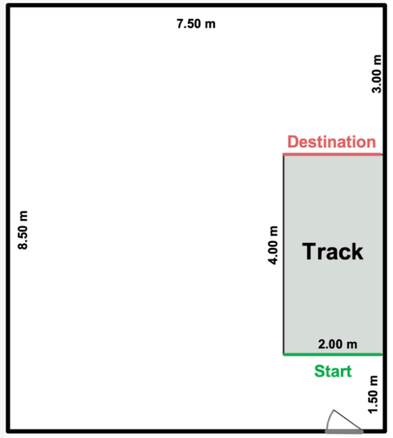


0.91	0.78	0.88	0.86	0.76
0.97	0.4	0.43	0.57	0.86
0.89	0.35	0.26	0.35	0.71
0.79	0.25	0.31	0.18	0.86
0.92	0.92	0.96	0.92	0.87

Navigating with sound







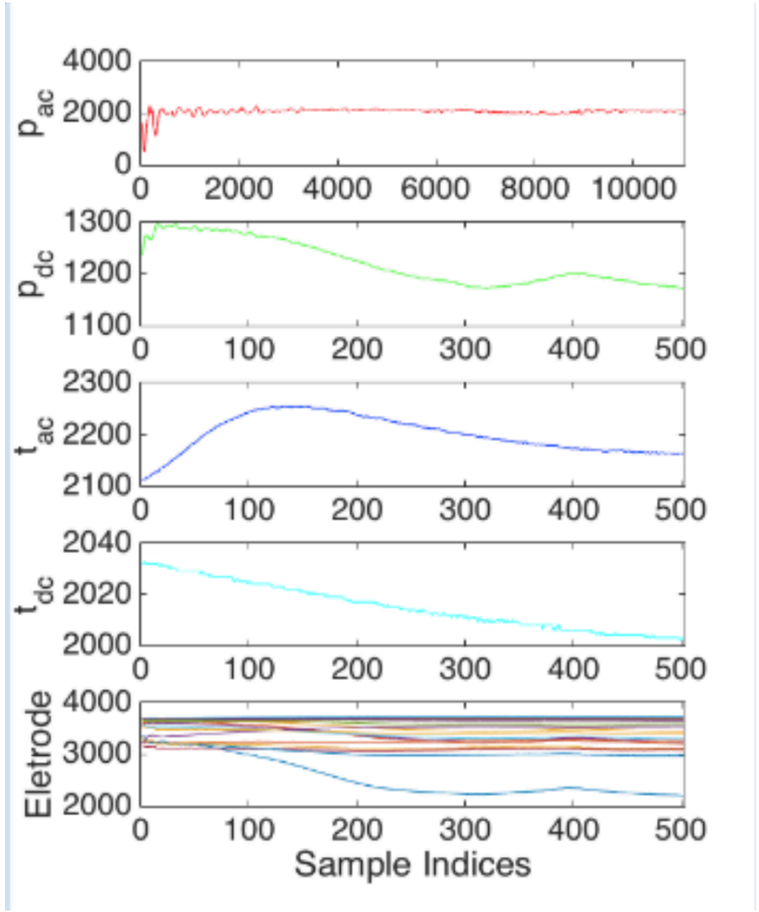
Touch sensing

Post-anesthetization
Performance

From the laboratory of
Dr. Roland Johansson
Dept. of Physiology
University of Umeå, Sweden

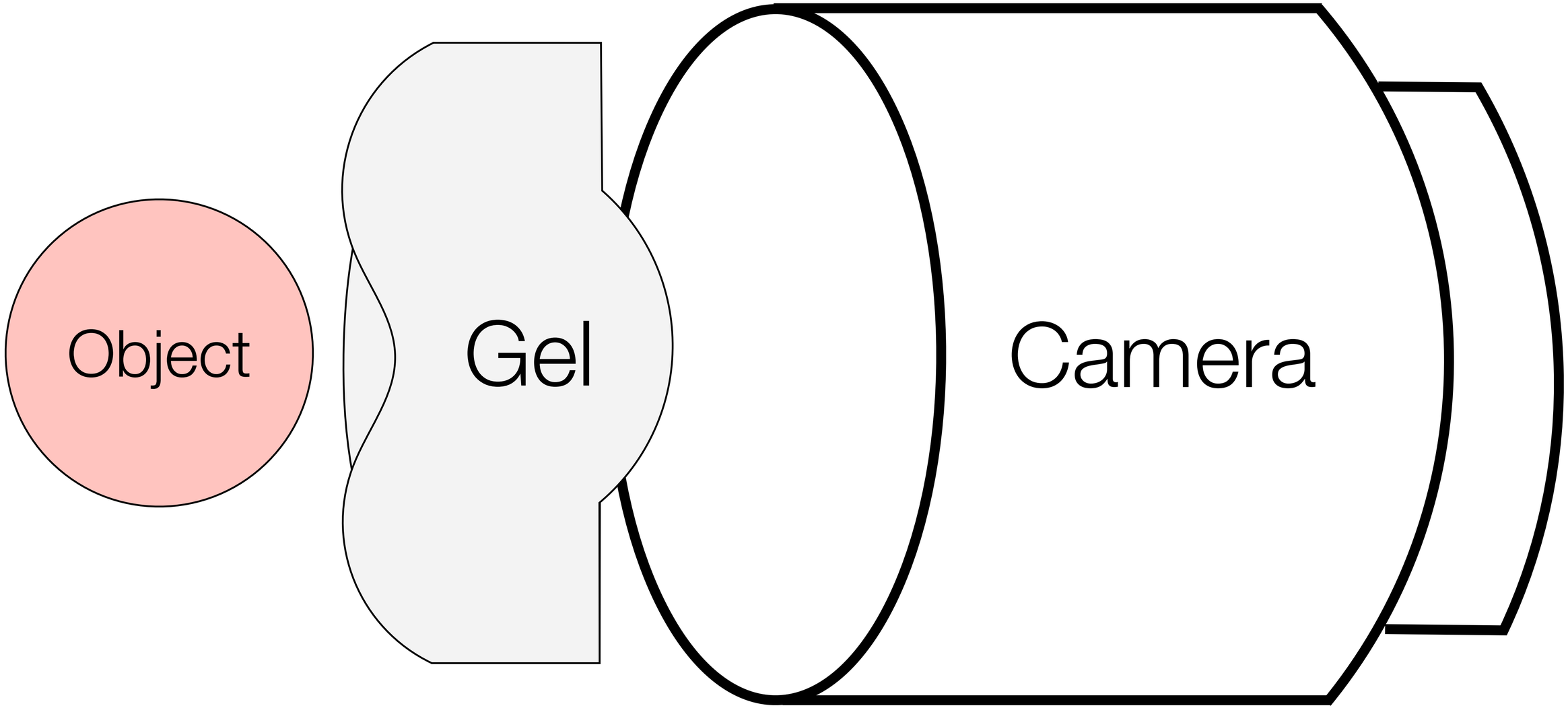
Simple touch sensors

Biotac sensor

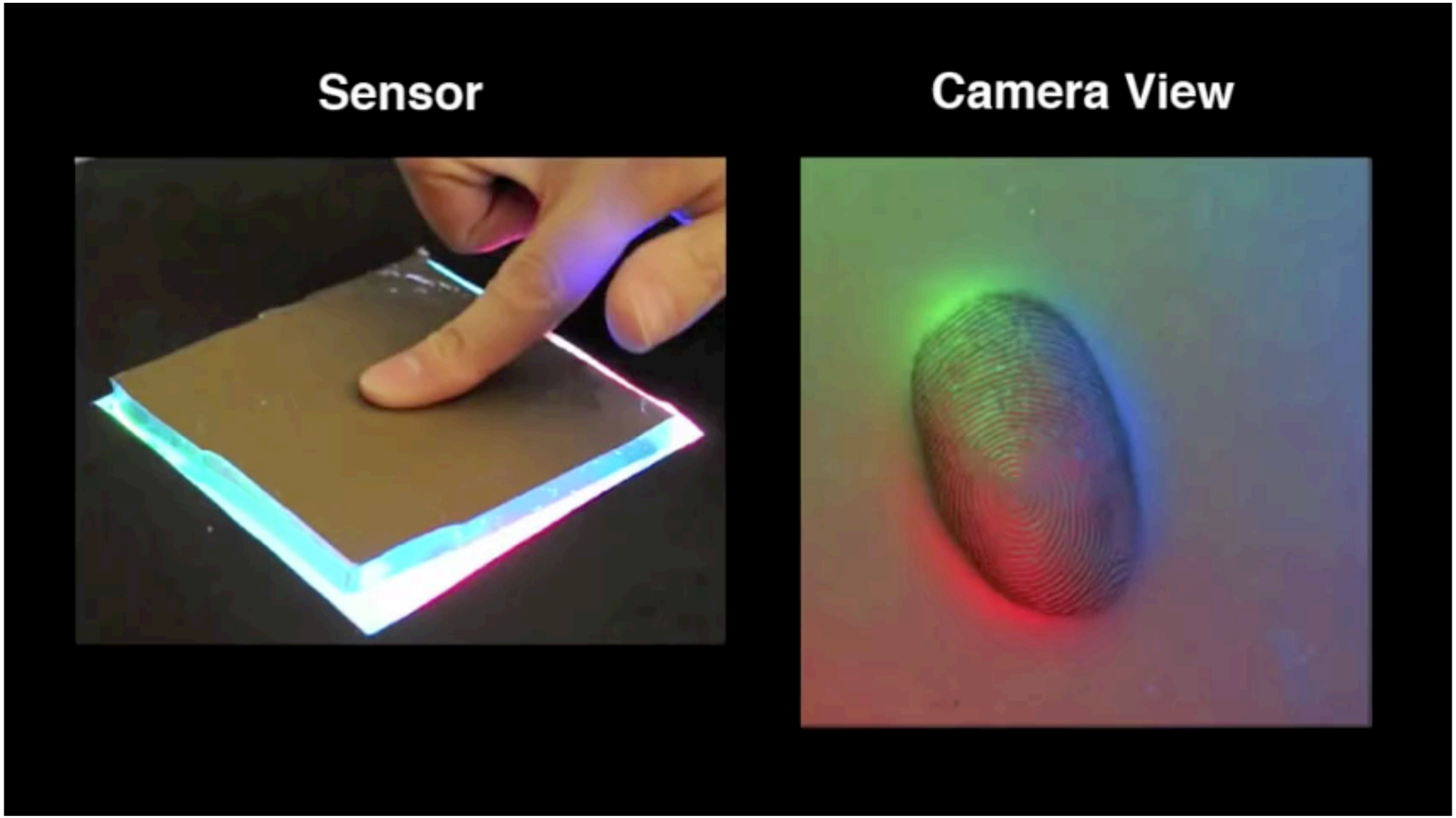


Measurements: e.g. force, vibration, temperature

Vision-based tactile sensing



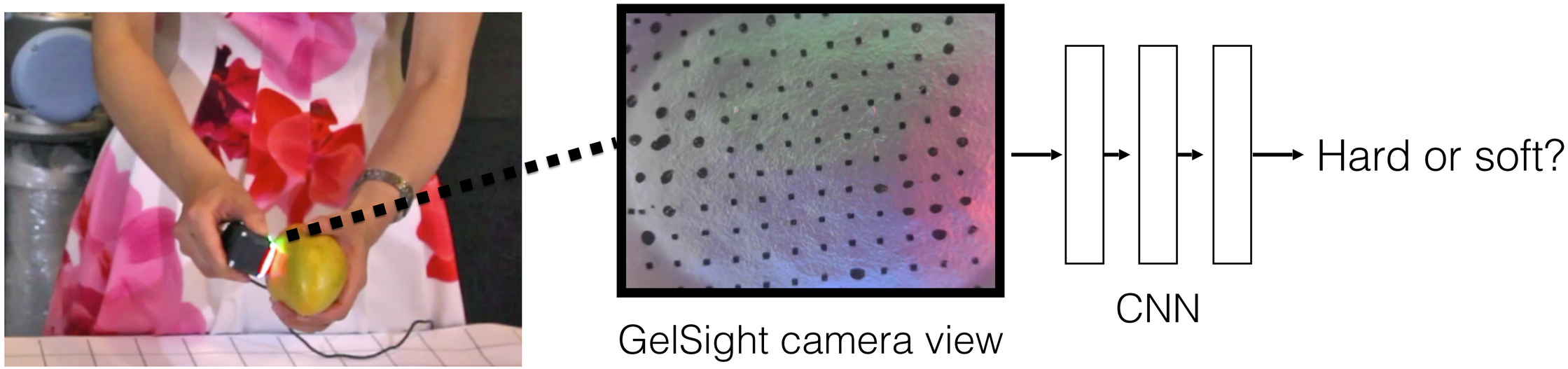
Converting touch to vision



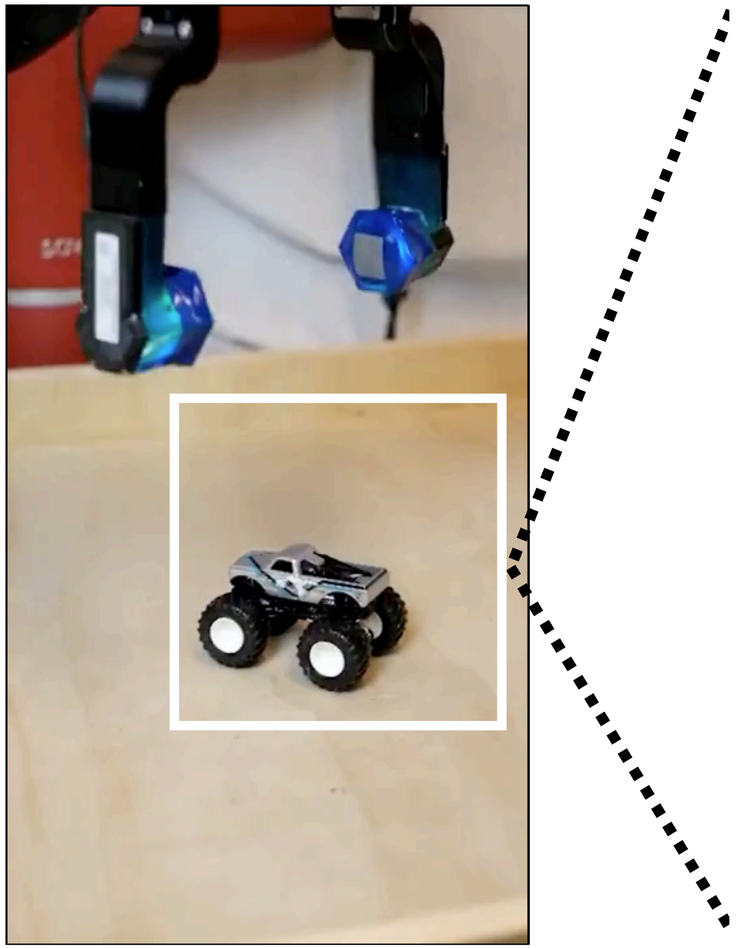
GelSight [Johnson et al. 2009]

GelSight demo

Converting touch to vision

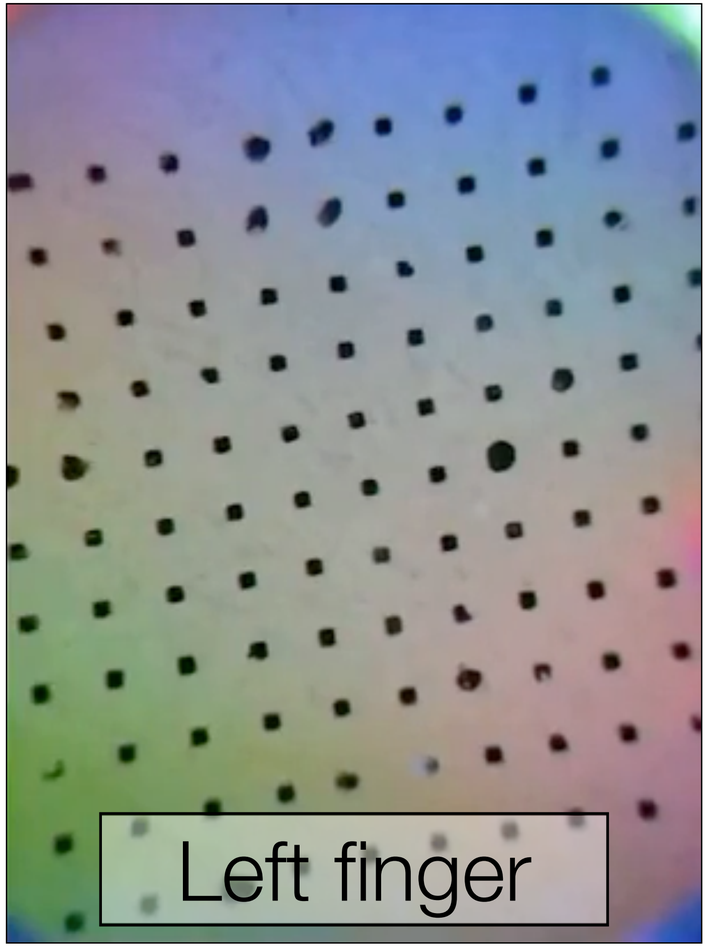


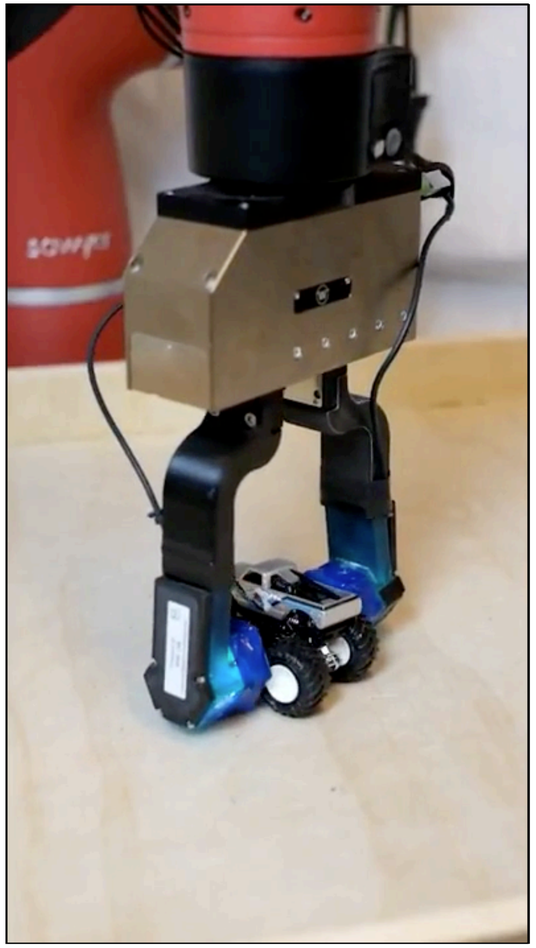
Grasping with vision and touch

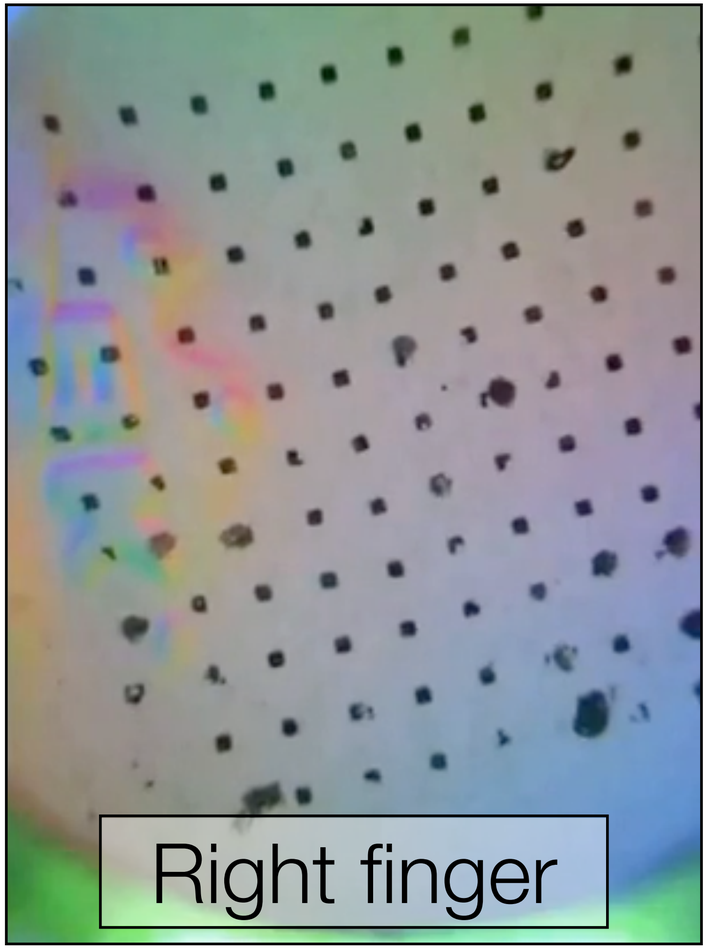


(R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. Adelson, S. Levine. RA-L 2018)
(R. Calandra, A. Owens, M. Upadhyaya, J. Lin, W. Yuan, E. Adelson, S. Levine. CoRL 2018)

Grasping with vision and touch

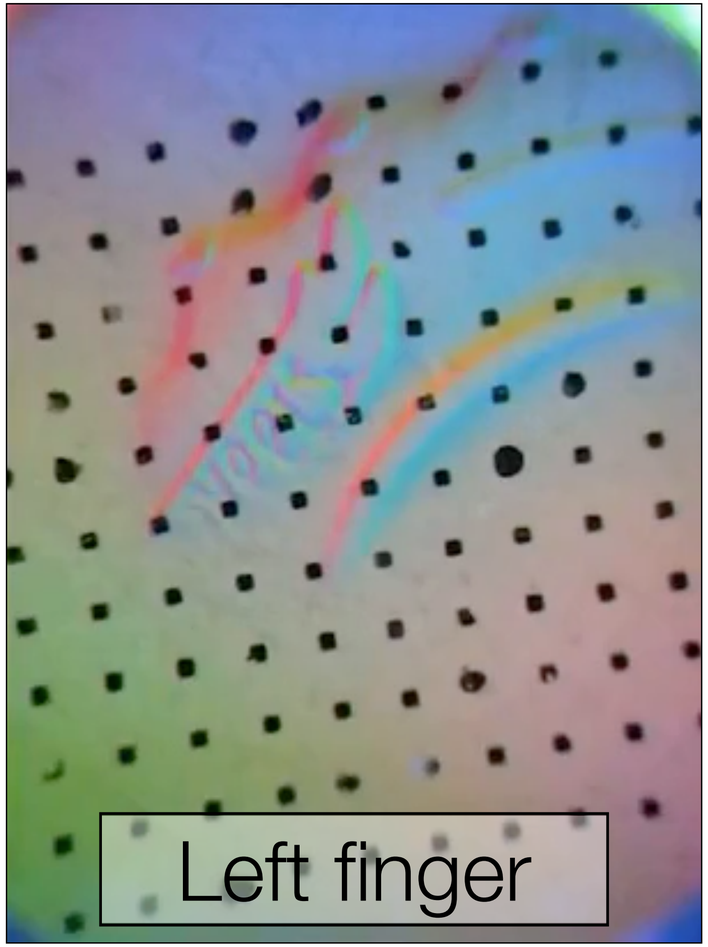


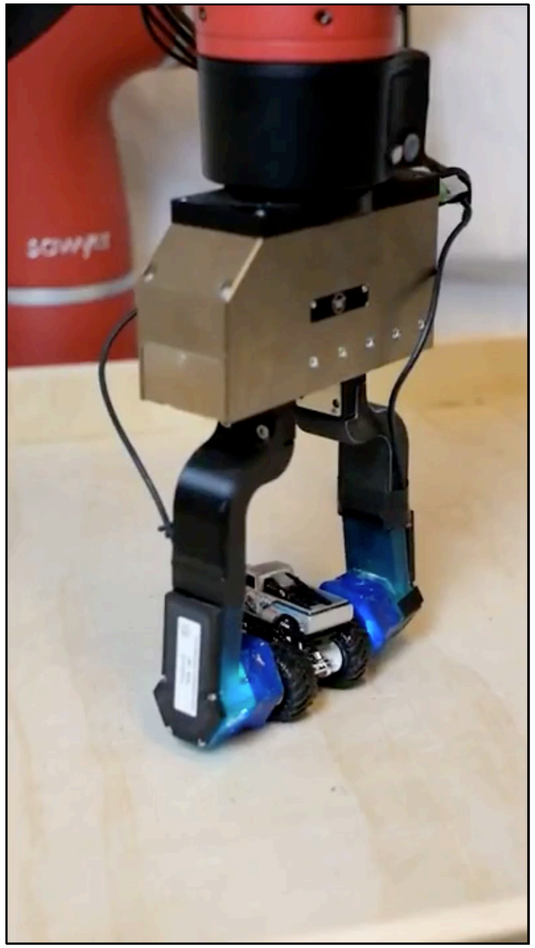


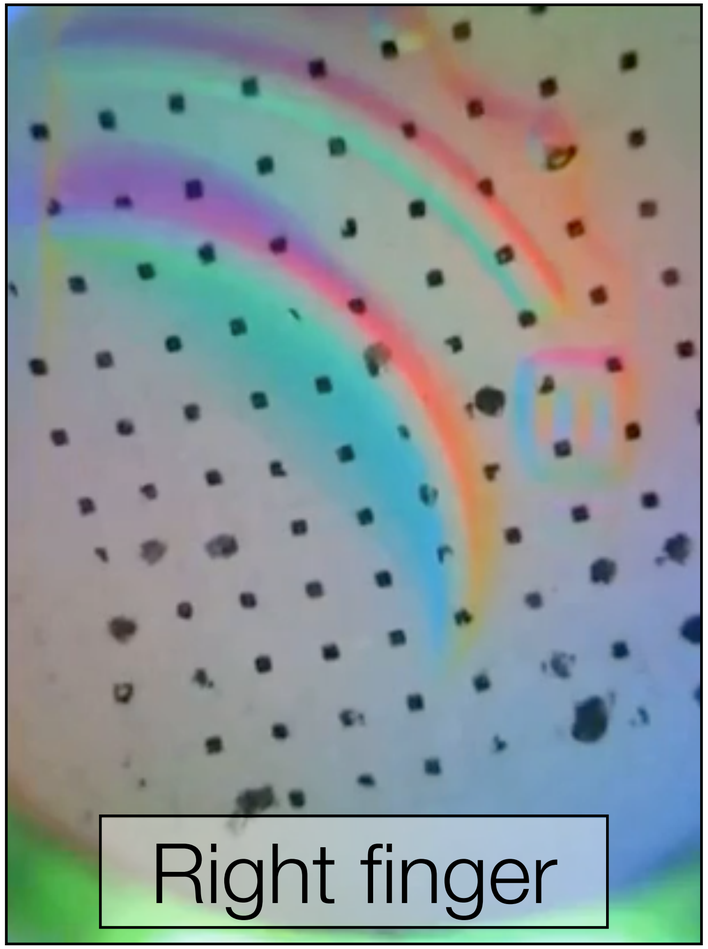


(R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. Adelson, S. Levine. RA-L 2018)
(R. Calandra, A. Owens, M. Upadhyaya, J. Lin, W. Yuan, E. Adelson, S. Levine. CoRL 2018)

Grasping with vision and touch



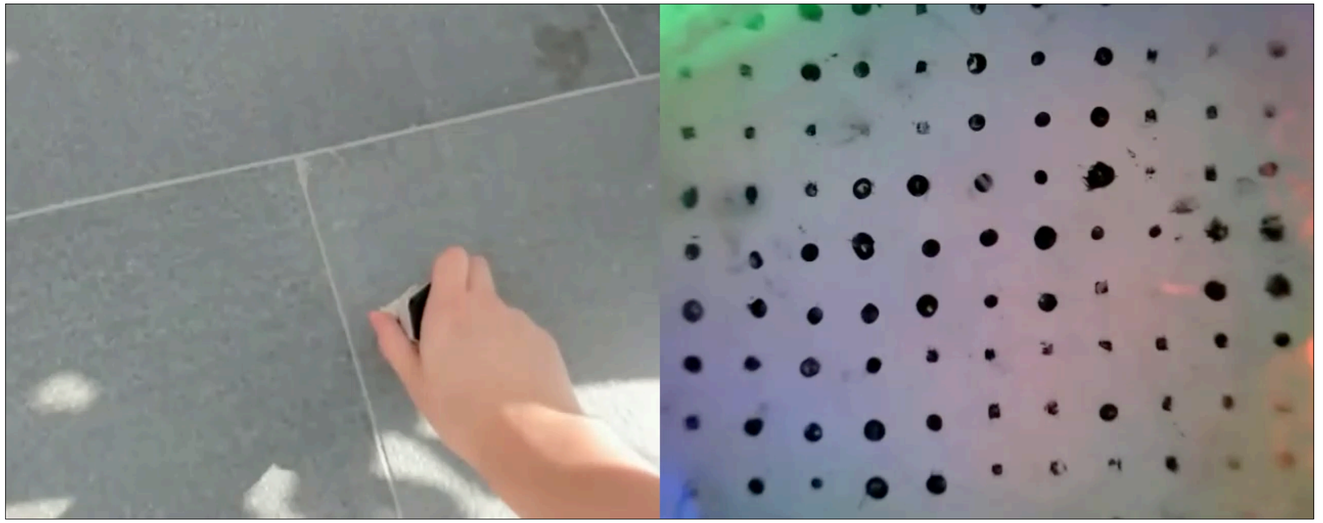


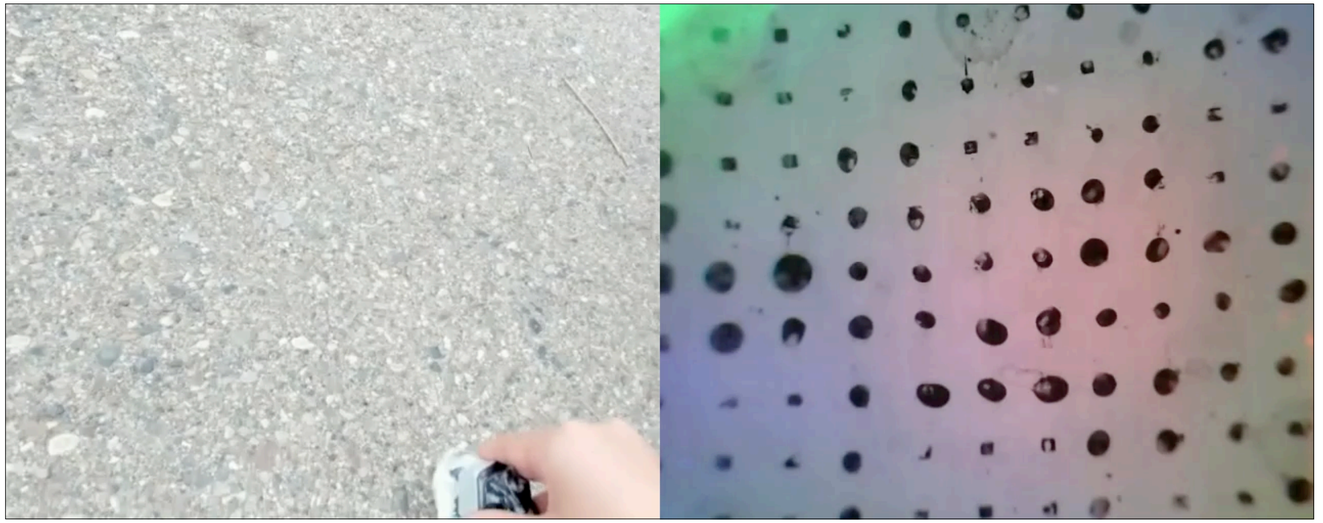


(R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. Adelson, S. Levine. RA-L 2018)
(R. Calandra, A. Owens, M. Upadhyaya, J. Lin, W. Yuan, E. Adelson, S. Levine. CoRL 2018)

Human-Powered Dataset Collection

Data Collection

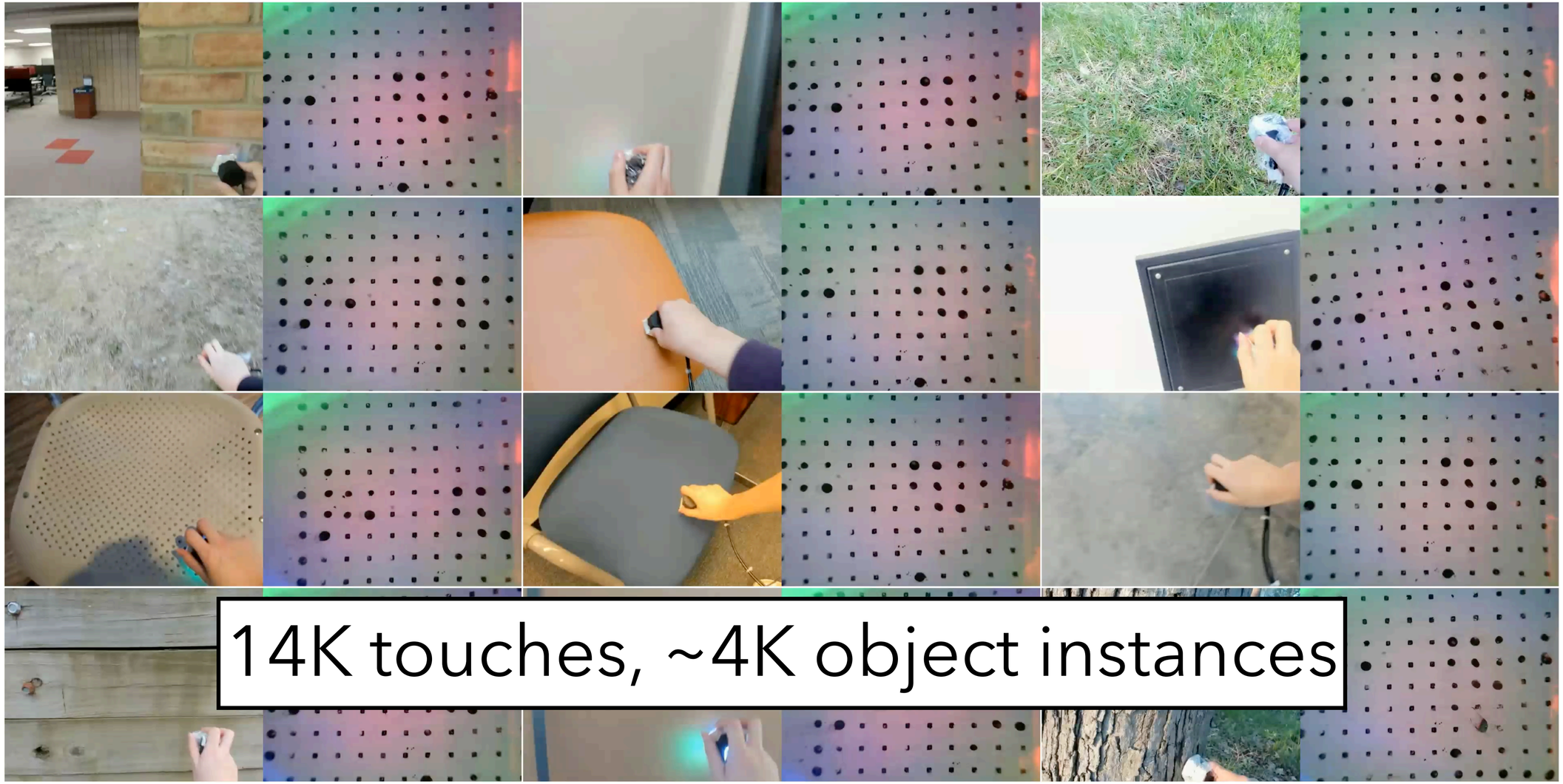




Visual

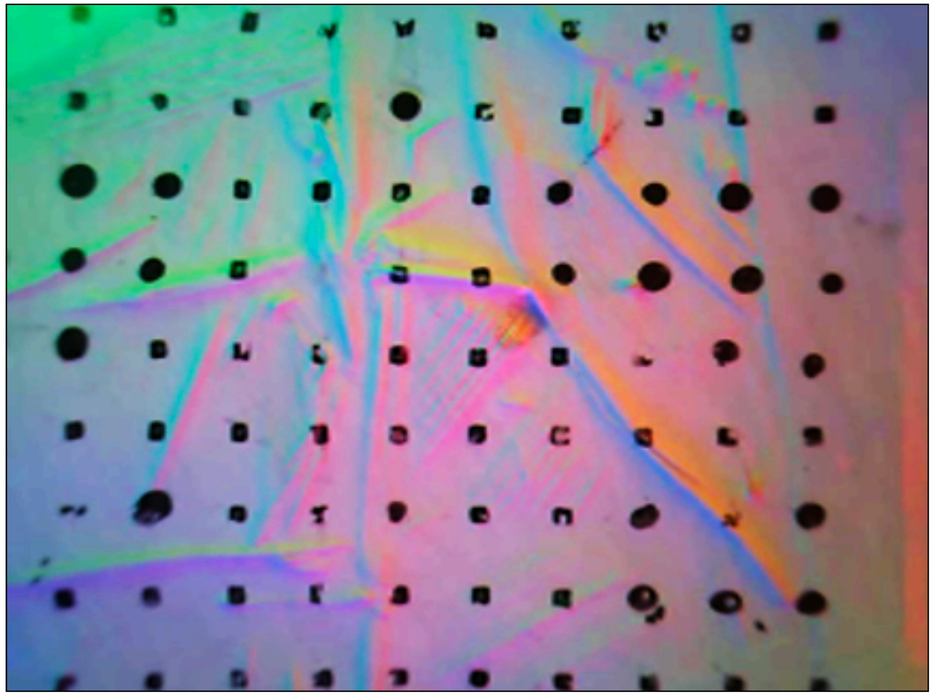
Tactile (GelSight)

Touch and Go



[Yang*, Ma*, Zhang, Zhu, Yuan, Owens, NeurIPS 2022 - Datasets and Benchmarks]

Contrastive visuo-tactile learning

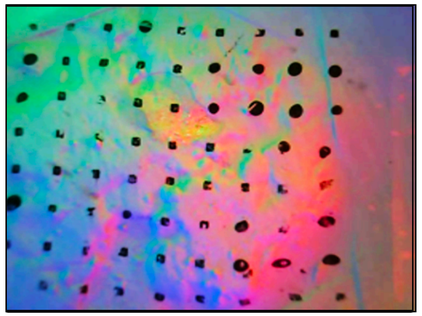
(, )

→ **real** or fake?

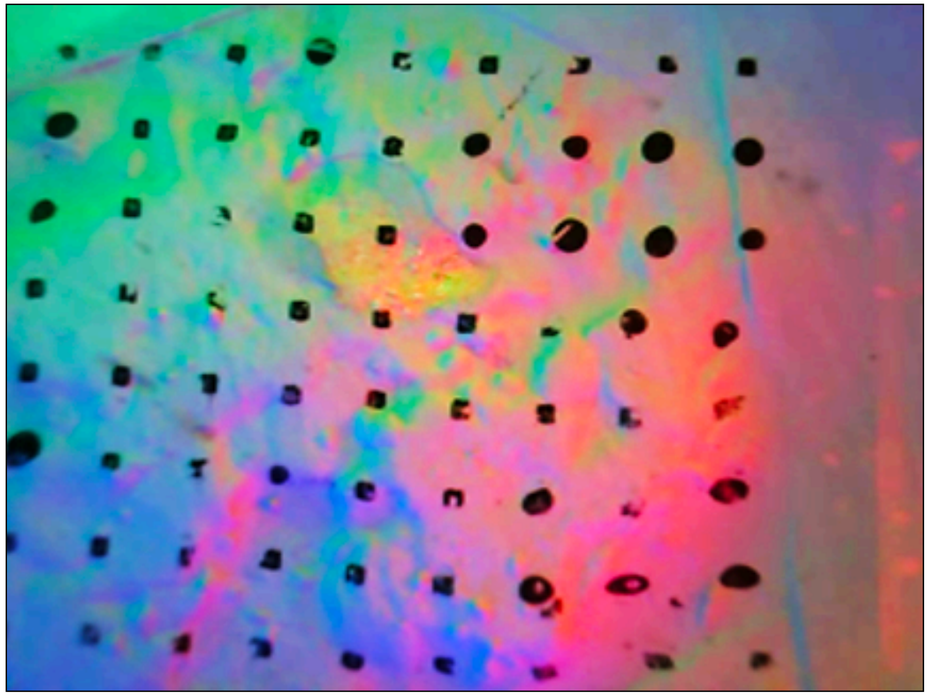
Contrastive visuo-tactile learning

(,)

→ real or **fake**?

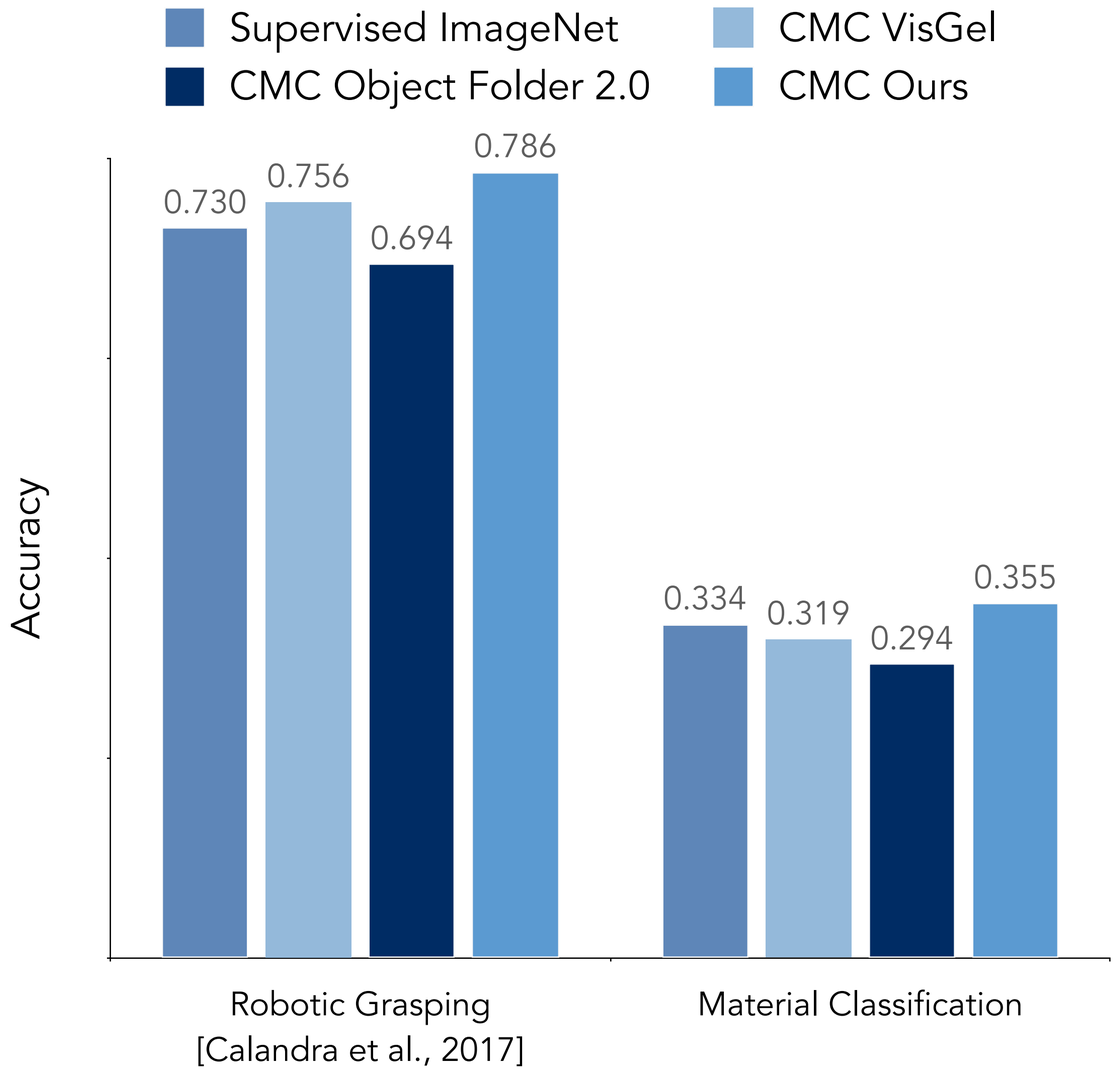


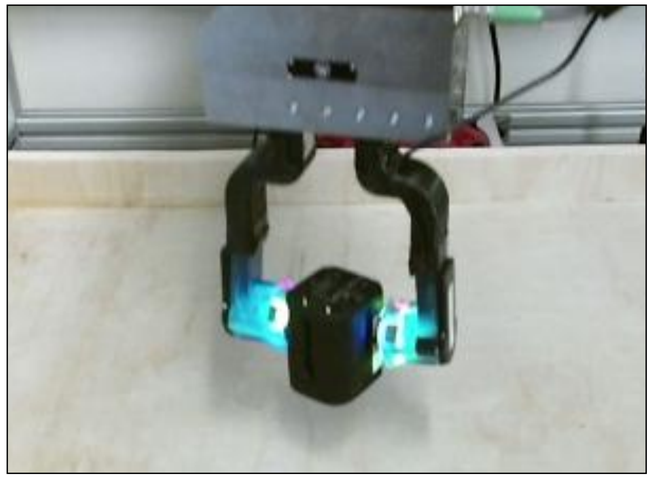
Contrastive visuo-tactile learning

(, )

→ real or **fake**?

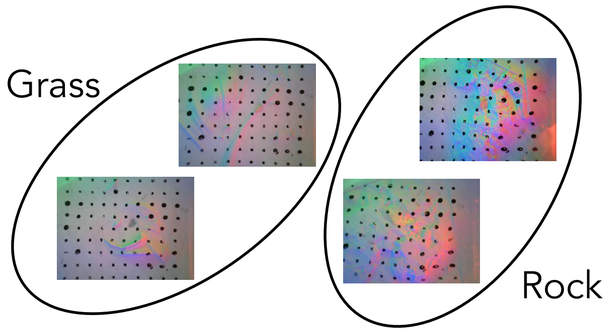
Touch features generalize to a robotic grasping task





Grasp success prediction

[Calandra, Owens, Upadhyaya, Yuan, Adelson, Levine, 2017]



Material recognition

Generating images from touch

Touch

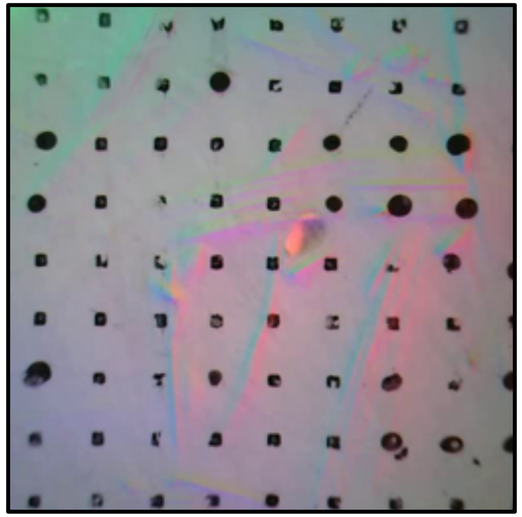
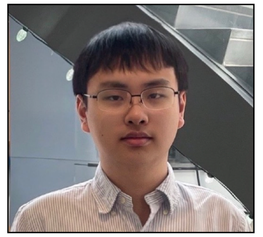


Image (reference)

Generated image

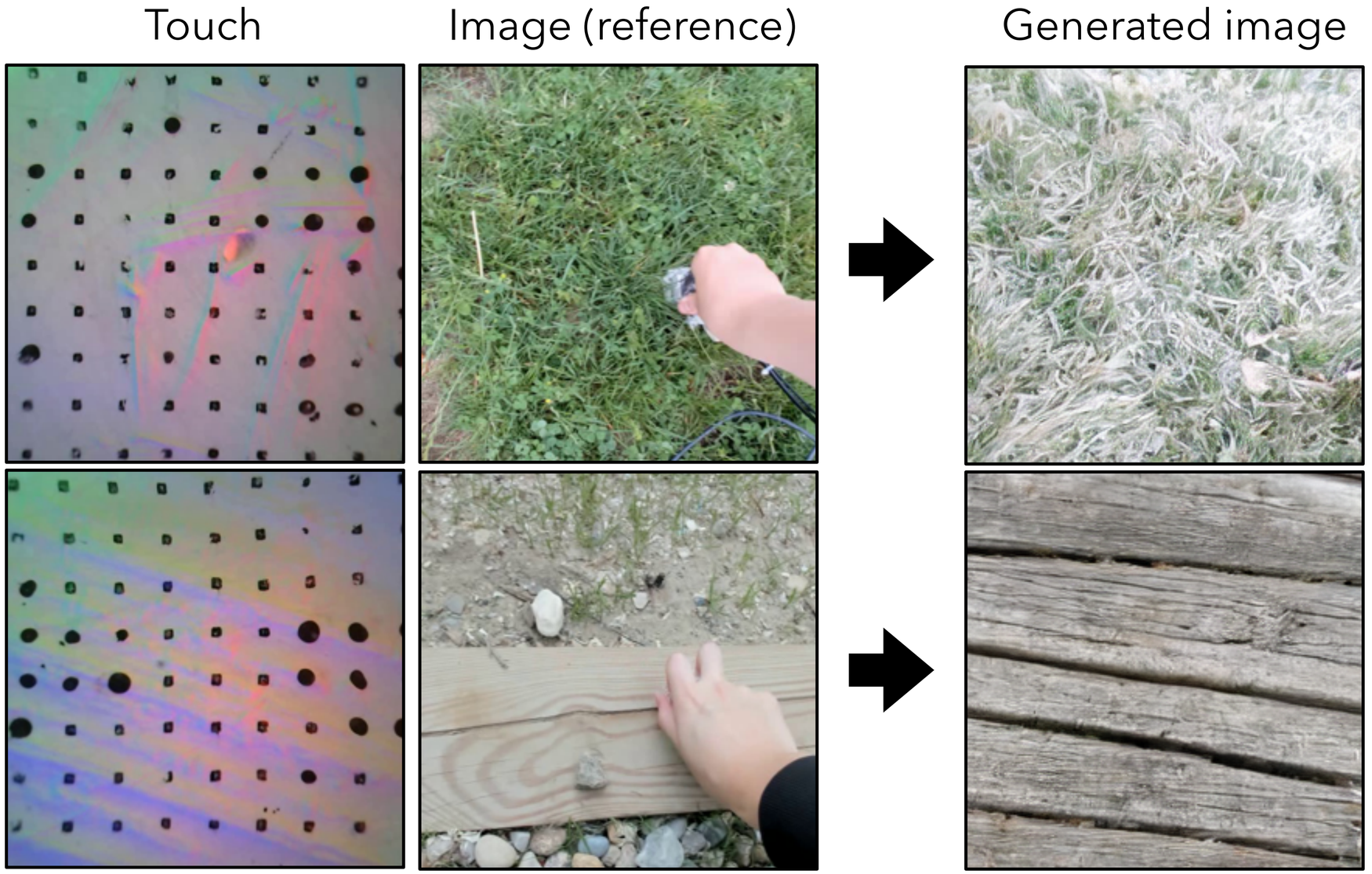
[Yang, Zhang, Owens, “Generating Visual Scenes from Touch”, ICCV 2023]



Fengyu Yang

Jiacheng Zhang

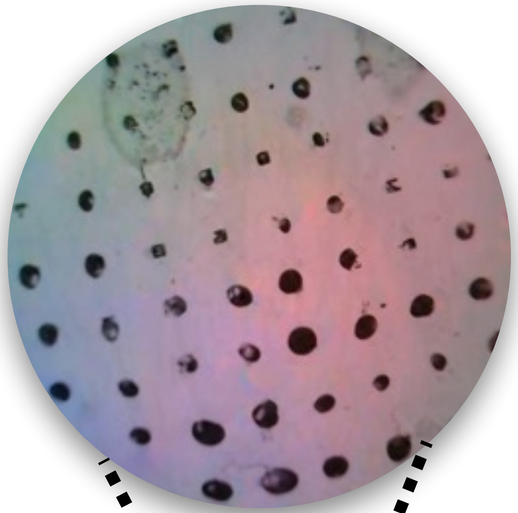
Generating images from touch



[Yang, Zhang, Owens, “Generating Visual Scenes from Touch”, ICCV 2023]

Tactile-driven stylization

Input Touch



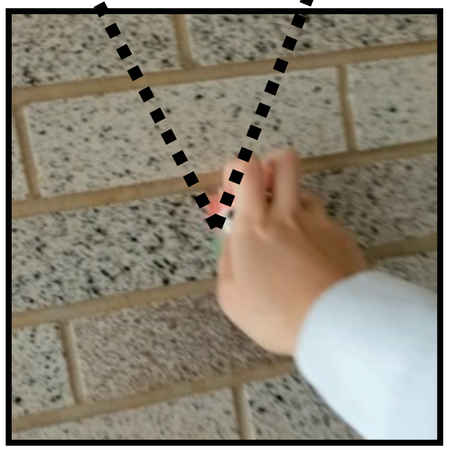


Image (not used)

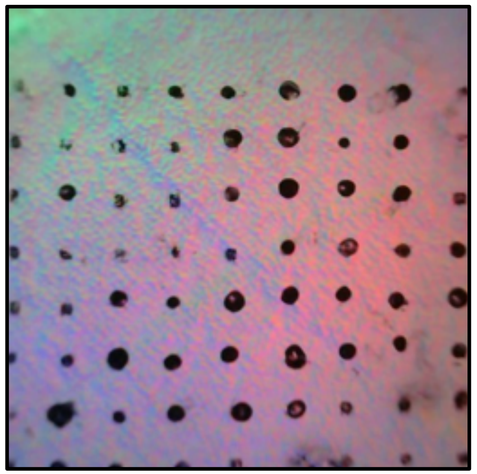
Input Image

SDEdit

Restyled Image

Based on SDEdit [Meng et al., 2022]

Tactile-driven stylization



Tactile input

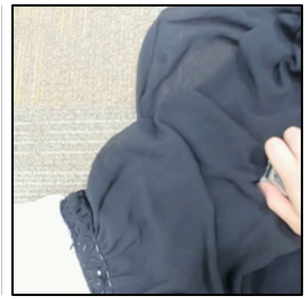
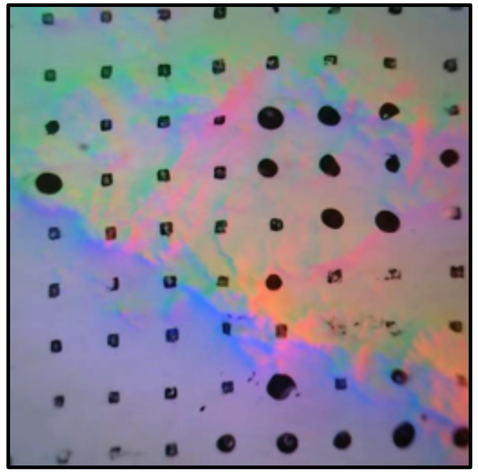


Image
(as reference)

Input image

Stylized image

Tactile-driven stylization



Tactile input

Image
(not used)

Input image

Stylized image

Summary

1. **Multimodal models**: different modalities provide complementary information
2. **Self-supervision**: One sensory signal can provide a learning signal to another
3. **Same methods, different signals:** We can reduce “non-visual” problems into problems that can be solved with neural nets designed for vision

Next time: image formation