CVPR LONG BEACH CALIFORNIA

June 16-20, 2019

Tutorial on Action Classification and Video Modeling

Self-Supervised Learning Using the Time Axis

Lorenzo Torresani

facebook research

DARTMOUTH

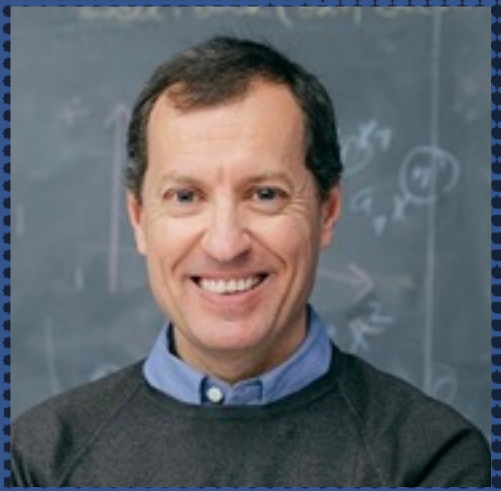
facebook
Artificial Intelligence Research

Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization

presented at NeurIPS 2018


Bruno Korbar


Du Tran

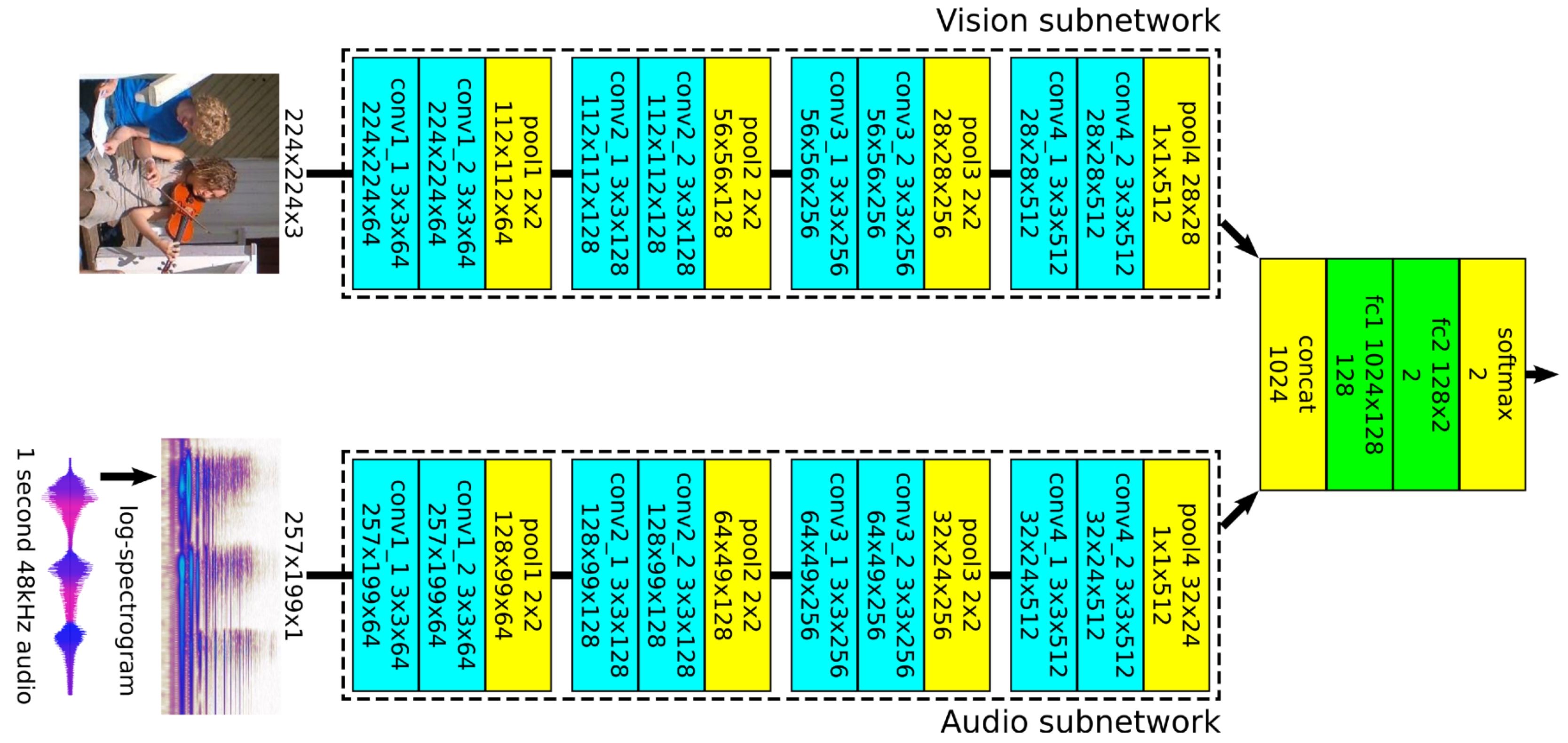

Lorenzo Torresani

Prior Work: Audio-Visual Correspondence

[Arandjelovic and Zisserman, ICCV 2017]:

Still frame

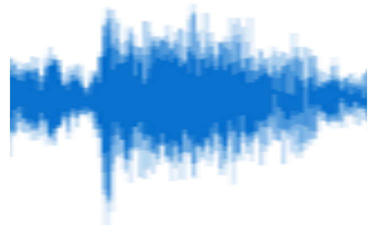
Audio Clip



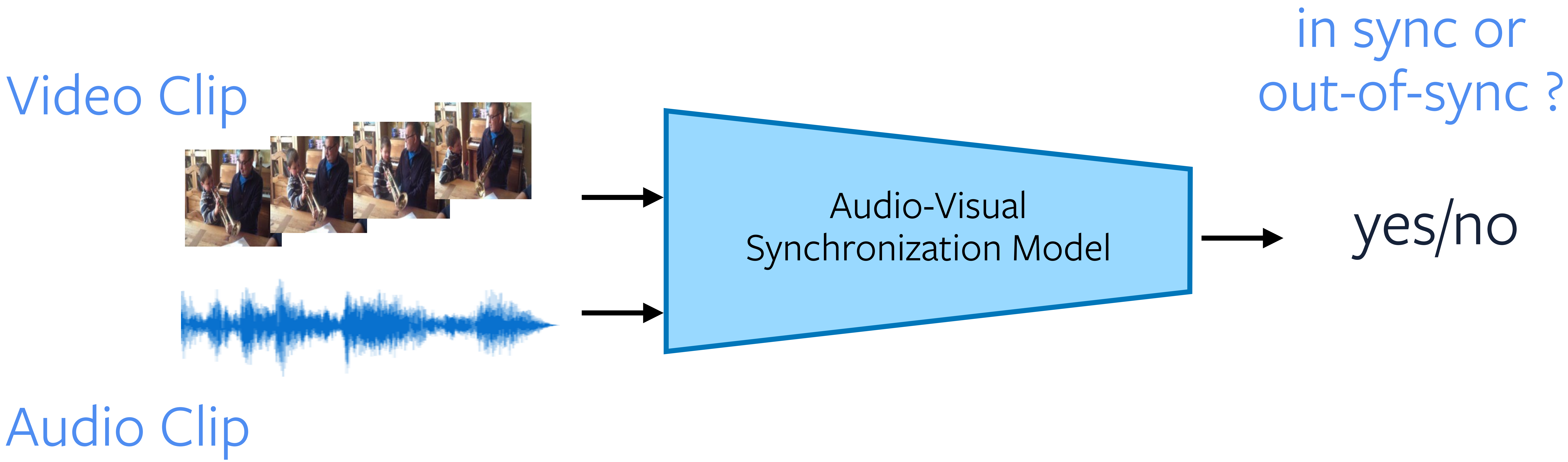
semantically matching?

yes/no

- Pretext task:
  - ✓ negative (frame, audio) pairs are sampled from different videos
  - ✓ network learns semantic correspondence

( , violin playing ) → yes

Learning Audio and Video Models from Self-Supervised Synchronization

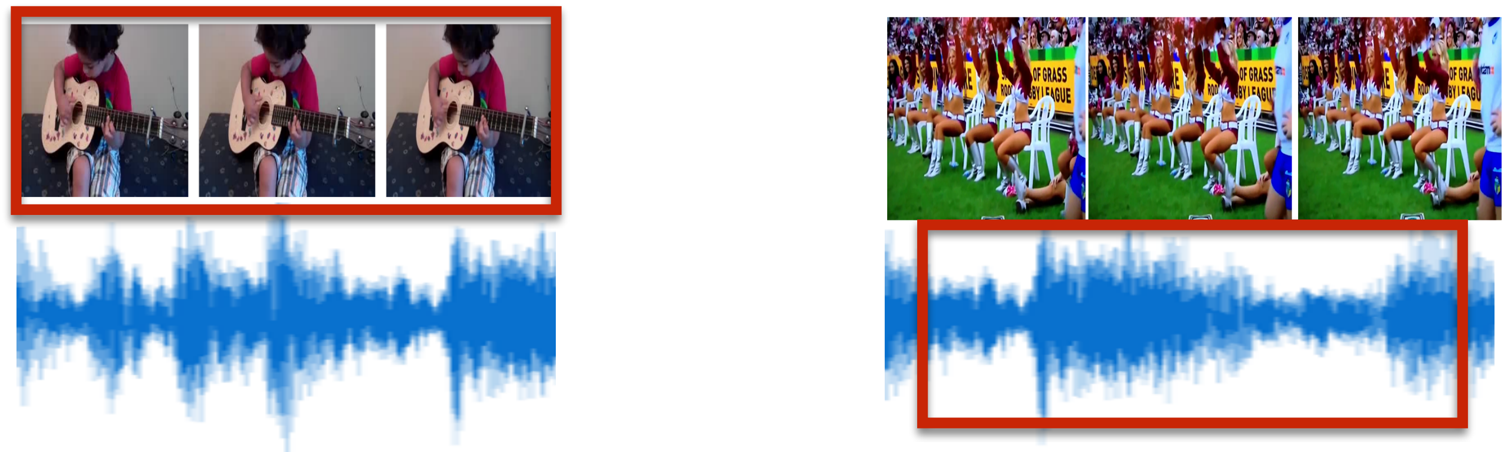


- Pretext task forces the network to learn temporal (sound/motion) representations useful for audio/video classification

Complexity of our pretext

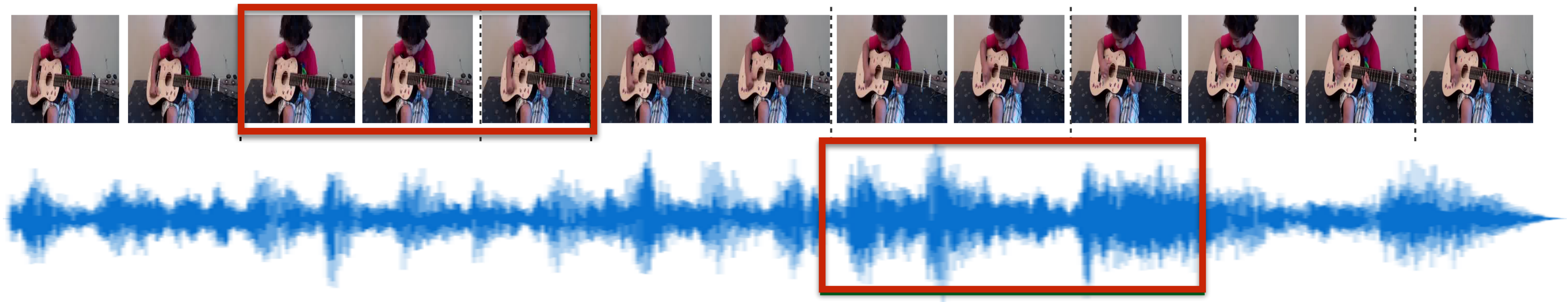
- Controlled by choice of negatives

✓ easy negatives: (video, audio) from distinct sequences



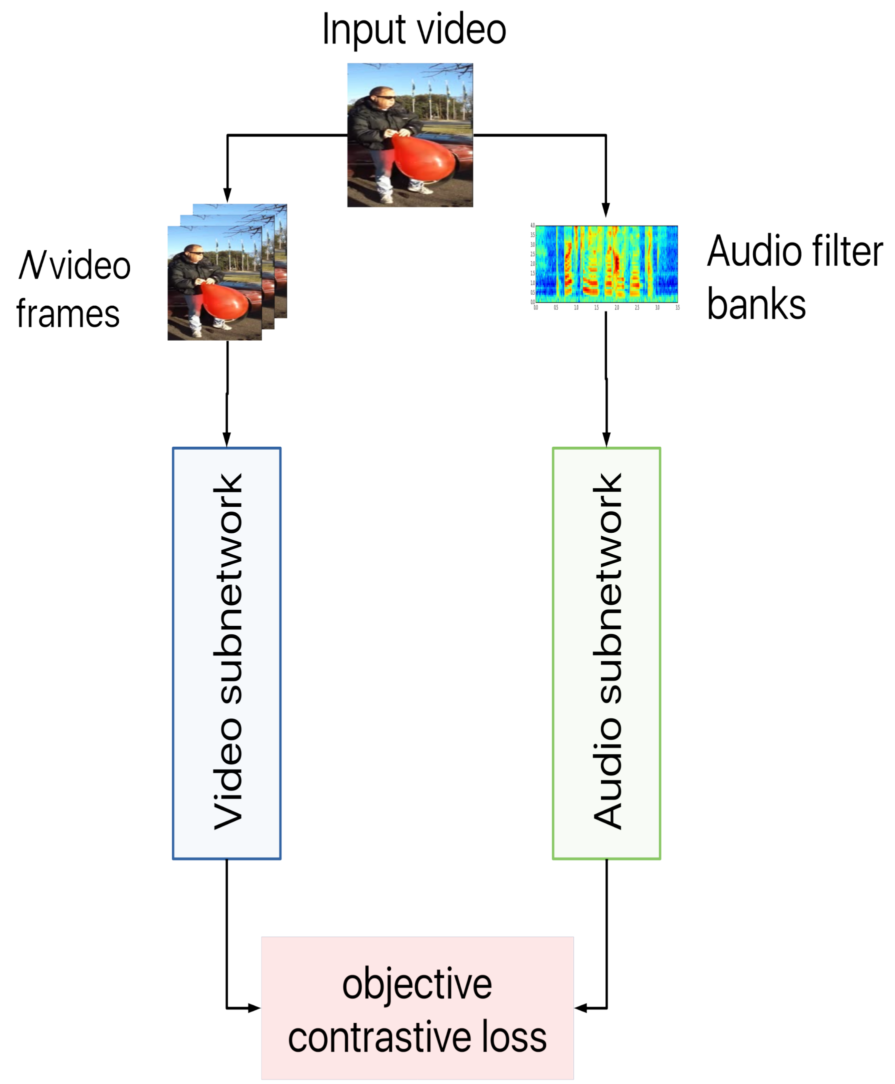
→ can be recognized from different semantics, temporal analysis is *not* needed

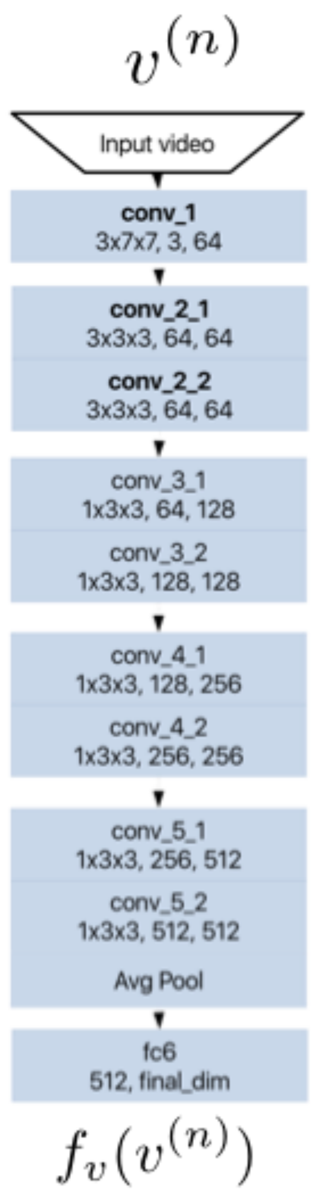
✓ hard negatives: (video, audio) sampled from same sequence but out-of-sync



→ force the learning of temporal features

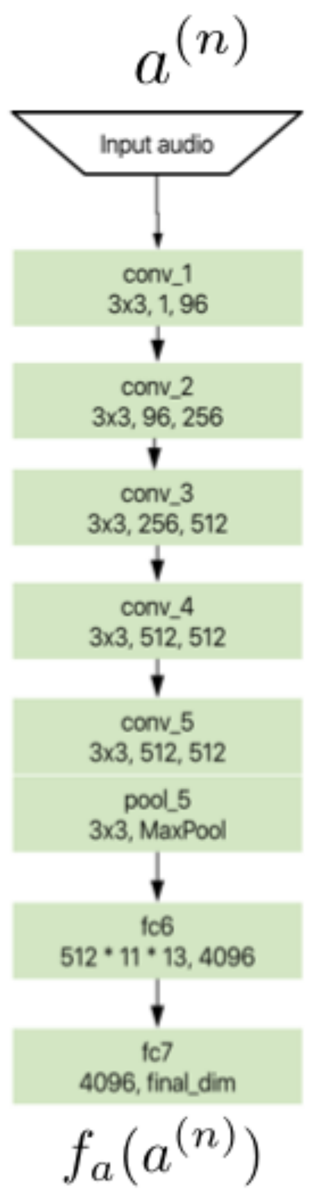
Architecture and learning objective





$v^{(n)}$ video clip of 1 sec

$f_v(\ )$ *MC3* 3D CNN [Tran et al., CVPR 18]



$a^{(n)}$ audio clip of 1 sec

$f_a(\ )$ *VGG* applied to MEL-spectrogram

- ✓ Contrastive loss:

$$E = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)})||f_v(v^{(n)}) - f_a(a^{(n)})||_2 + (1 - y^{(n)})\max(\eta - ||f_v(v^{(n)}) - f_a(a^{(n)})||_2, 0)^2$$

$$y^{(n)} = \begin{cases} 1: & \text{if the examples are in sync} \\ 0: & \text{otherwise} \end{cases}$$

Accuracy on pretext task (in-sync vs out-of-sync)

Evaluation on Kinetics dataset
(230K training videos, action labels are *not* used):

- ✓ training sets of varying difficulty (easy vs hard negatives)
- ✓ test set includes easy negatives only

Method	Negative type	Epochs	Accuracy (%)
Single learning stage	easy	1 - 90	69.0
	75% easy, 25% hard	1 - 90	58.9
	hard	1 - 90	52.3

Accuracy on pretext task (in-sync vs out-of-sync)

Evaluation on Kinetics dataset
(230K training videos, action labels are *not* used):

- ✓ training sets of varying difficulty (easy vs hard negatives)
- ✓ test set includes easy negatives only

Method	Negative type	Epochs	Accuracy (%)
Single learning stage	easy	1 - 90	69.0
	75% easy, 25% hard	1 - 90	58.9
	hard	1 - 90	52.3
	easy	1 - 50	67.2
Curriculum learning (i.e., second learning stage applied after a first stage of 1-50 epochs with easy negatives only)	75% easy, 25% hard	51 - 90	**78.4**
	hard	51 - 90	65.7

Curriculum learning yields better models even for downstream tasks

AVTS: Audio-Video Temporal Synchronization
[Korbar et al., NeurIPS 2018]

	pretext task	audio classification		video (action) classification	
Method	AVTS-Kinetics	ESC-50	DCASE	HMDB51	UCF101
Our AVTS - single stage	69.8	70.6	89.2	46.4	77.1
Our AVTS - curriculum	78.4	82.3	94.1	56.9	85.8
L^3-Net	74.3	79.3	93	40.2	72.3

Audio-Video Semantic Correspondence [Arandjelovic and Zisserman, ICCV 2017]

Audio-video synchronization as a pretraining scheme for action recognition

Video Network Architecture	Pretraining Dataset	Pretraining Supervision	UCF101	HMDB51
MC3	none	n/a	69.1	43.9
MC3	Kinetics	self-supervised	85.8	56.9
MC3	Audioset	self-supervised	89.0	61.6
MC3	Kinetics	fully supervised	90.5	66.8

almost 20% better than learning from scratch!

Audio-video synchronization as a pretraining scheme for action recognition

Video Network Architecture	Pretraining Dataset	Pretraining Supervision	UCF101	HMDB51
MC3	none	n/a	69.1	43.9
MC3	Kinetics	self-supervised	85.8	56.9
MC3	Audioset	self-supervised	89.0	61.6
MC3	Kinetics	fully supervised	90.5	66.8

nearly on-par with fully-supervised pretraining!

Audio classification with AVTS features

Method	Auxiliary dataset	Auxiliary supervision	# auxiliary examples	ESC-50 accuracy (%)	DCASE2014 accuracy (%)
Our audio subnet	none	none	none	61.6	72
SoundNet [2]	SoundNet	self	2M+	74.2	88
L^3-Net [1]	SoundNet	self	2M+	79.3	93
Our AVTS features	Kinetics	self	230K	76.7	91
Our AVTS features	AudioSet	self	1.8M	80.6	93
Our AVTS features	SoundNet	self	2M+	**82.3**	**94**

learning-from-scratch
vs
AVTS pretraining

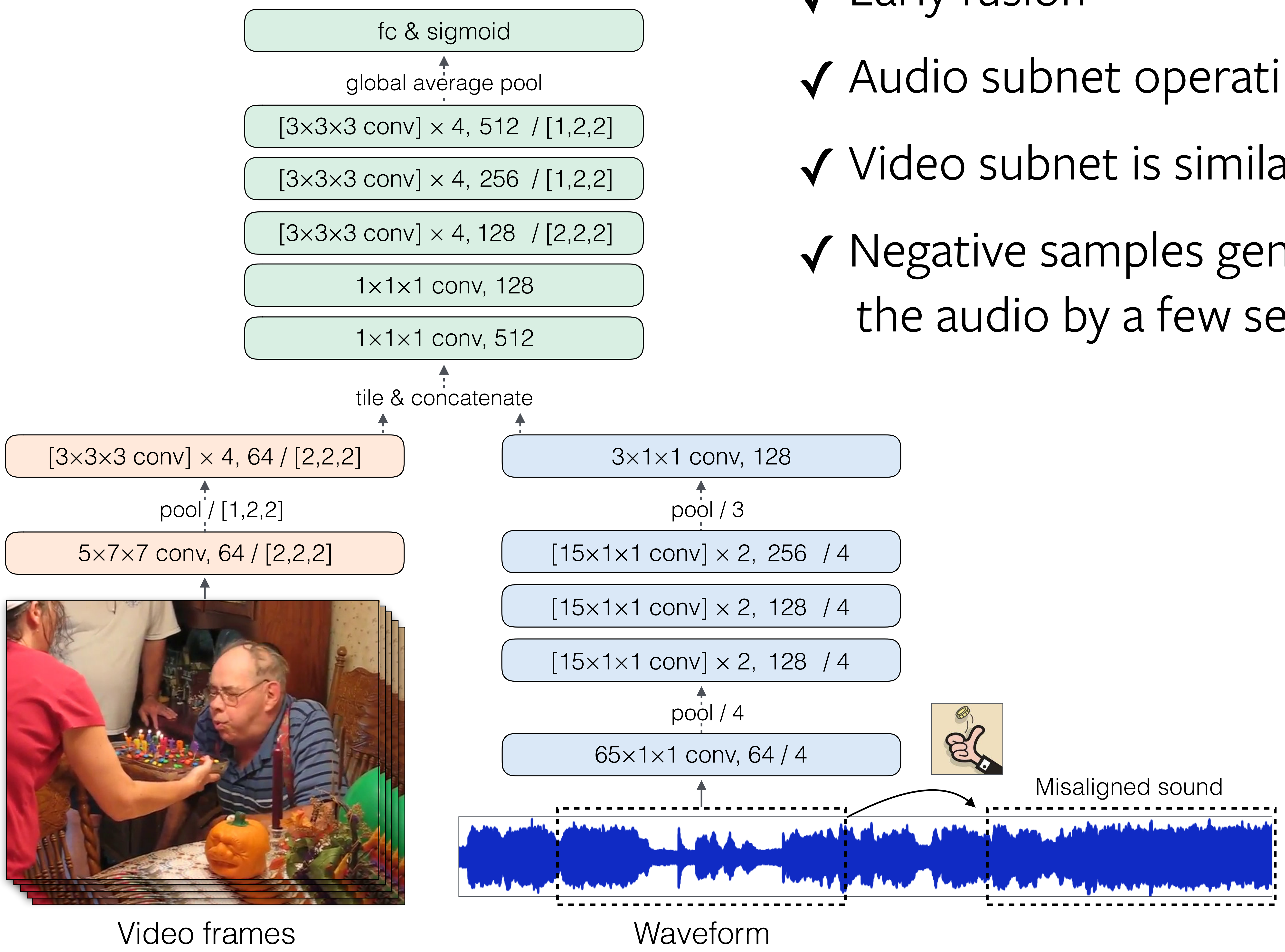
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

[Owens and Efros, ECCV 2018]

Concurrent work showing the use of self-supervised audio/video synchronization features for several applications:

(a) Sound localization (b) Action recognition (c) On/off-screen audio separation

Multisensory network design [Owens and Efros, ECCV 2018]



- ✓ Early fusion
- ✓ Audio subnet operating on raw waveform
- ✓ Video subnet is similar to ResNet3D-18
- ✓ Negative samples generated by shifting the audio by a few seconds

Action recognition by finetuning on UCF101

[Owens and Efros, ECCV 2018]

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%

Action recognition by finetuning on UCF101

[Owens and Efros, ECCV 2018]

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%

14% better than learning from scratch!

Action recognition by finetuning on UCF101

[Owens and Efros, ECCV 2018]

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%

audio increases accuracy by more than 5%

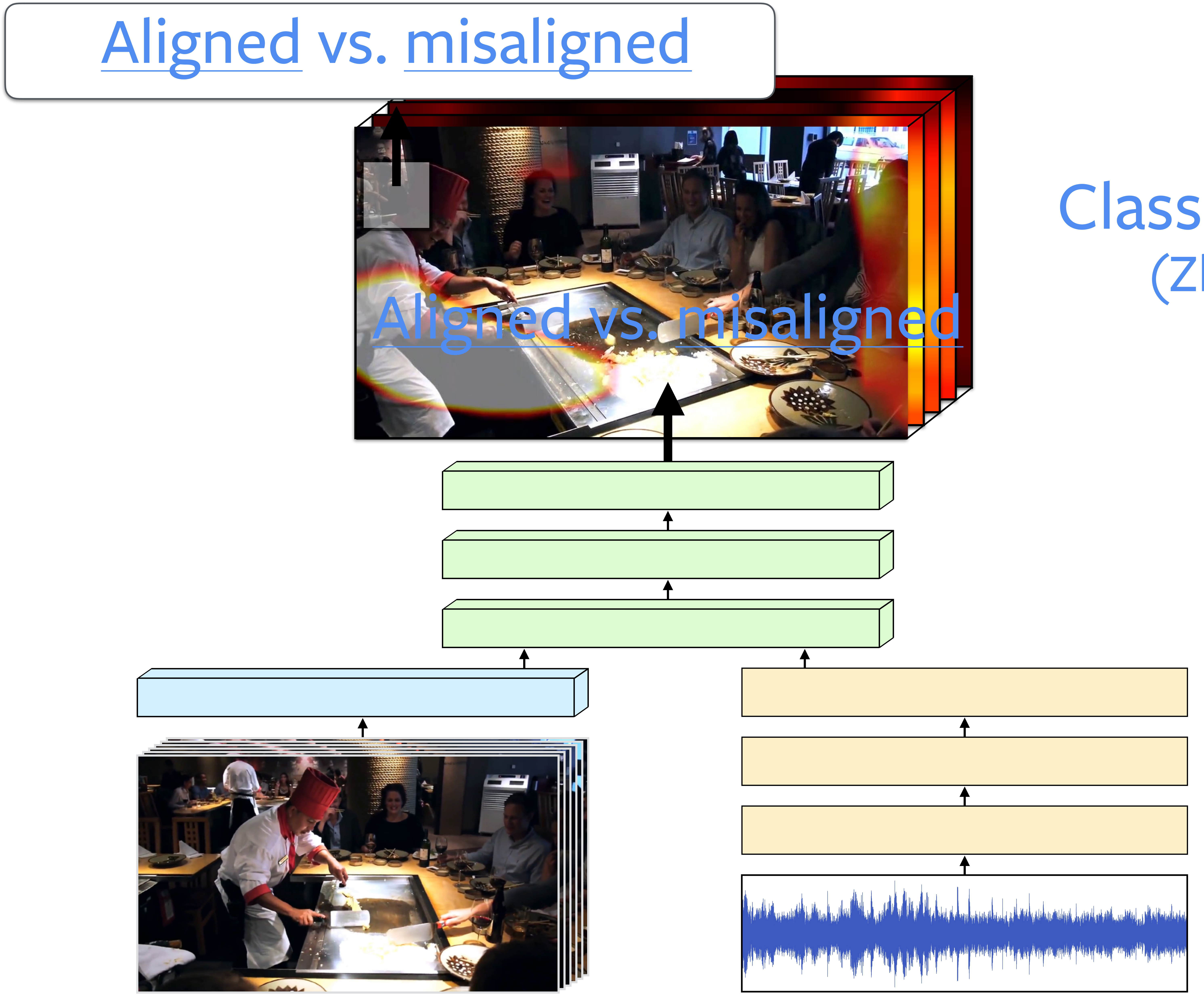
Action recognition by finetuning on UCF101

[Owens and Efros, ECCV 2018]

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%

comparison of objectives: synchronization vs semantic correspondence

What does the network learn?



Class activation map
(Zhou et al. 2016)

Slide credit: A. Owens

Slide credit: A. Owens

Top responses in test set

Top responses per category (speech examples omitted)

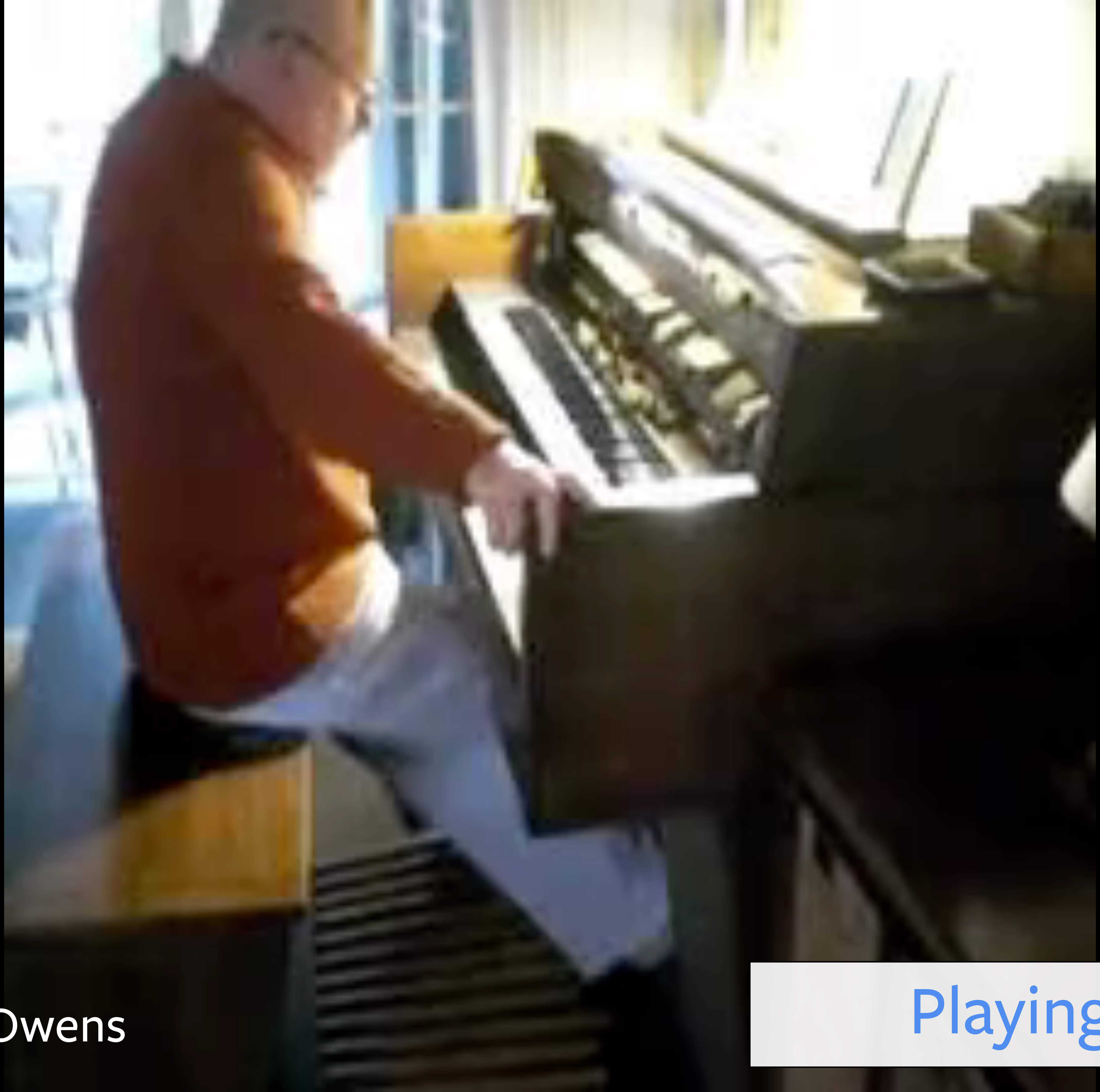
Slide credit: A. Owens

Slide credit: A. Owens

Dribbling basketball

Slide credit: A. Owens

Slide credit: A. Owens



Slide credit: A. Owens


Playing organ
Slide credit: A. Owens


Slide credit: A. Owens
Chopping wood

Slide credit: A. Owens

Chopping wood

Slide credit: A. Owens

Input video

Slide credit: A. Owens

On-screen prediction

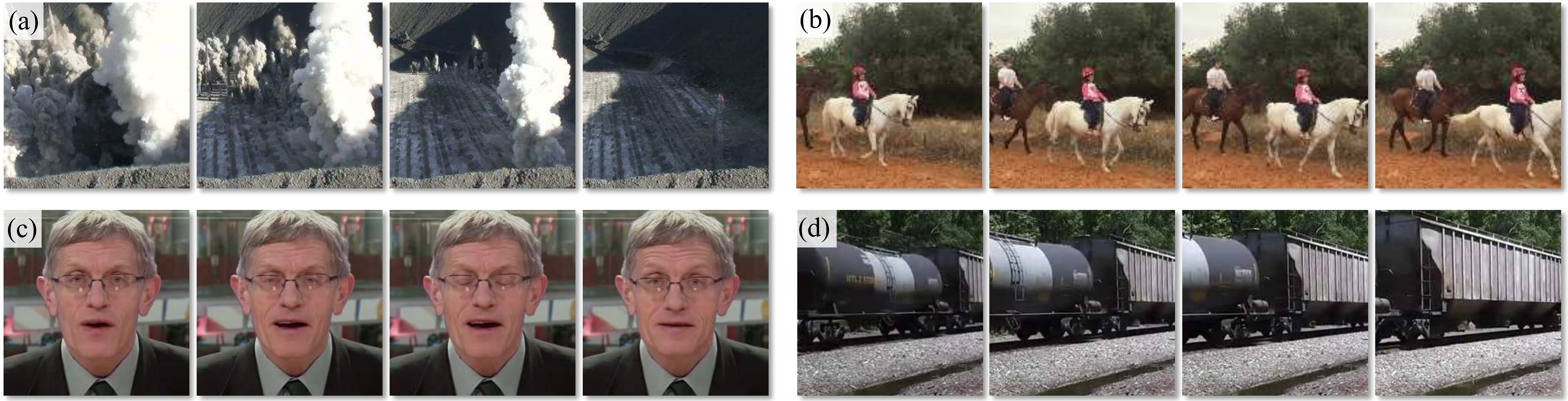
Slide credit: A. Owens

Off-screen prediction

Slide credit: A. Owens

Learning and using the arrow of time

[Wei, Lim, Zisserman and Freeman , CVPR 2018]



Can you tell from these ordered frames if the video is played forward or backward?

✓ Is it possible to train a "arrow of time" classifier from large-scale natural videos while avoiding artificial cues?

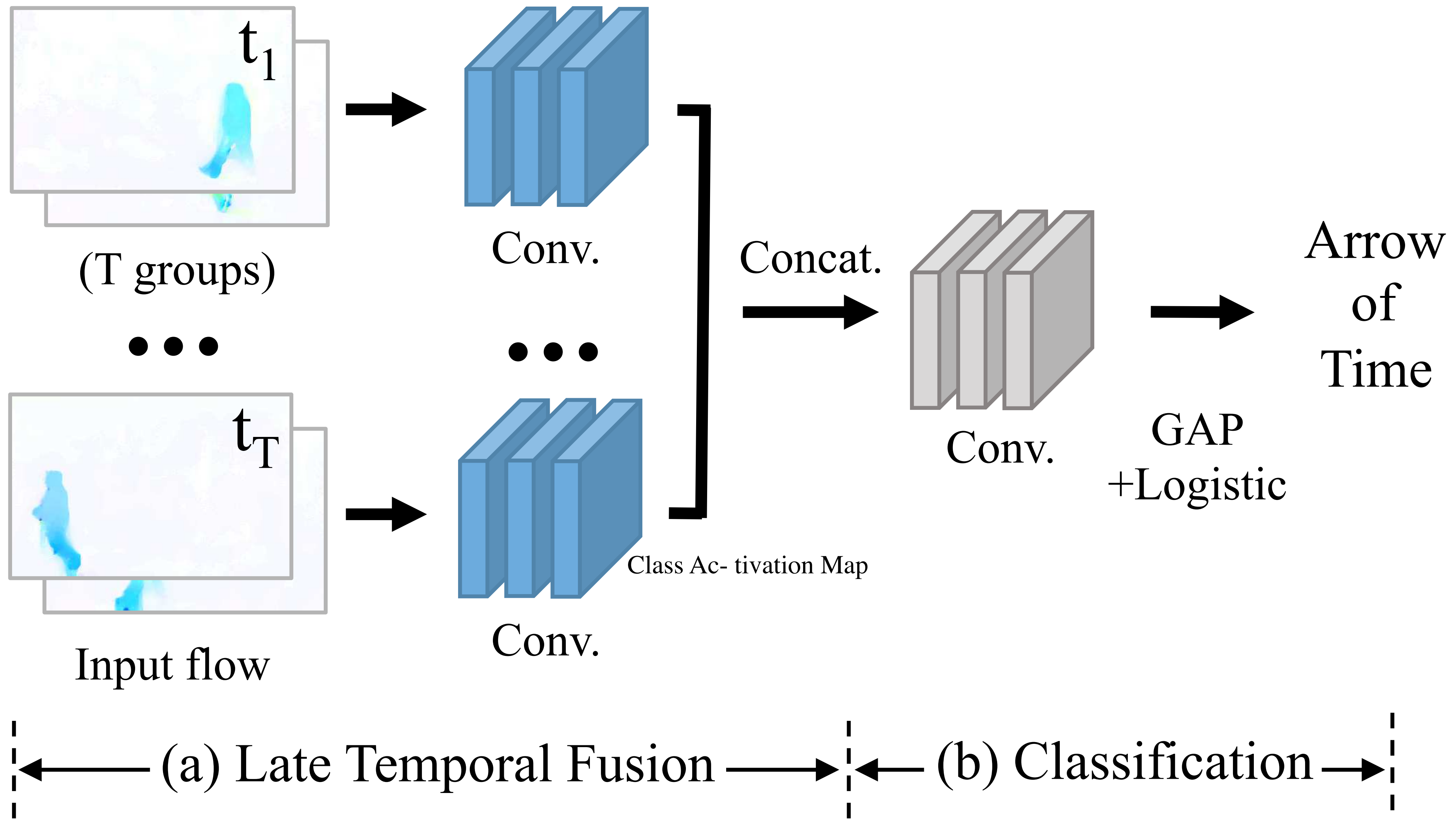
✓ What does the model learn about the visual world in order to solve this task?

✓ Is it possible to apply such learned commonsense knowledge to other video analysis tasks?

Forwards: (b), (c); backwards: (a), (d). Though in (d) the train can move in either direction

Design of an "arrow of time" classifier

[Wei, Lim, Zisserman and Freeman , CVPR 2018]



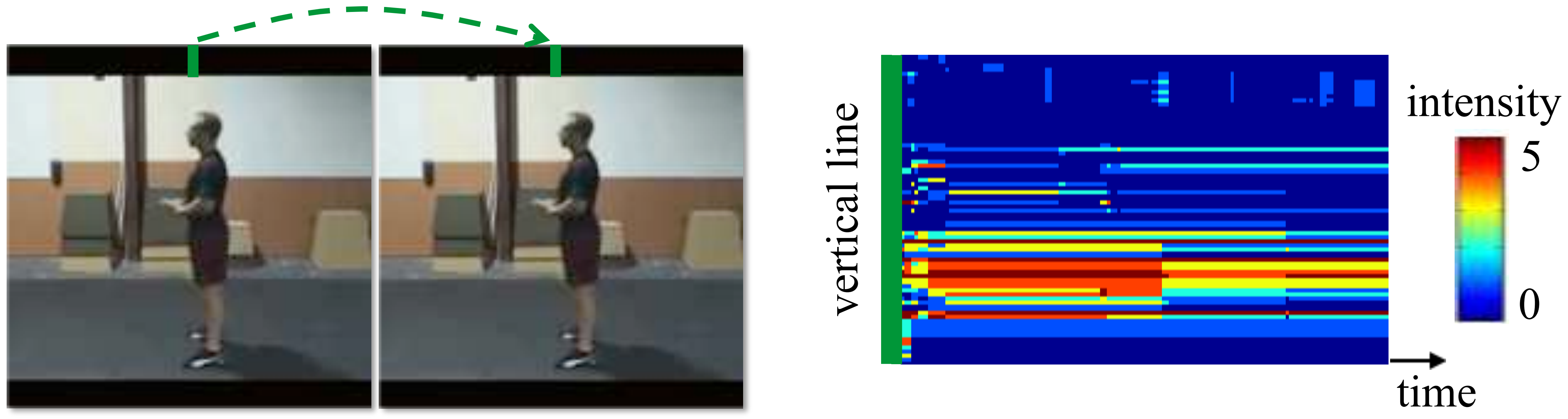
- ✓ Optical flow as input to focus on temporal aspects in the video
- ✓ Extended temporal span by concatenation of conv5 VGG features computed over T segments
- ✓ Global average pooling layer (GAP) for better activation localization via Class Activation Map (CAM) [Zhou et al., CVPR 2016]

Avoid "cheating"

[Wei, Lim, Zisserman and Freeman , CVPR 2018]

Deep networks can leverage artificial cues to solve the task

Black framing



Consistent camera motion

Tilt down

Zoom-in

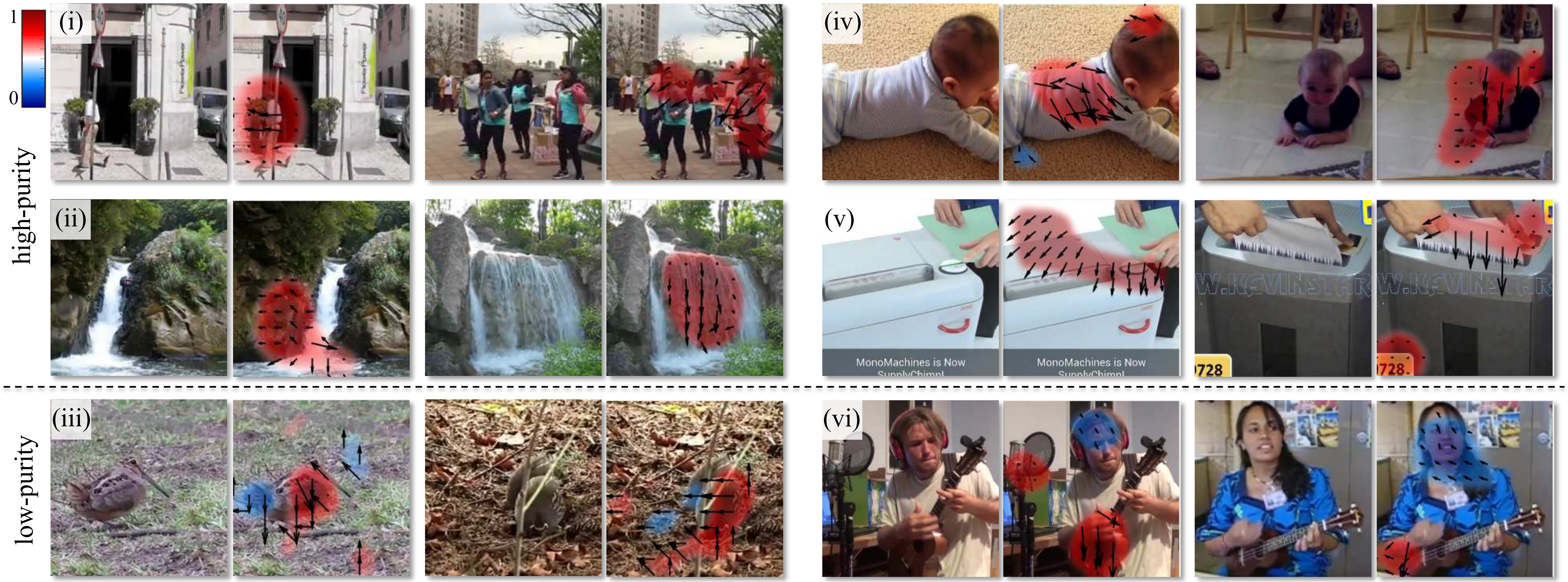
Avoid "cheating"

[Wei, Lim, Zisserman and Freeman , CVPR 2018]

Deep networks can leverage artificial cues to solve the task

		Black frame	+Camera motion
Percent of videos		46%	73%
Acc.	before removal	98%	88%
	after removal	90%	75%

Localization results [Wei, Lim, Zisserman and Freeman , CVPR 2018]



(a) Clusters in Flickr-AoT (b) Action classes in Kinetics-AoT

Finetuning for action classification

[Wei, Lim, Zisserman and Freeman , CVPR 2018]

✓ Results on UCF101:

Initialization		Fine-tune		
		Last layer	After fusion	All layers
Random	[24]	-	-	81.7%
	T-CAM	38.0%	53.1%	79.3%
ImageNet	[24]	-	-	85.7%
	T-CAM	47.9%	68.3%	84.1%
AoT (ours)	UCF101	58.6%	**81.2%**	**86.3%**
	Flickr	57.2%	79.2%	84.1%
	Kinetics	55.3%	74.3 %	79.4%