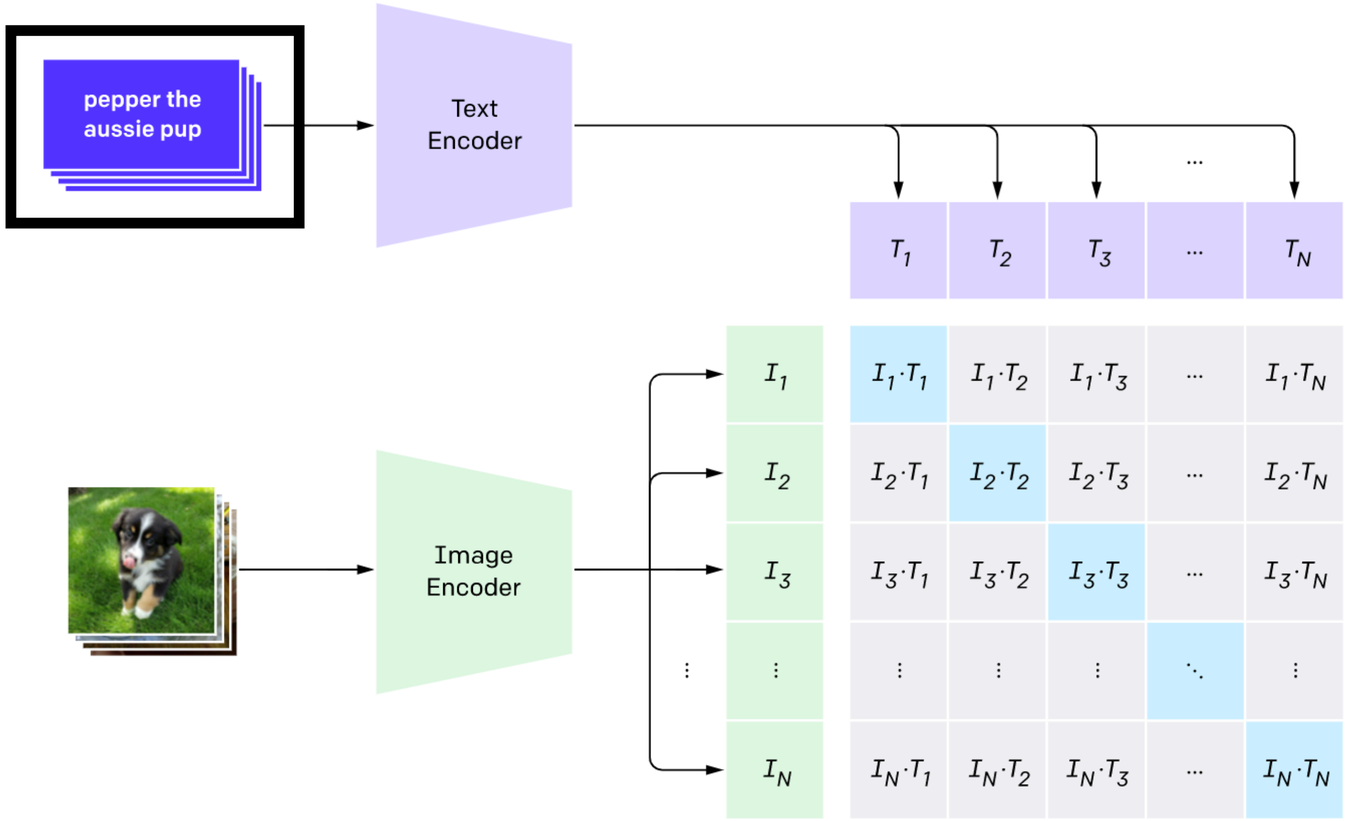
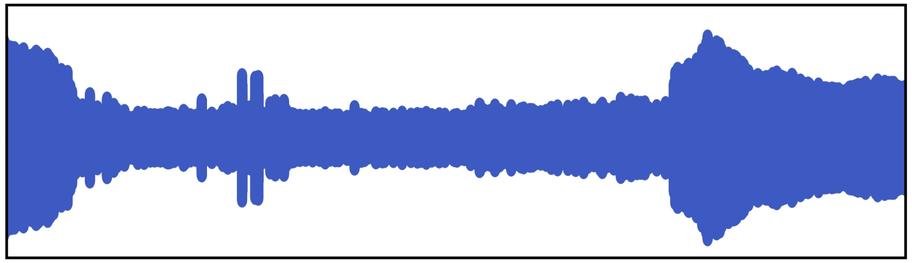
Lecture 11: Sound

Vision and language



Let's focus on *low level modalities* instead

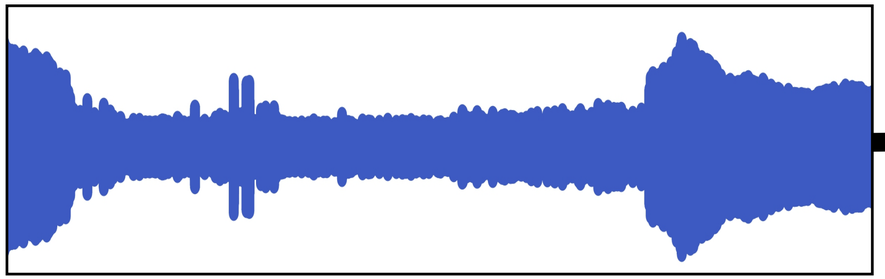
Self-supervision



Self-supervision

Vision

...



Hearing

...

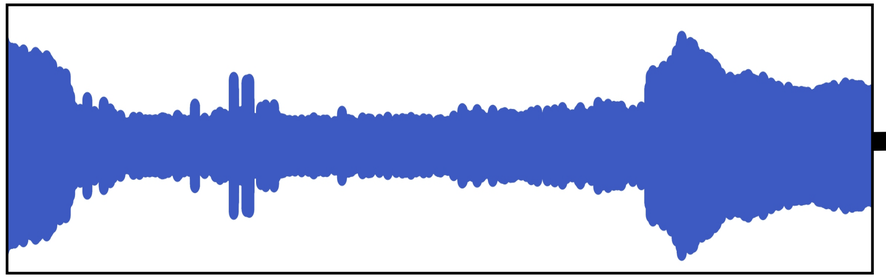
(de Sa 1994, Smith 2005)

Self-supervision

Vision

...

Prediction



Hearing

...

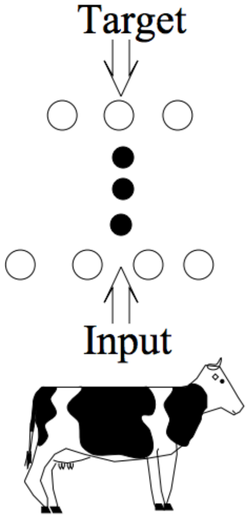
(de Sa 1994, Smith 2005)

Self-supervision

Supervised

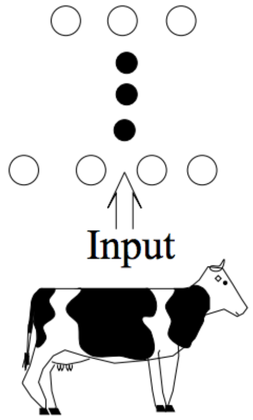
- implausible label

"COW"



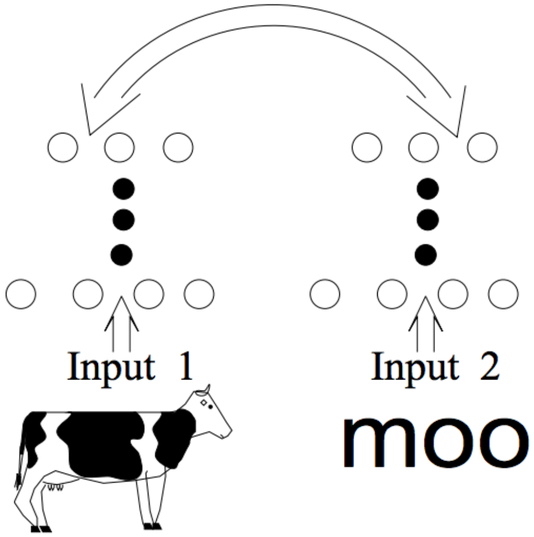
Unsupervised

- limited power

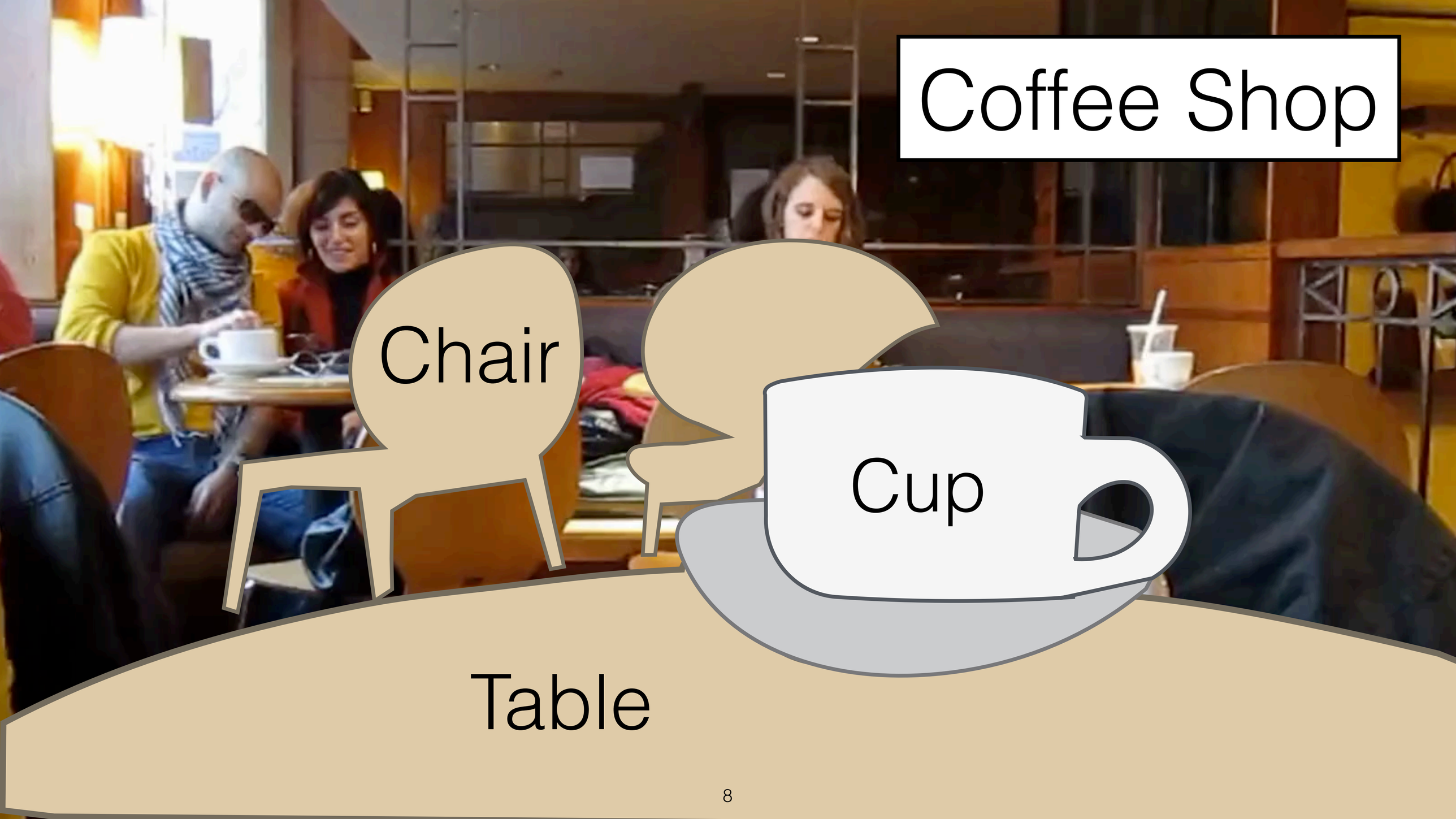


Self-Supervised

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



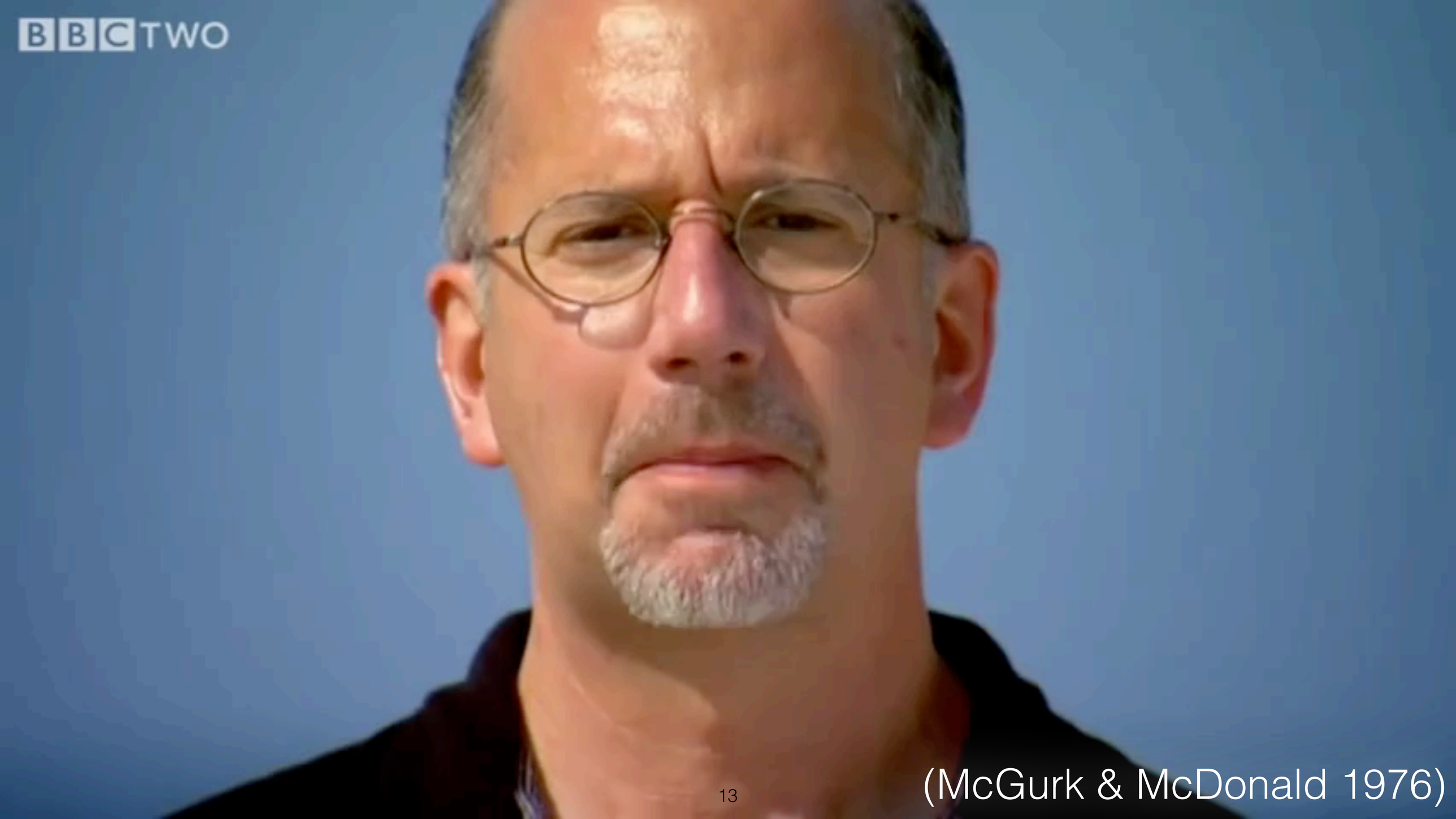
Virginia de Sa. Learning Classification with Unlabeled Data. NIPS 1994.


Coffee Shop
Chair
Cup
Table

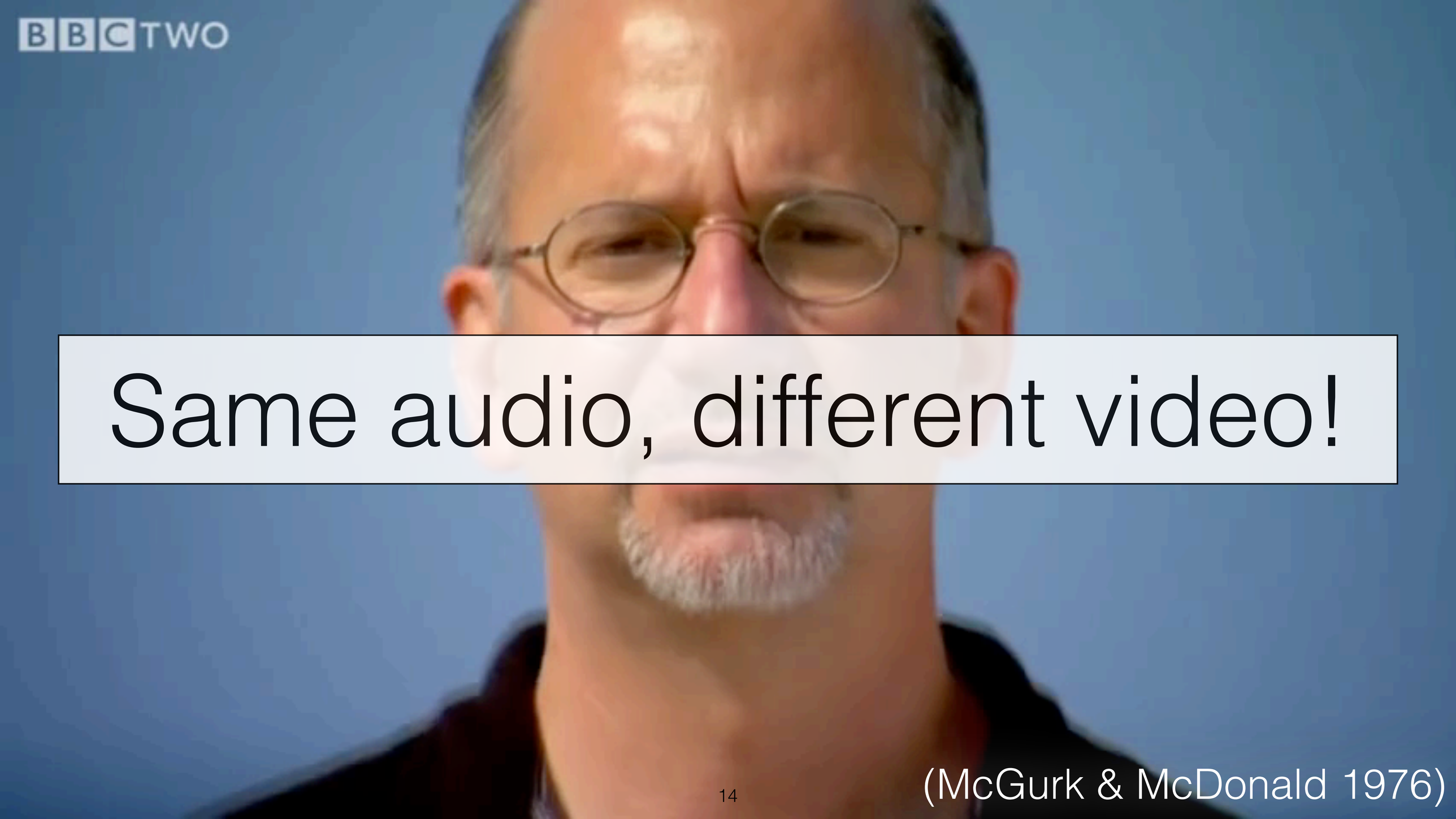

SHIPPING
RECEIVING


SHIPPING
RECEIVING

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



(McGurk & McDonald 1976)



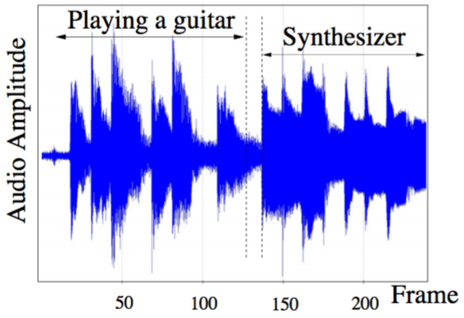
Same audio, different video!

(McGurk & McDonald 1976)

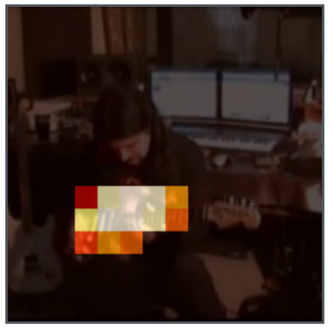
(McGurk & McDonald 1976)


How do we get ground-truth correspondences?

Sound source localization

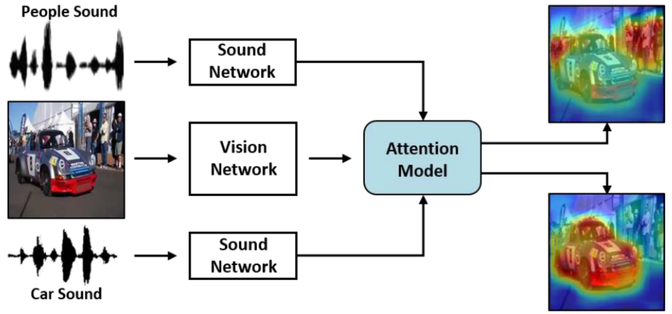


Kidron, Schechner, Elad. Pixels that Sound. CVPR 2005.

Arandjelović and Zisserman. Objects that Sound. ECCV 2018.

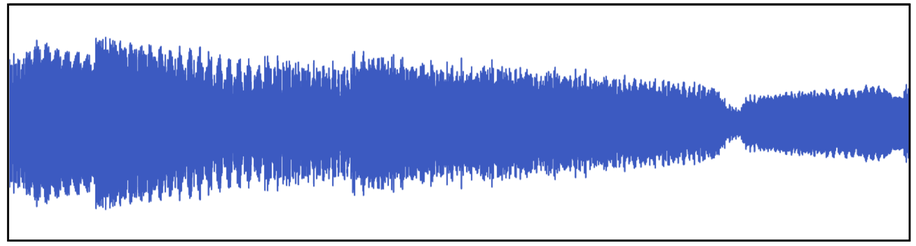

Hershey and Movellan. Audio vision: Using audio-visual synchrony to locate sounds. NIPS 2000.



Senocak, Oh, Kim, Yang, Kweon. Learning to Localize Sound Source in Visual Scenes. CVPR 2018.

Contrastive audio-visual learning

(

, )

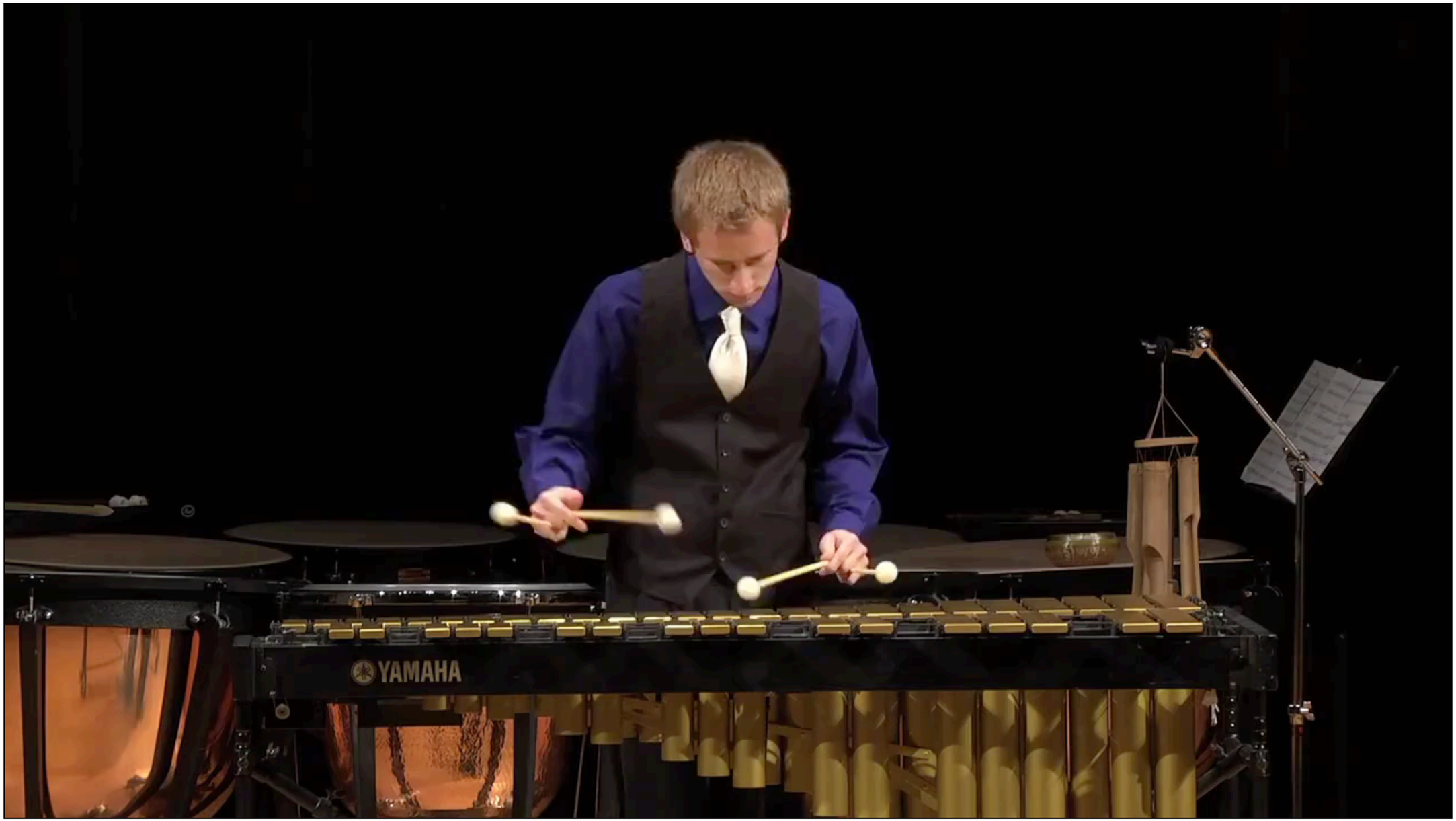
→ **real** or fake ?

Idea #1: correspondence problem

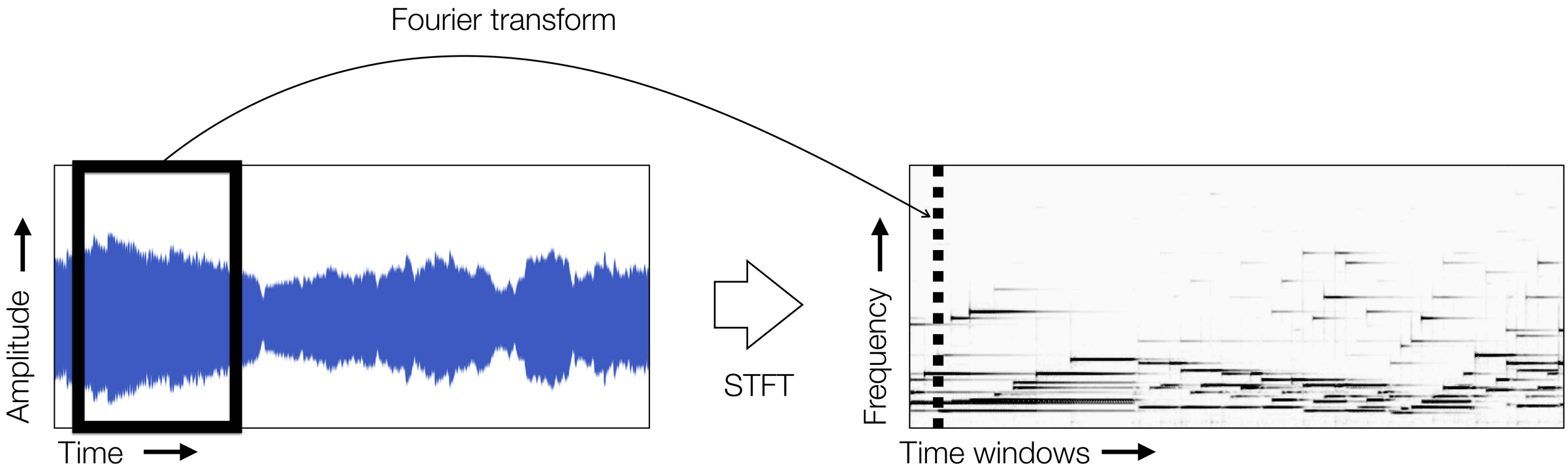
,

(Arandjelović & Zisserman 2017)

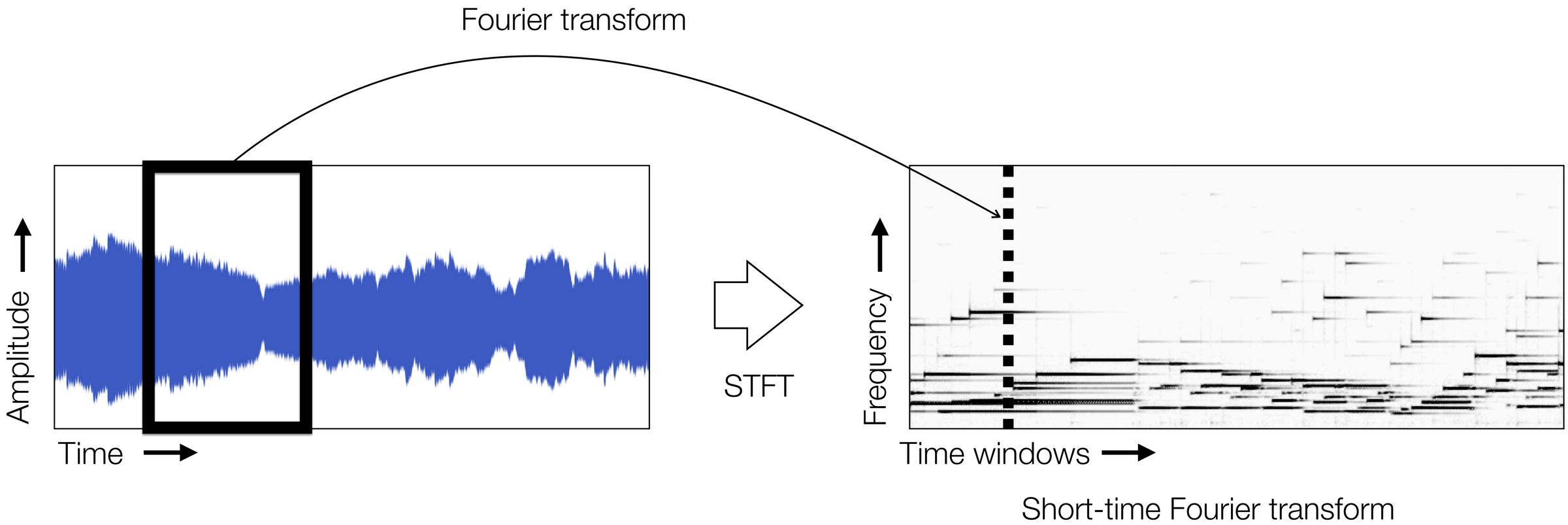
Idea #1: correspondence problem



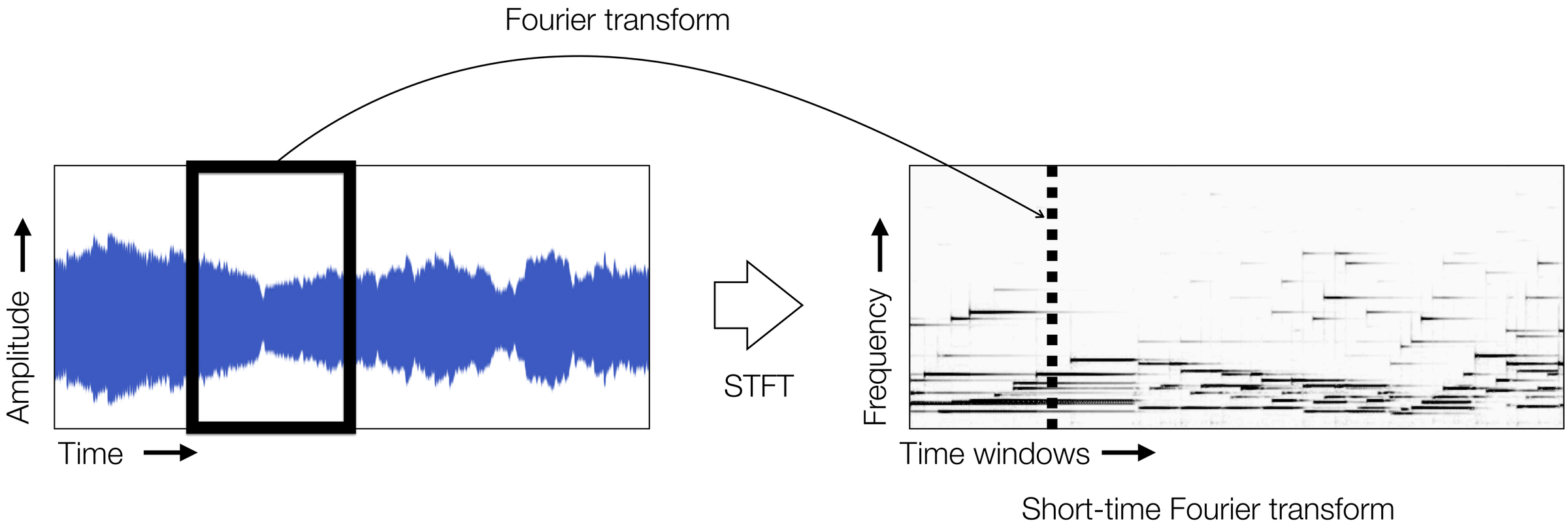
Spectrogram



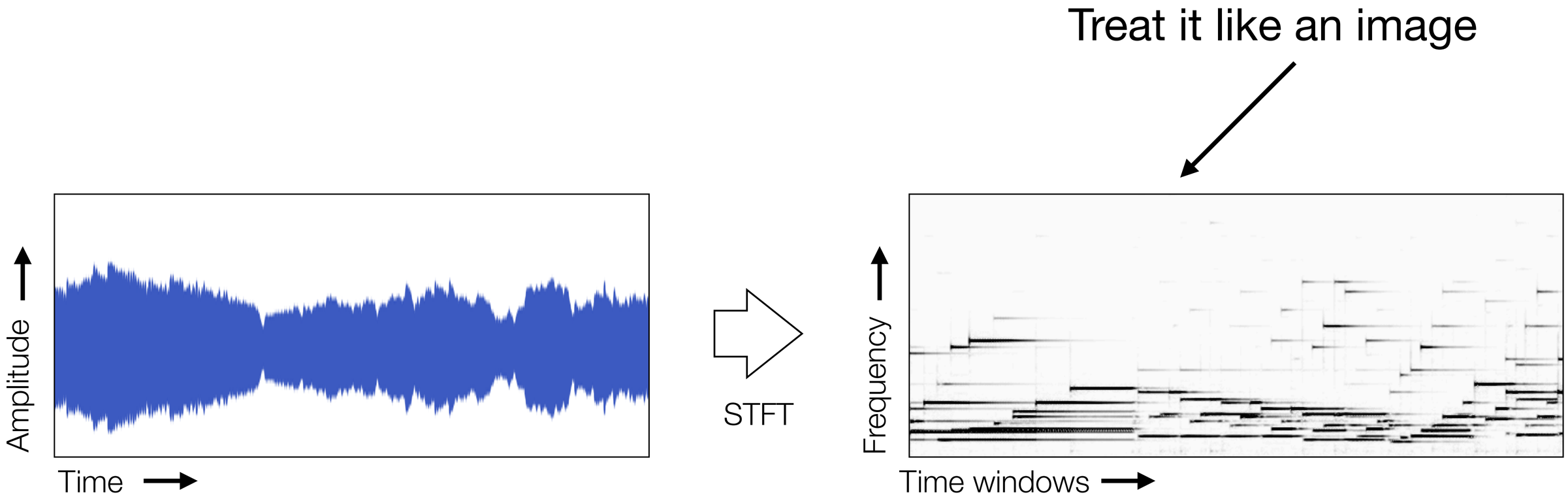
Spectrogram



Spectrogram

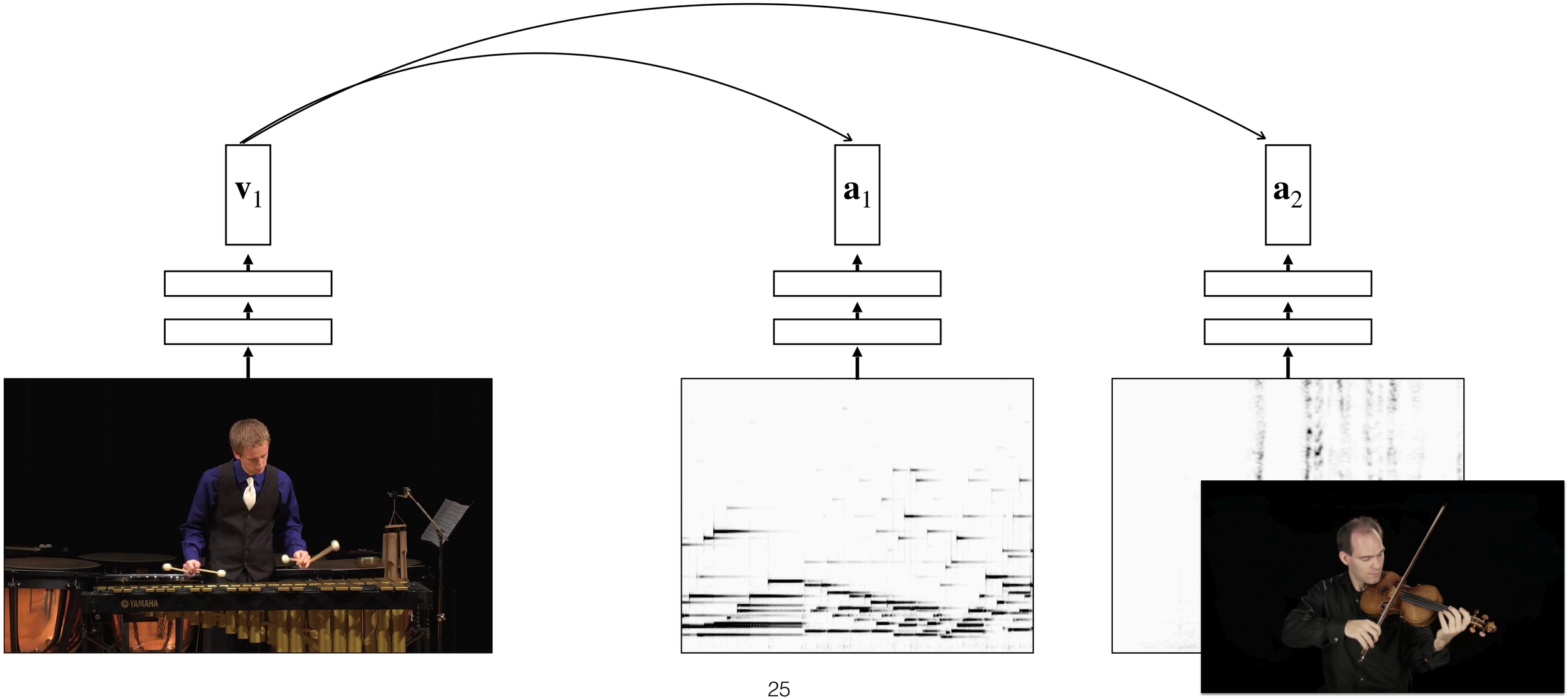


Spectrogram



Short-time Fourier transform

Contrastive audio-visual learning

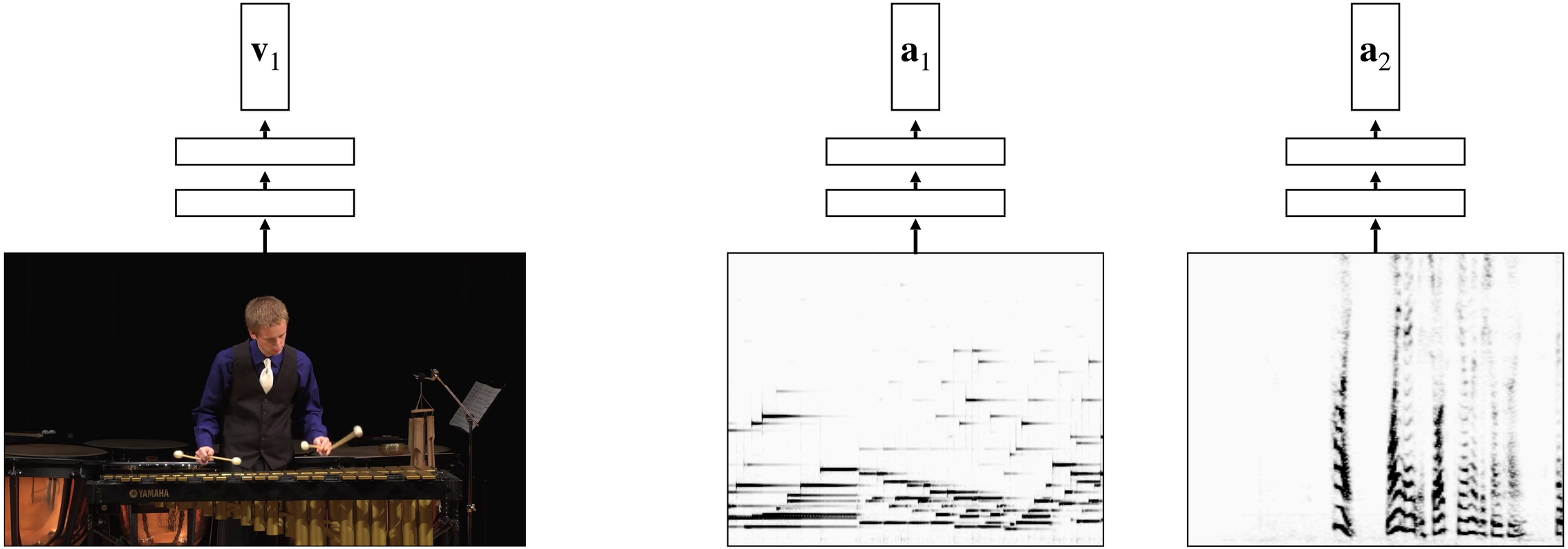


Contrastive audio-visual learning

Learn networks where:

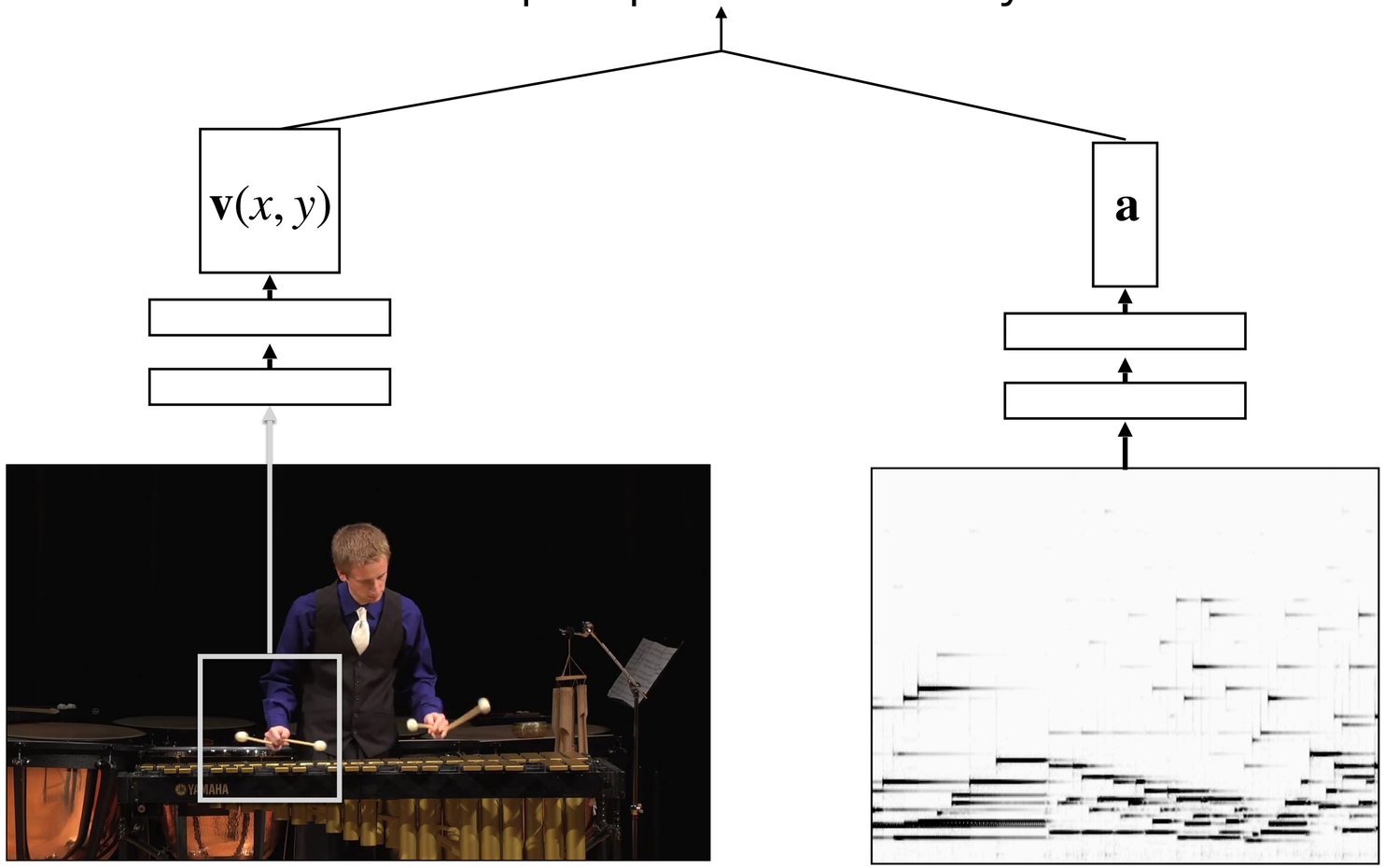
$\mathbf{v}_1^\top \mathbf{a}_1 \longrightarrow$ High dot product with self

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



Sound source localization

Combine per-patch similarity scores



Overall image-audio similarity:

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

Or mean:

$$\frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \mathbf{v}(x, y)^\top \mathbf{a}$$

[R. Arandjelovic and A. Zisserman, 2018]

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

Input single frame

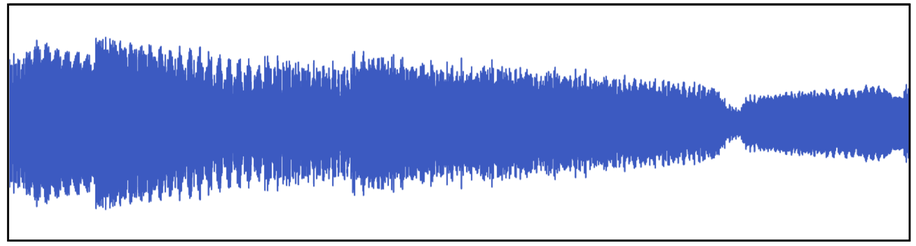
Frame/ Localization overlaid

Localization

[R. Arandjelovic and A. Zisserman, 2018]

Idea #2: synchronization problem

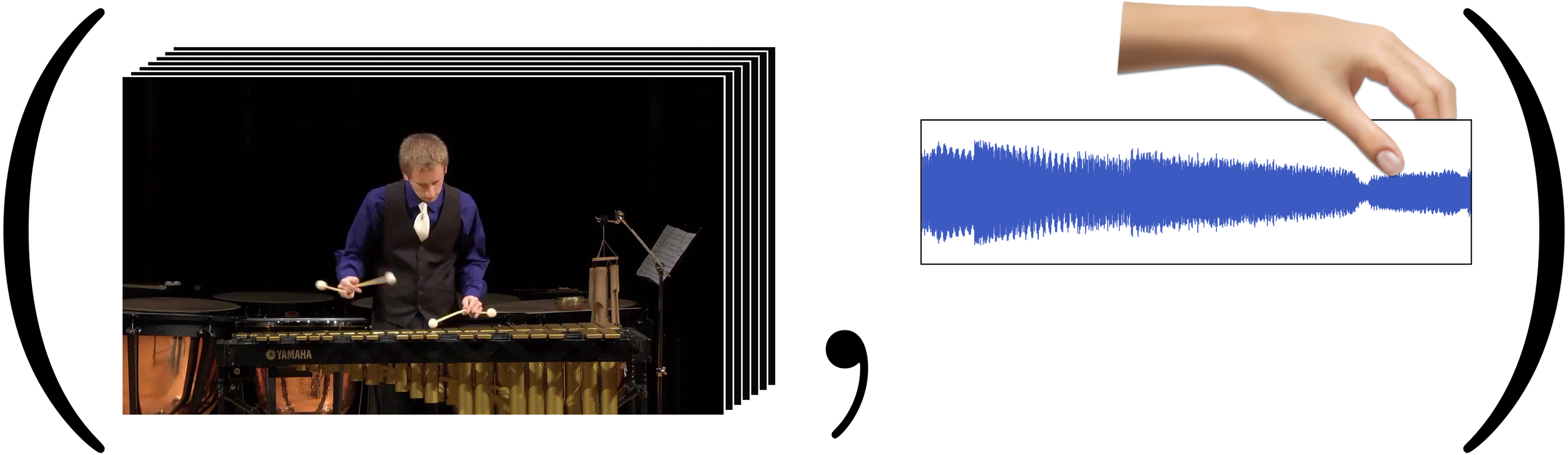
(

, )

→ **real** or fake ?

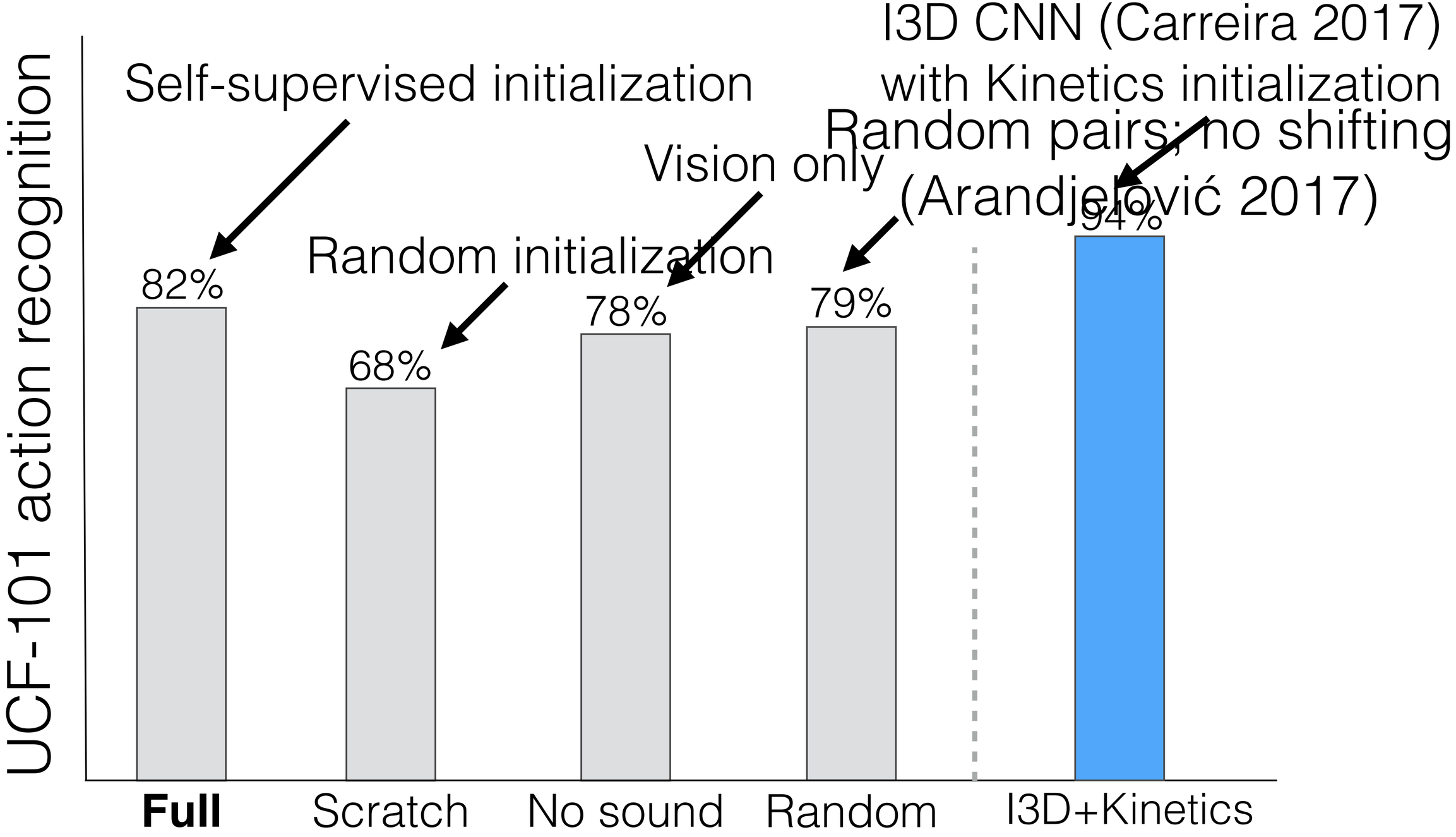
[Owens & Efros, 2018]

Idea #2: synchronization problem



→ real or **fake**?

[Owens & Efros, 2018]


UCF-101 action recognition
Self-supervised initialization
Random initialization
Vision only
Random pairs; no shifting
(Arandjelović 2017)
I3D CNN (Carreira 2017)
with Kinetics initialization
82%
68%
78%
79%
94%
Full
Scratch
No sound
Random
I3D+Kinetics


Doesn’t group into discrete **instances**.
Localization is not very **precise**.

Self-Supervised Learning of Audio-Visual Objects from Video

Triantafyllos Afouras

Andrew Owens

Joon Son Chung

Andrew Zisserman

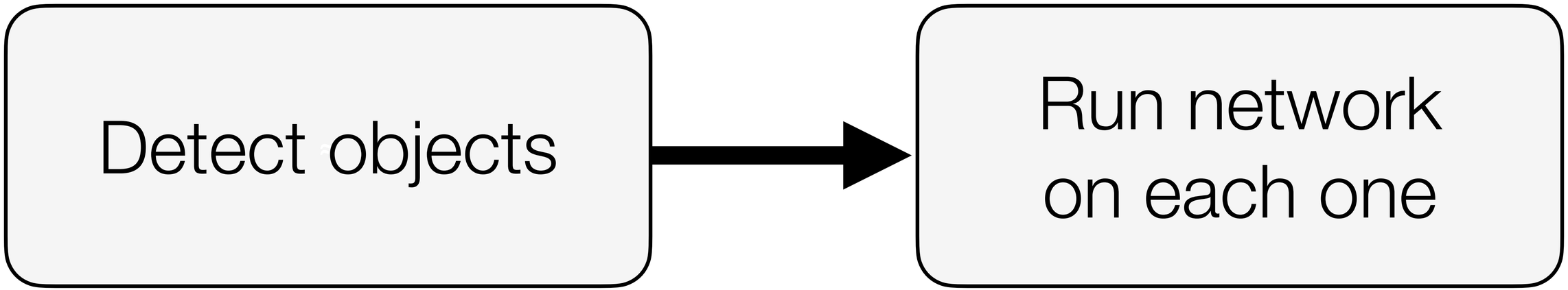
VGG University of Oxford

University of Michigan

NAVER

ECCV 2020

Group-and-Process Pipeline



Group-and-Process Pipeline

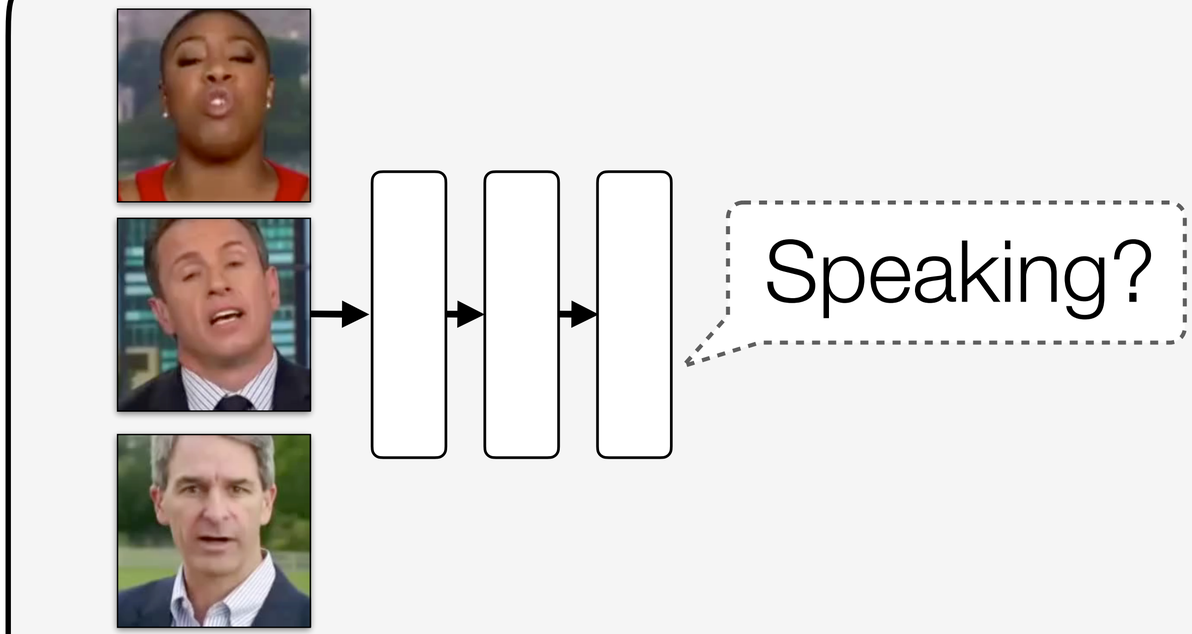
Detect objects

Run network on each one

Example: Active speaker detection [Chung & Zisserman 2016]

Group-and-Process Pipeline

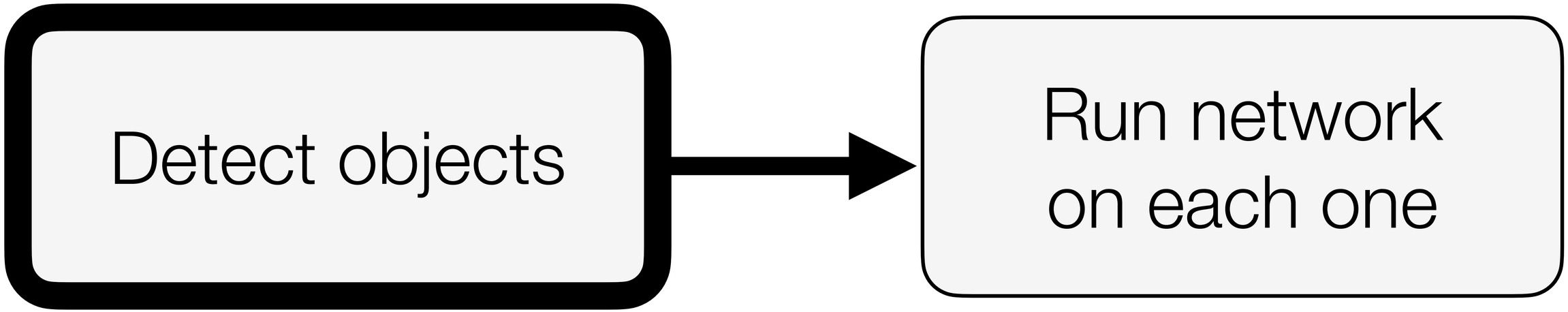
Detect objects



Run network on each one

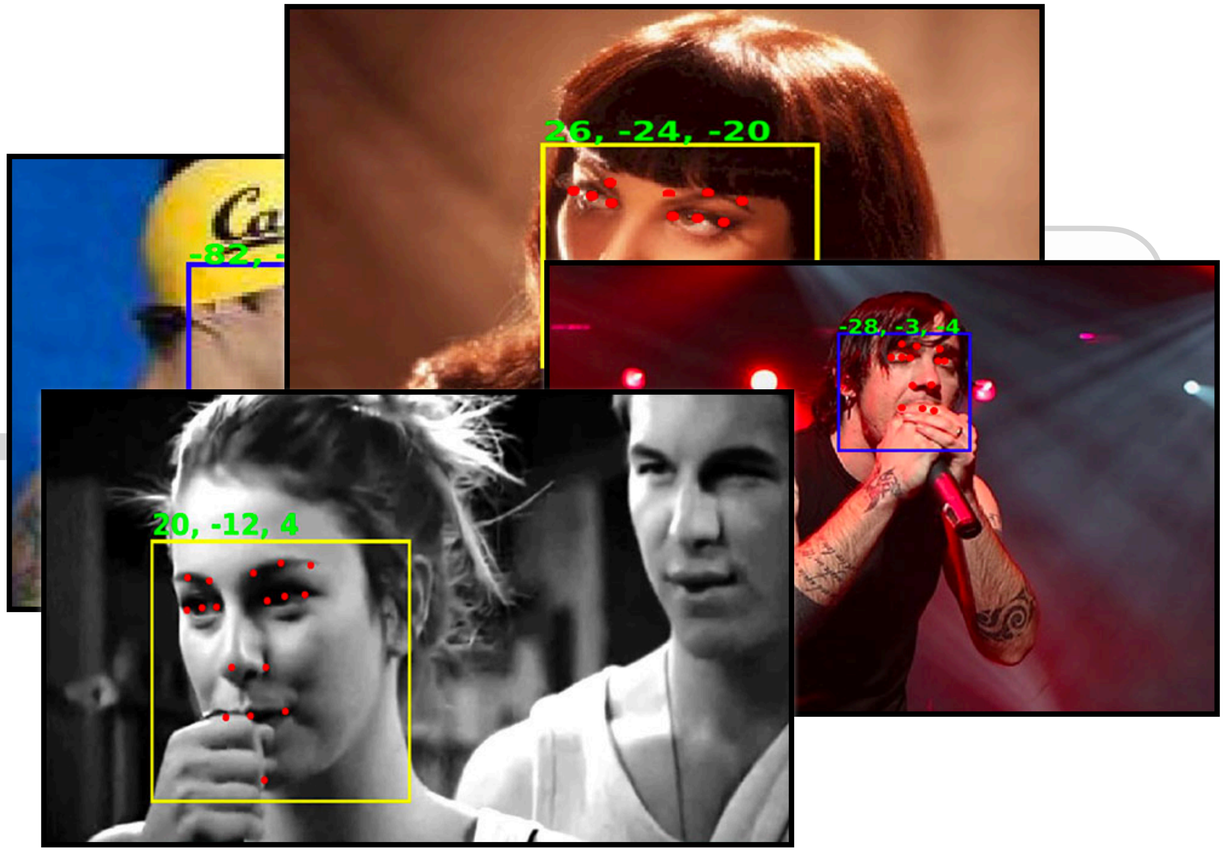
Example: Active speaker detection [Chung & Zisserman 2016]

Limitations of supervised grouping



Limitations of supervised grouping

Detect objects



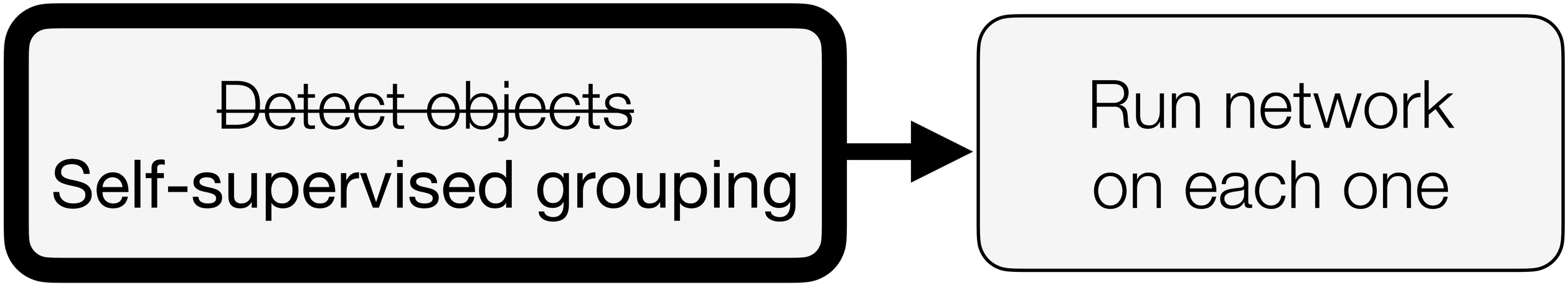
Requires lots of annotation

Limitations of supervised grouping

~~Detect objects~~

Difficult to adapt

What we want instead



Perceptual grouping is multimodal

Balls appear to bounce

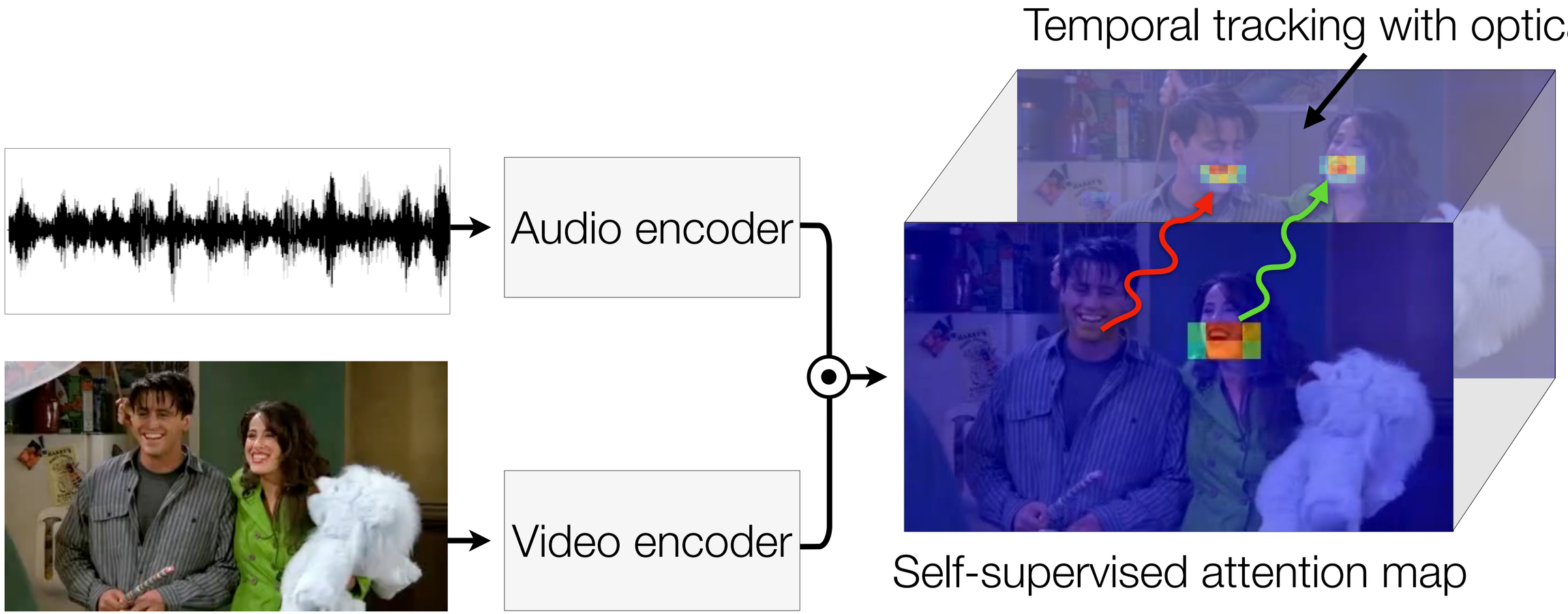
Motion-bounce illusion [Sekuler et al. 1997]

Perceptual grouping is multimodal

Balls appear to move through each other

Motion-bounce illusion [Sekuler et al. 1997]

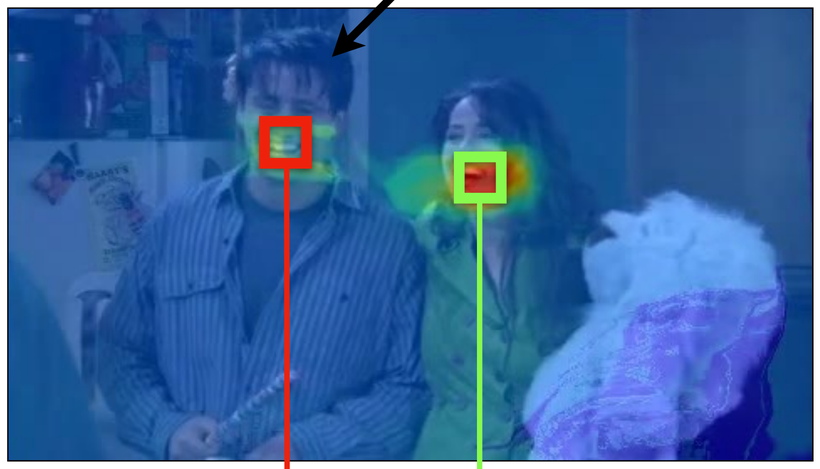
Grouping audio-visual objects



Grouping audio-visual objects

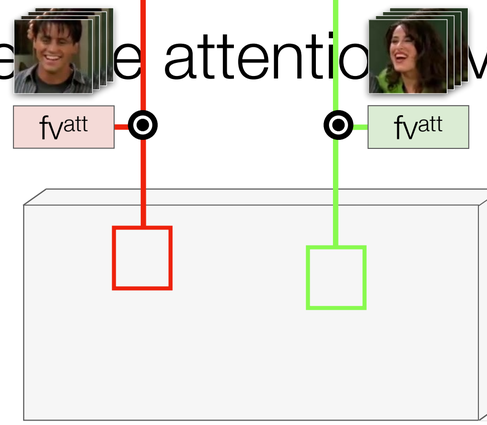
Self-supervised attention map

Find peaks + non-max suppression



Aggregate attention over time

Object embeddings

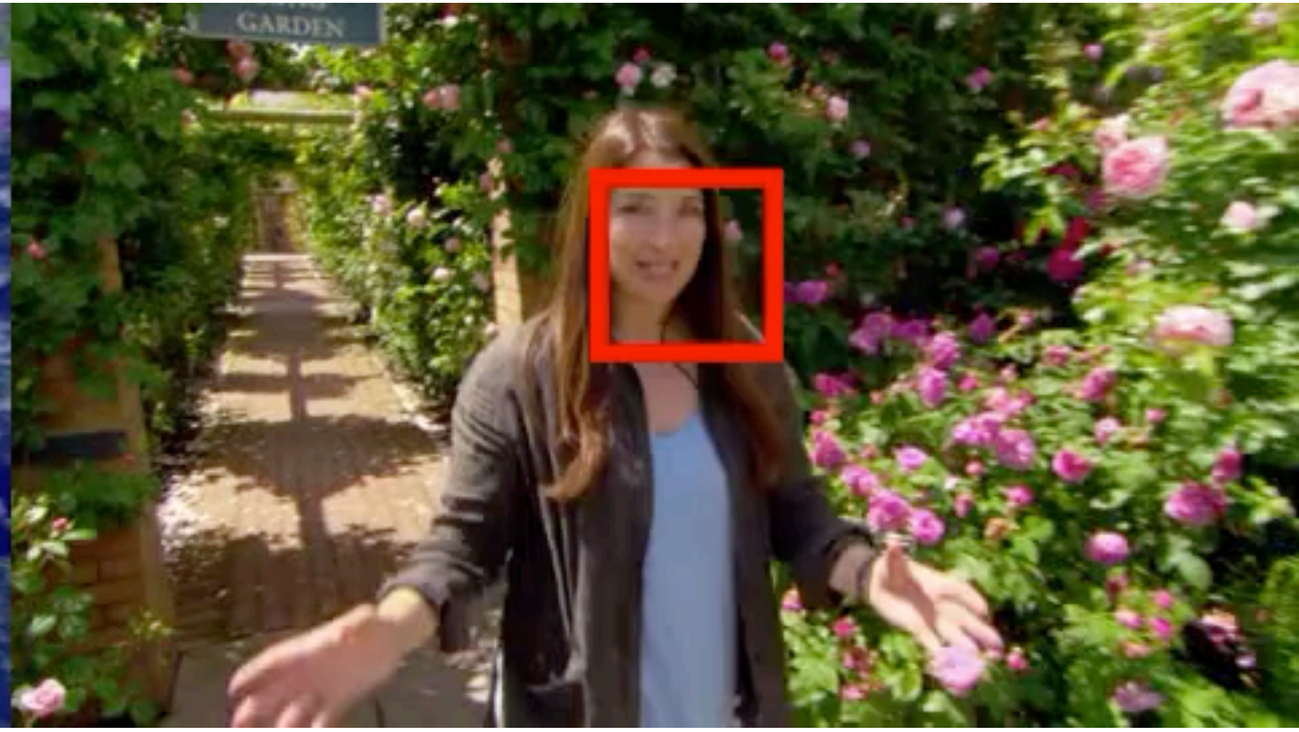


Video CNN

Audio-Visual Objects

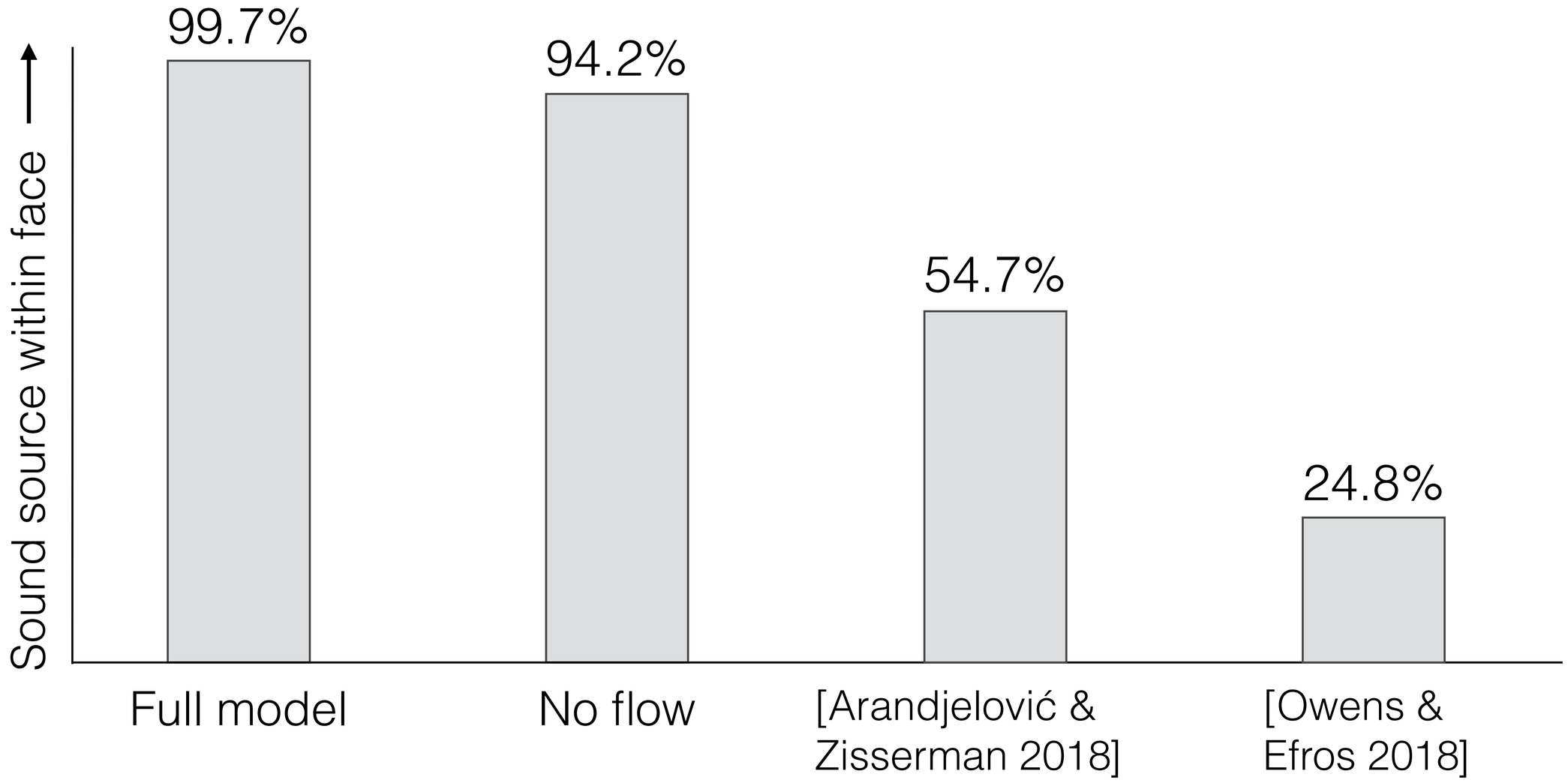
Examples from the LRS2 dataset

S_{AV} attention map

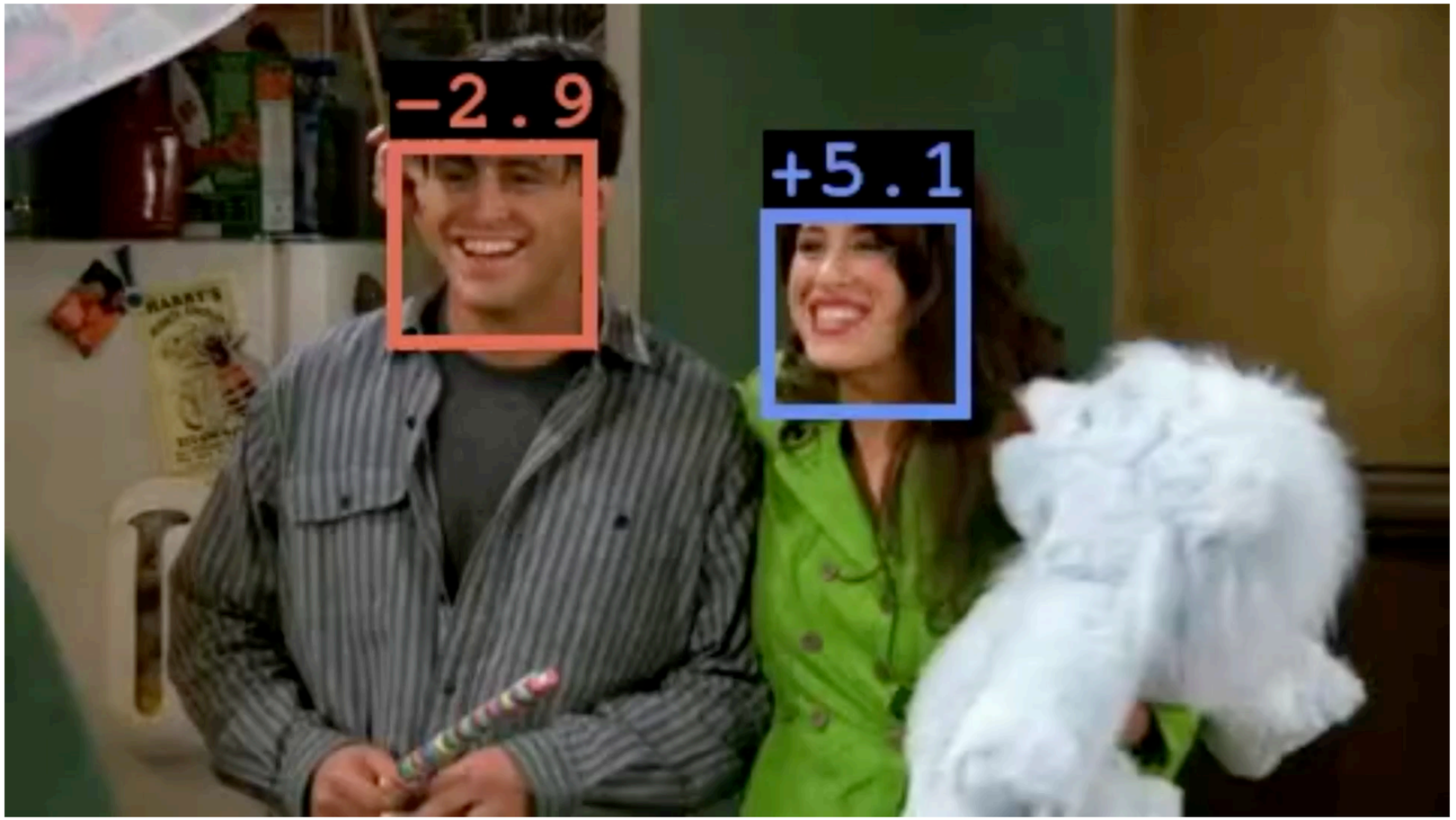


Audio-visual object

Sound source localization on LRS3 dataset

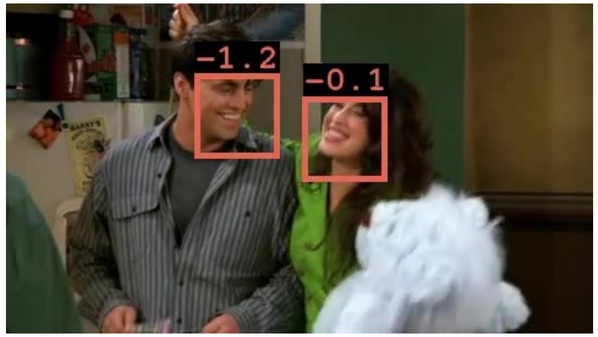


Downstream applications

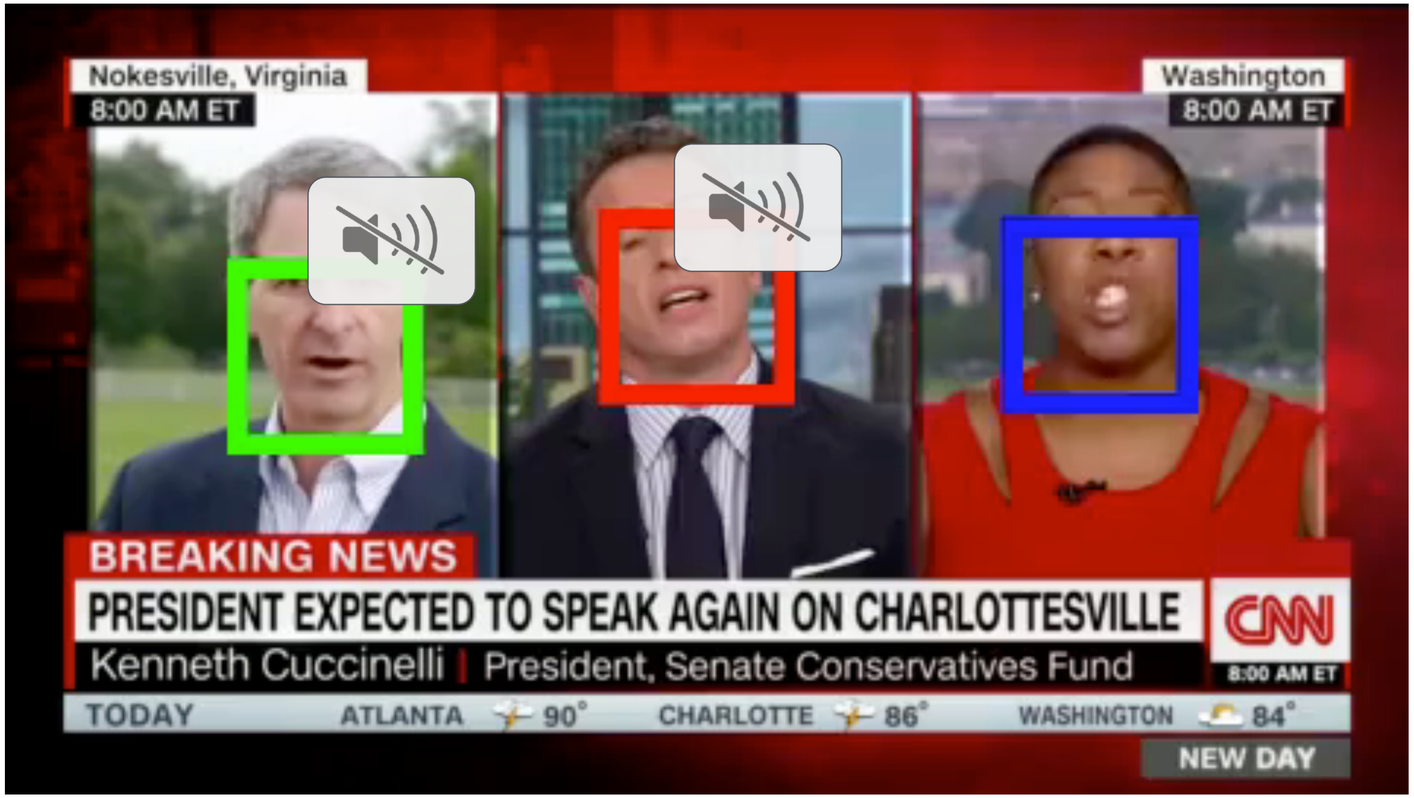


Active Speaker
Detection

Downstream applications

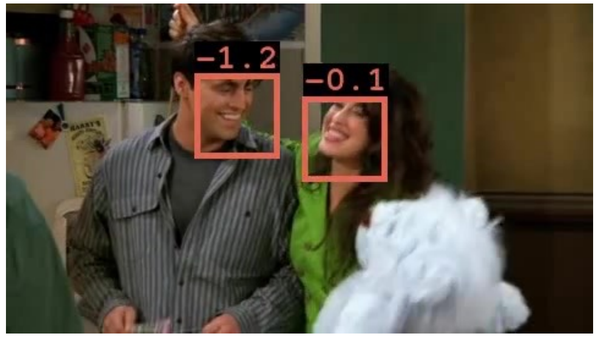


Active Speaker Detection



Multi-speaker source separation

Downstream applications



Active Speaker Detection

Multi-speaker source separation

Correcting temporal misalignment

Adapting to new domains

Sesame Street

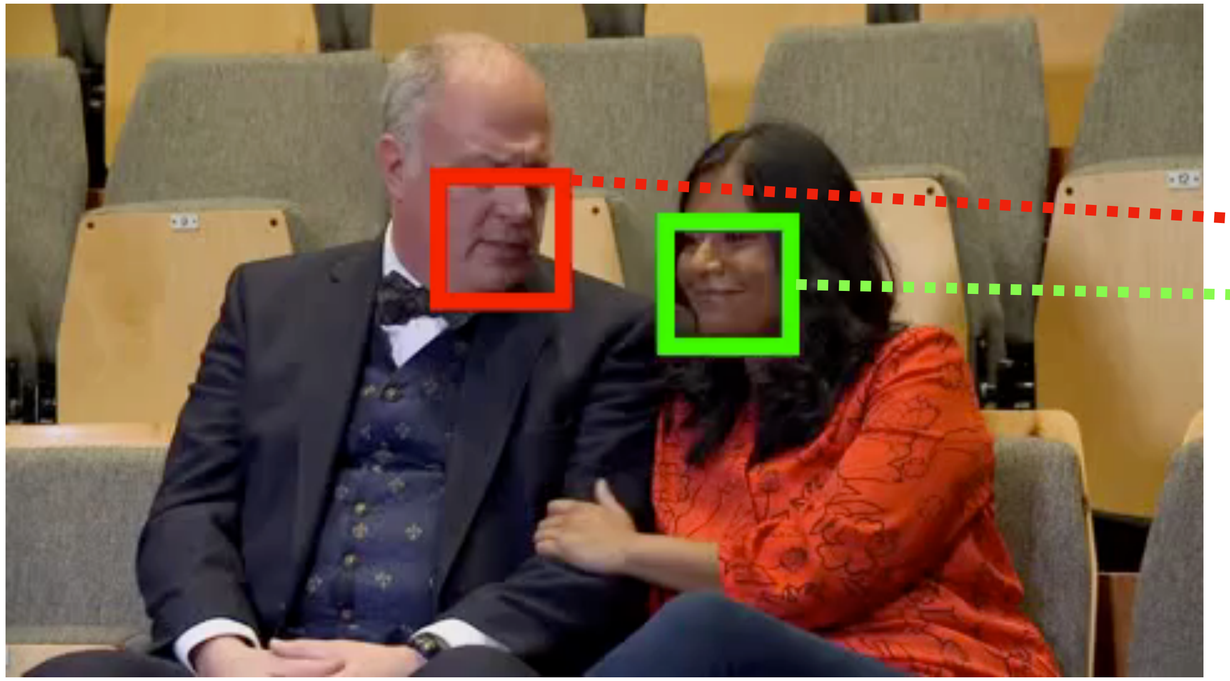
The Simpsons

Since everything's self-supervised, just finetune!

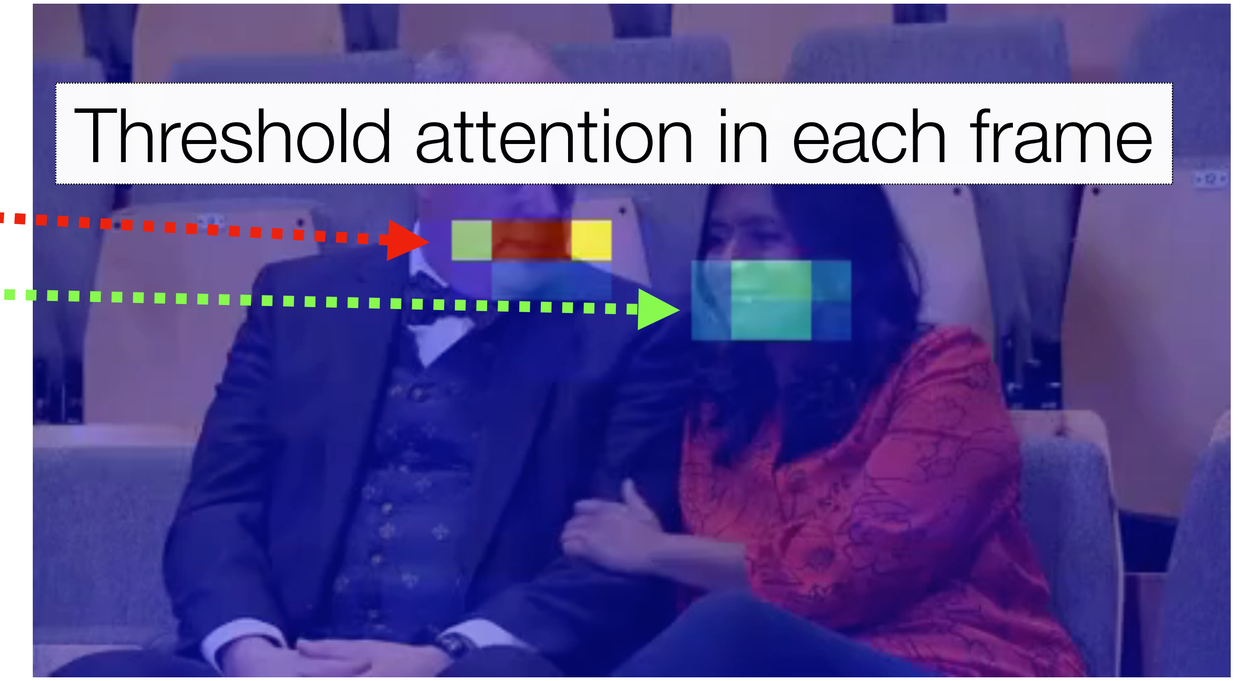
- One difference: use random audio-visual pairs as negatives [Arandjelović and Zisserman, 2017]

Downstream applications

Active Speaker Detection



Audio-visual objects



Attention map

- Displaying hand-chosen number of audio-visual objects

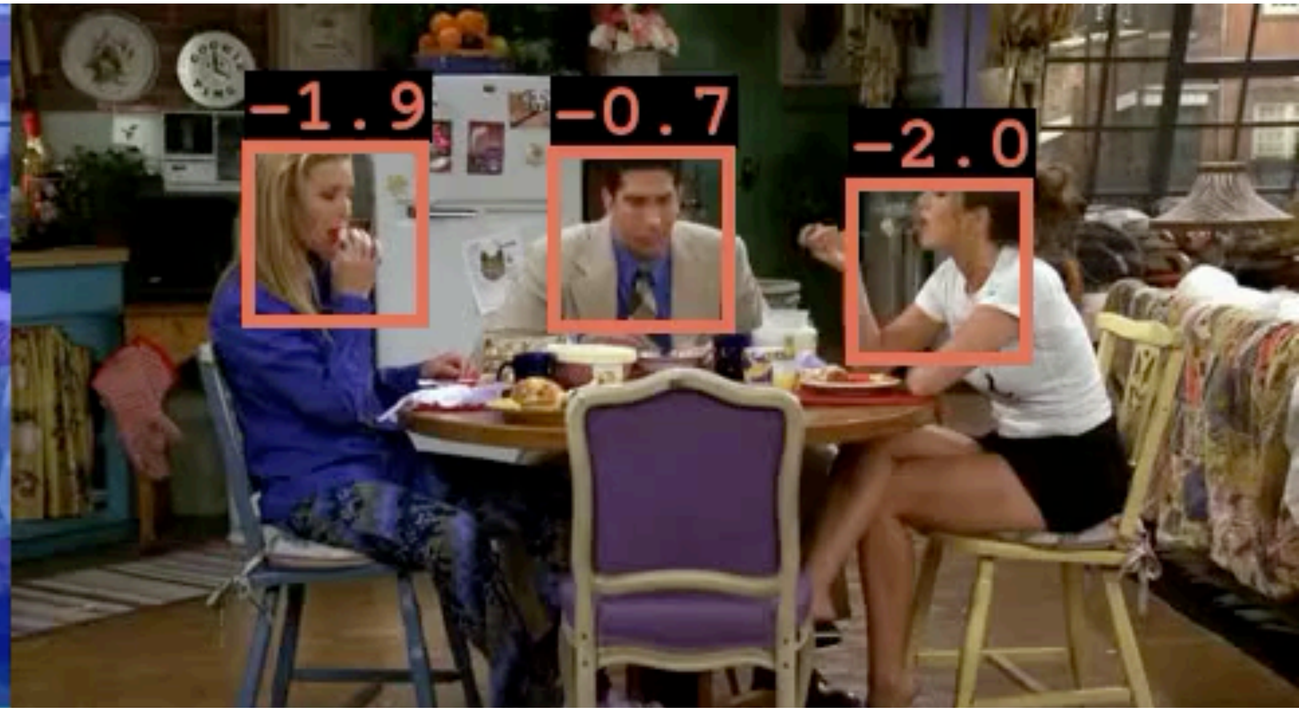
Active Speaker Detection

Examples from the Columbia benchmark dataset

Blue = active speaker
Red = inactive speaker

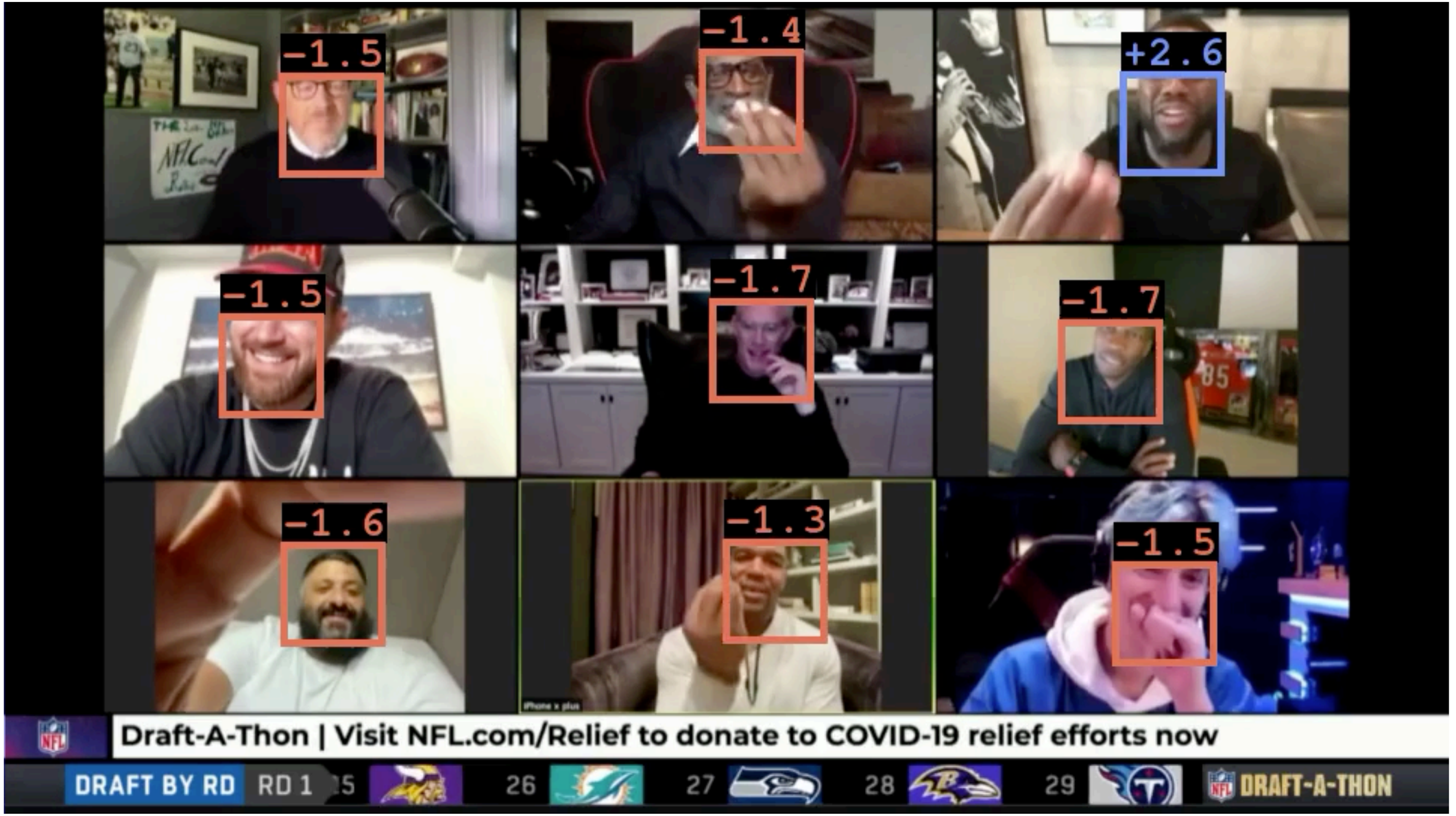
Active Speaker Detection

Examples from the *Friends* series



Blue = active speaker
Red = inactive speaker

Active Speaker Detection



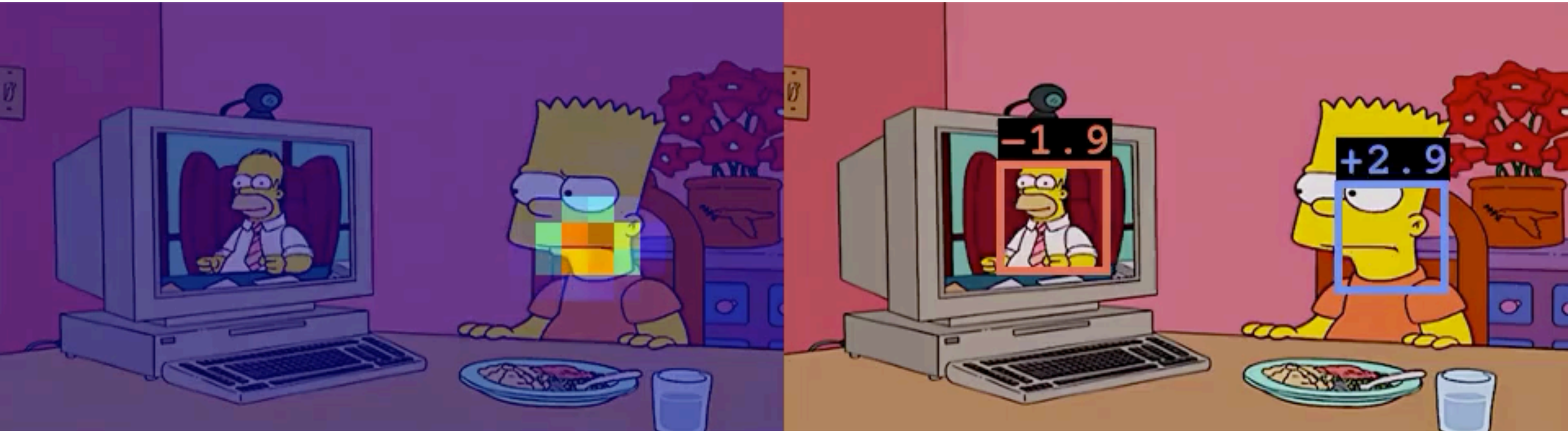
Active Speaker Detection

Examples from *Sesame Street*

Blue = active speaker
Red = inactive speaker

Active Speaker Detection

Examples from *The Simpsons*

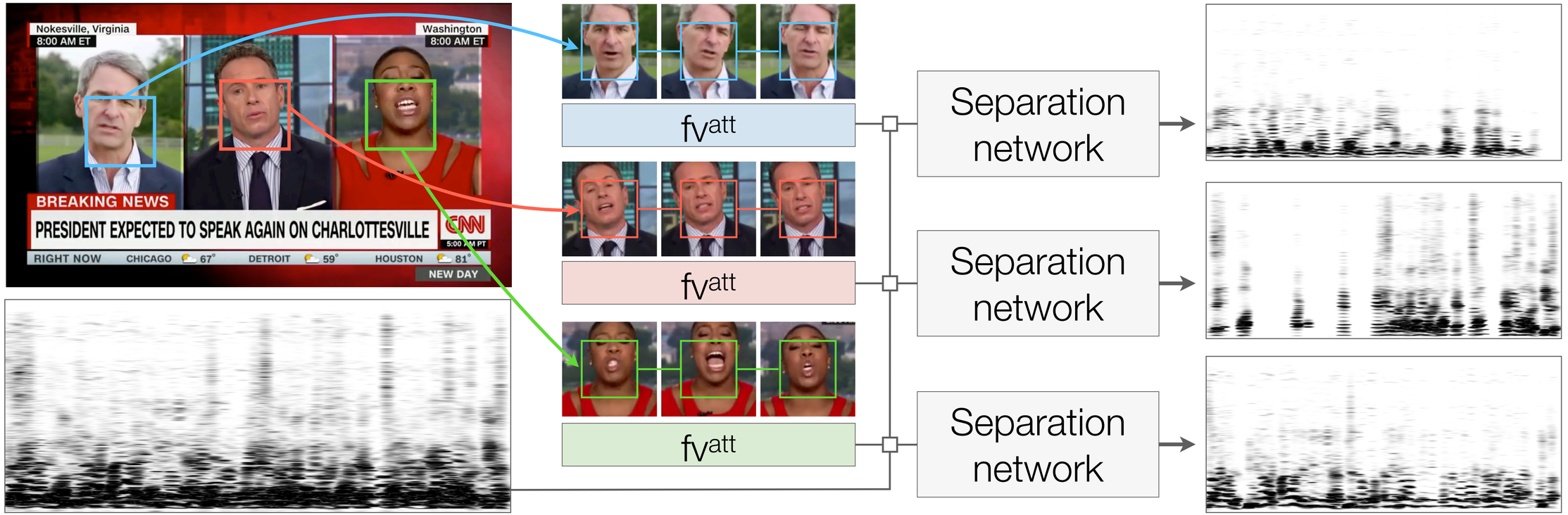


Blue = active speaker

Red = inactive speaker

Multi-speaker Source Separation

Multi-speaker Source Separation



Multi-speaker Source Separation

S_{AV} attention map

Detected
AV objects

Separated voices:
AV object 1

S_{AV} attention map

Detected
AV objects

Separated voices:
AV object 2

S_{AV} attention map

Detected
AV objects

Separated voices:
AV object 3

S_{AV} attention map

Detected
AV objects

Separated voices:
AV object 1

Multi-speaker Source Separation

S_{AV} attention map

Detected
AV objects

Original clip

Multi-speaker Source Separation

S_{AV} attention map

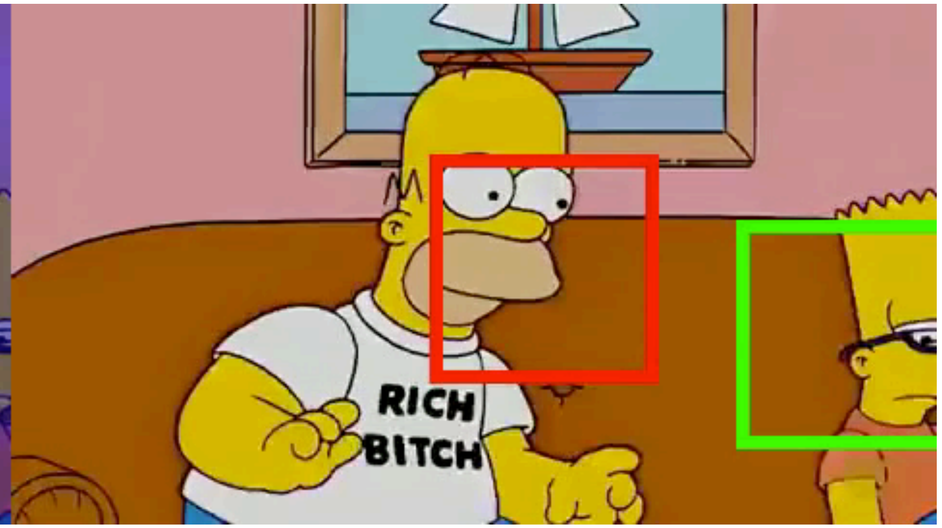
Detected
AV objects

Separated voices:
AV object 1

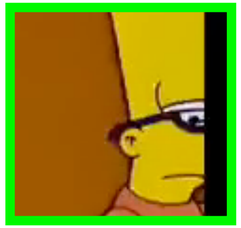
Multi-speaker Source Separation

S_{AV} attention map

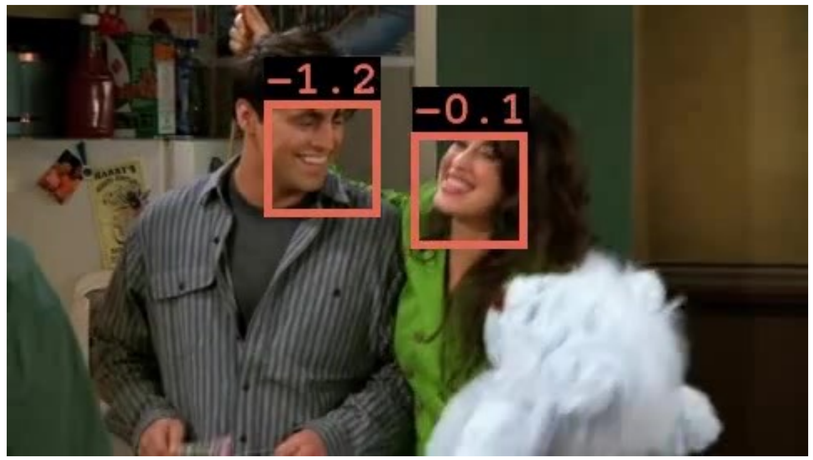
Detected
AV objects



Separated voices:
AV object 2



Applications of audio-visual objects



Active speaker detection

Multi-speaker source separation

Correcting temporal misalignment

Discussion

- The method only detects points, e.g. the mouth. Why is this?
- Is the problem well-defined?
- How dependent is the method on motion?

Labelling unlabelled videos from scratch with multi-modal self-supervision

Yuki M. Asano*, Mandela Patrick*, Christian Rupprecht, Andrea Vedaldi

NeurIPS 2020

Goal: cluster videos without supervision.

Self-labeling

Labeling function

Feature function

$$y(\Phi(x))$$

Self-labeling

Solve for $y_i = y(x_i)$

$$E(p|y_1, \ldots, y_N) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i).$$

Trivial solution! Give everything the same label.

Need more constraints.

Self-labeling

Jointly solve for $y_i = y(x_i)$ and features x_i

$$E(p|y_1, \ldots, y_N) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i).$$

Trivial solution! Give everything the same label.

Need more constraints.

[Asano et al. “SeLa”, 2020]

Self-labeling

Extra assumption: each label is equally likely.

$$E(p, q) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{K} q(y|\boldsymbol{x}_i) \log p(y|\boldsymbol{x}_i)$$

$$\min_{p,q} E(p, q) \quad \textbf{subject to} \quad \forall y : q(y|\boldsymbol{x}_i) \in \{0, 1\} \textbf{ and } \sum_{i=1}^{N} q(y|\boldsymbol{x}_i) = \frac{N}{K}$$

- This is an optimal transport problem that can be solved (somewhat) efficiently using linear programming.
- In practice, solve an approximation via Sinkhorn-Knopp by adding a regularization term.

[Asano et al. “SeLa”, 2020]

Learning

1. **Representation learning:** update the features given the labels y. This is doing using SGD with a cross-entropy loss.
2. **Self-labeling:** update labels y using Sinkhorn-Knopp clustering.

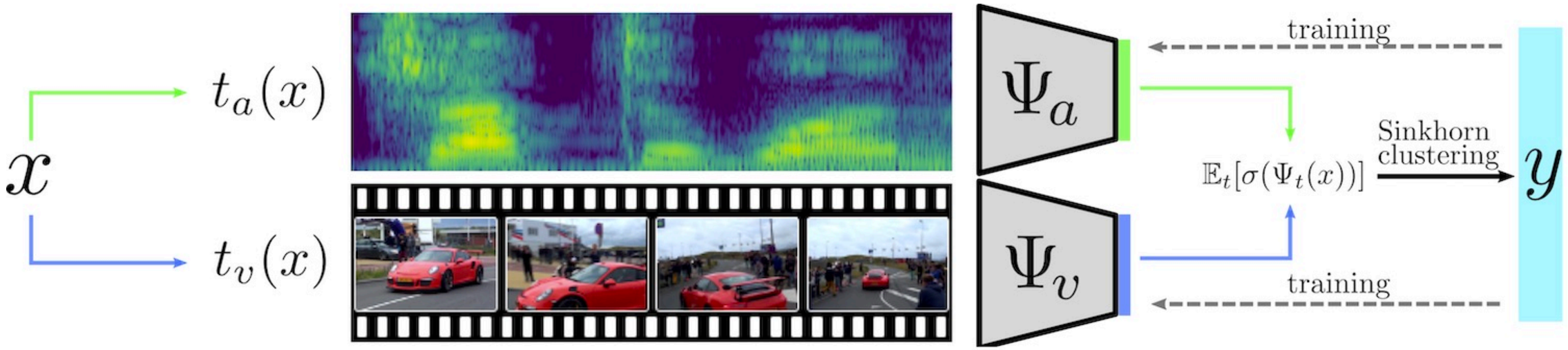
- Quite similar to other clustering methods, like DeepCluster [Caron et al., 2018], but there's a well-defined objective and less susceptible to degeneracies. Both steps optimize the same objective.

[Asano et al. "SeLa", 2020]

Extensions

- Do data augmentation and force examples to belong to the same cluster
  - Similar idea in contrastive learning
  - In practice, just augment the data during the representation learning step
- Use multiple clusterings with one shared representation.

Deep multimodal clustering



"SeLaVi"

Multimodal clustering

- Drop out modalities
  - Treat as special type of *data augmentation*
  - Either use image or audio (but not both)
  - Plus do regular data augmentation, too
- Use special initialization, since both modalities need same label.
- Another trick: clusters shouldn’t be the same size! Allow arbitrary distributions, e.g. Zipf

Results



Cluster #0

For more, please see: https://www.robots.ox.ac.uk/~vgg/research/selavi/

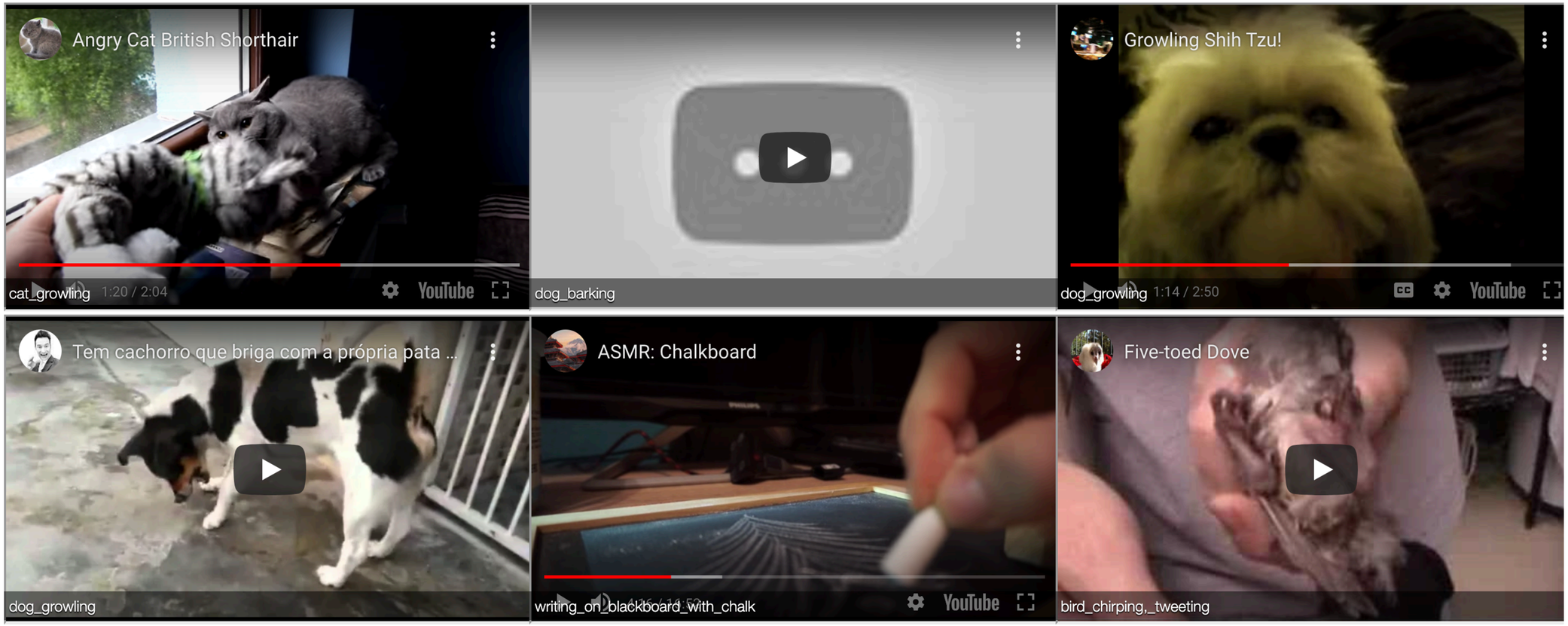
Results

Cluster #188

Results

VGG-Sound Cluster 68 (entropy: 5.020)

+ SHOW CLASSES



Cluster #68

Do clusters = classes?

Method	🔊	📹	NMI	ARI	Acc.	⟨H⟩	$\langle p_{max} \rangle$
Random	✗	✓	10.2	4.0	2.2	4.9	3.5
Supervised	✗	✓	46.5	15.6	24.3	2.9	30.8
DPC	✗	✓	15.4	0.7	3.2	4.7	4.9
MIL-NCE	✗	✓	48.5	12.5	22.0	2.6	32.9
XDC	✗	✓	16.7	1.0	3.9	4.5	6.4
	✓	✗	14.0	0.8	2.9	4.6	4.4
	✓	✓	18.1	1.2	4.5	4.41	7.4
SeLaVi	✗	✓	52.8	19.7	30.1	2.6	35.6
	✓	✗	47.5	15.2	26.5	2.8	32.9
	✓	✓	**56.7**	**22.5**	**32.3**	**2.4**	**38.0**

Cluster accuracy on VGGSound (K = 310, I think?)

Representation learning

Method	Architecture	Pretrain Dataset	Top-1 Acc%	
			UCF	HMDB
Full supervision [2]	R(2+1)D-18	ImageNet	82.8	46.7
Full supervision [2]	R(2+1)D-18	Kinetics-400	**93.1**	**63.6**
Weak supervision, CPD [49]†	3D-Resnet50	Kinetics-400	88.7	57.7
MotionPred [73]	C3D	Kinetics-400	61.2	33.4
RotNet3D [39]	3D-ResNet18	Kinetics-600	62.9	33.7
ST-Puzzle [42]	3D-ResNet18	Kinetics-400	65.8	33.7
ClipOrder [78]	R(2+1)D-18	Kinetics-400	72.4	30.9
DPC [30]	3D-ResNet34	Kinetics-400	75.7	35.7
CBT [66]	S3D	Kinetics-600	79.5	44.6
Multisensory [58]	3D-ResNet18	Kinetics-400	82.1	-
XDC [2]	R(2+1)D-18	Kinetics-400	84.2	47.1
AVTS [43]	MC3-18	Kinetics-400	85.8	56.9
AV Sync+RotNet [76]	AVSlowFast	Kinetics-400	87.0	54.6
GDT [60]	R(2+1)D-18	Kinetics-400	88.7	57.8
SeLaVi	R(2+1)D-18	Kinetics-400	83.1	47.1
SeLaVi	R(2+1)D-18	VGG-Sound	87.7	53.1

Discussion

- What information *doesn't* audio provide?
- What are the benefits of "low level" modalities vs. supervised data/language
- Could adding more modalities help?