Google DeepMind

UNIVERSITY OF OXFORD

Self-Supervised Learning

Andrew Zisserman

Slides from: Carl Doersch, Ishan Misra, Andrew Owens, Carl Vondrick, Richard Zhang

The ImageNet Challenge Story …

1000 categories

- Training: 1000 images for each category
- Testing: 100k images

Flute

Strawberry

Traffic light

Backpack

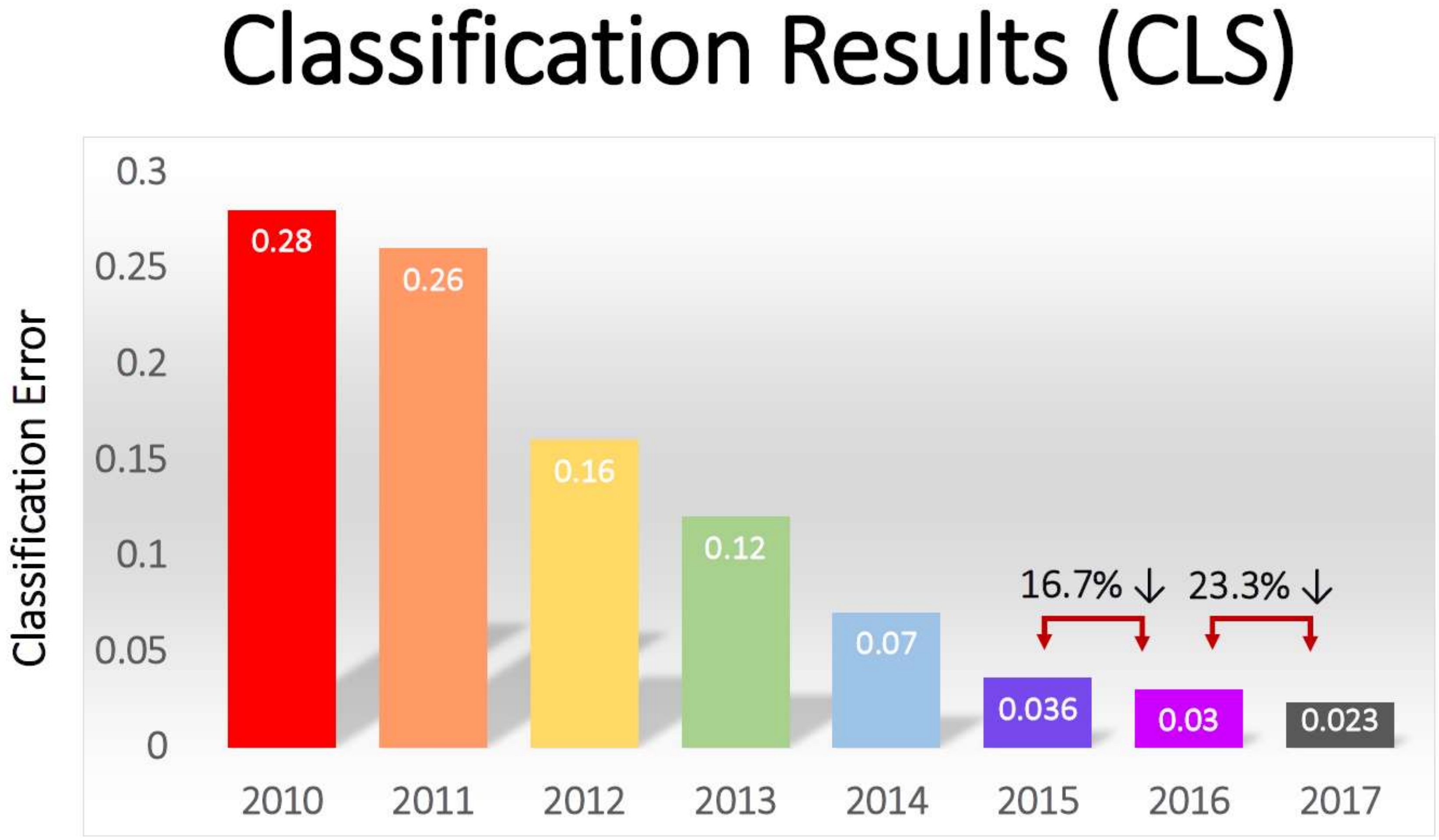
Matchstick

Bathing cap

Sea lion

Racket

The ImageNet Challenge Story … strong supervision



The ImageNet Challenge Story … outcomes

Strong supervision:

- Features from networks trained on ImageNet can be used for other visual tasks, e.g. detection, segmentation, action recognition, fine grained visual classification

- To some extent, any visual task can be solved now by:
    1. Construct a large-scale dataset labelled for that task
    2. Specify a training loss and neural network architecture
    3. Train the network and deploy

- Are there alternatives to strong supervision for training? Self-Supervised learning ….

Why Self-Supervision?

1. Expense of producing a new dataset for each new task

2. Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotation

3. Untapped/availability of vast numbers of unlabelled images/videos
   - Facebook: one billion images uploaded per day
   - 300 hours of video are uploaded to YouTube every minute

4. How infants may learn …

Self-Supervised Learning

The Scientist in the Crib: What Early Learning Tells Us About the Mind
by Alison Gopnik, Andrew N. Meltzoff and Patricia K. Kuhl

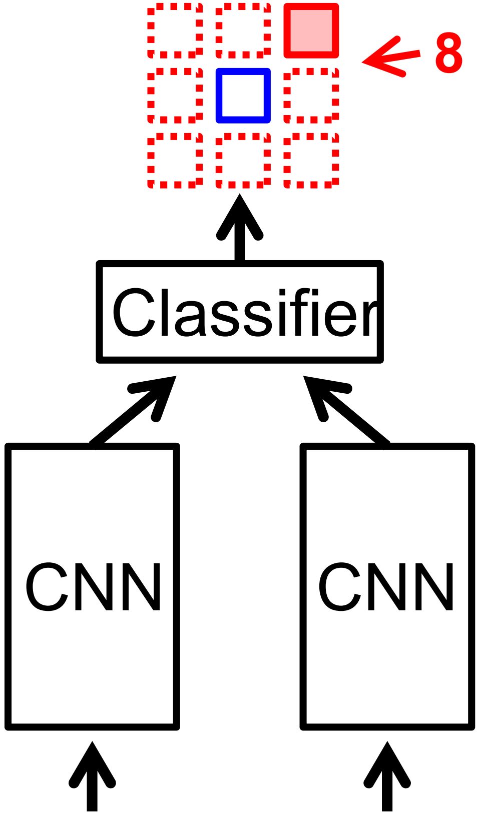
The Development of Embodied Cognition: Six Lessons from Babies
by Linda Smith and Michael Gasser

What is Self-Supervision?

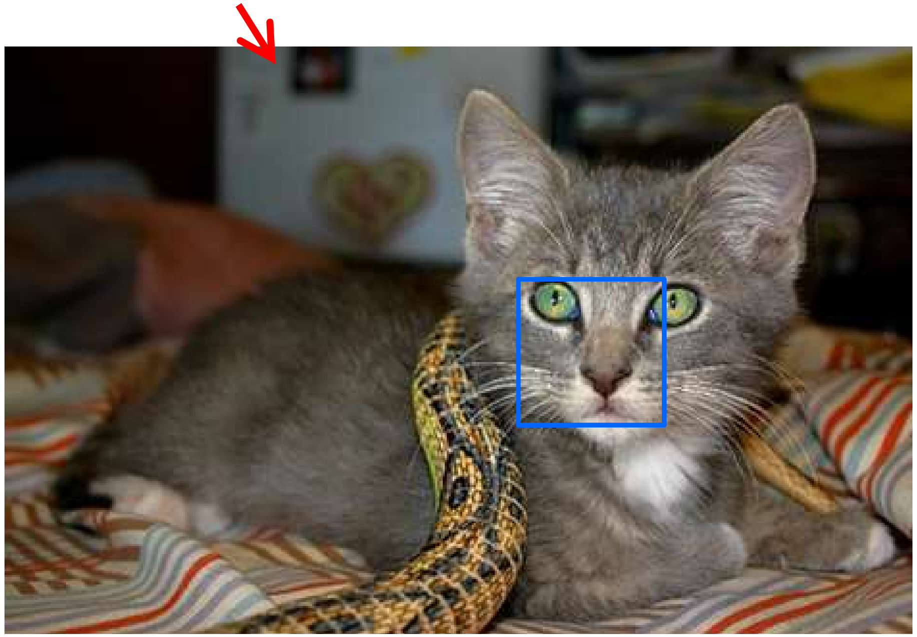
- A form of unsupervised learning where the data provides the **supervision**
- In general, withhold some part of the data, and task the network with predicting it
- The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to solve it

Example: relative positioning

Train network to predict relative position of two regions in the same image



8 possible locations

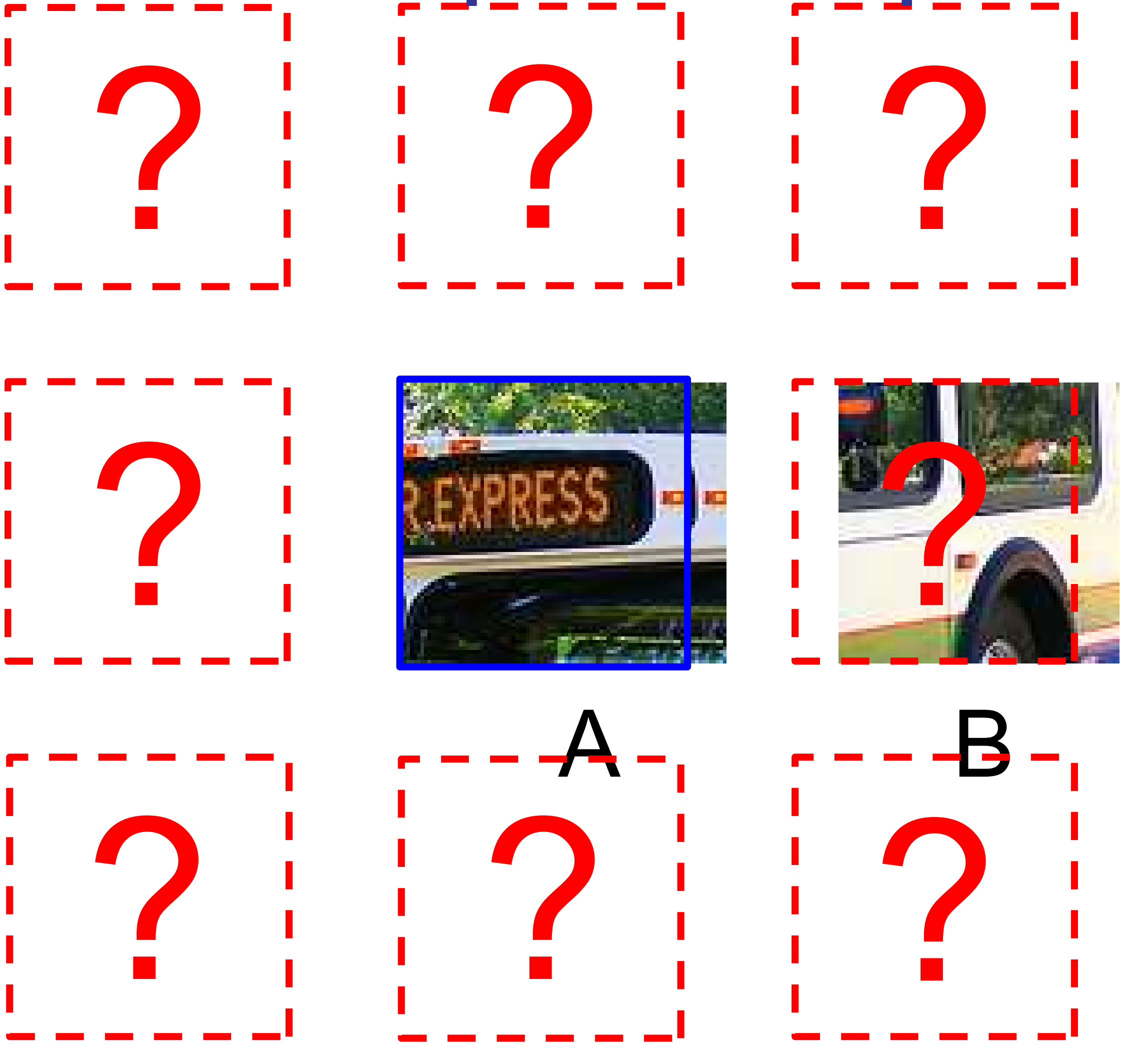


Randomly Sample Patch

Sample Second Patch

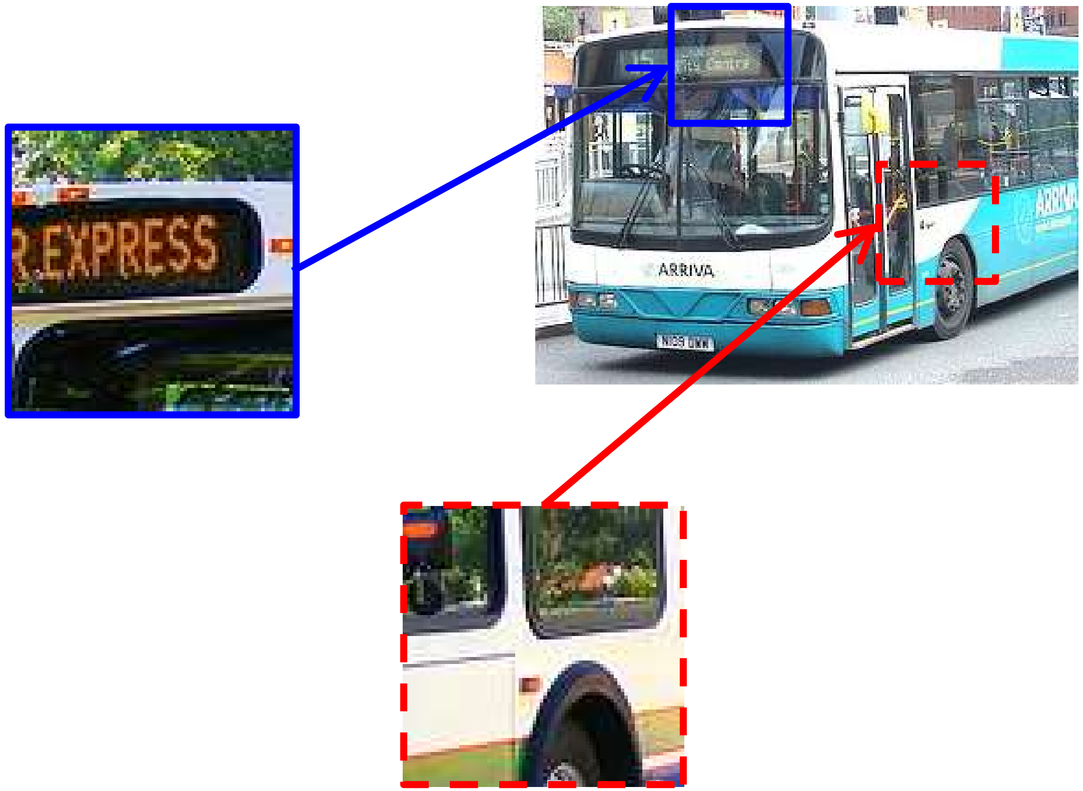
Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Example: relative positioning



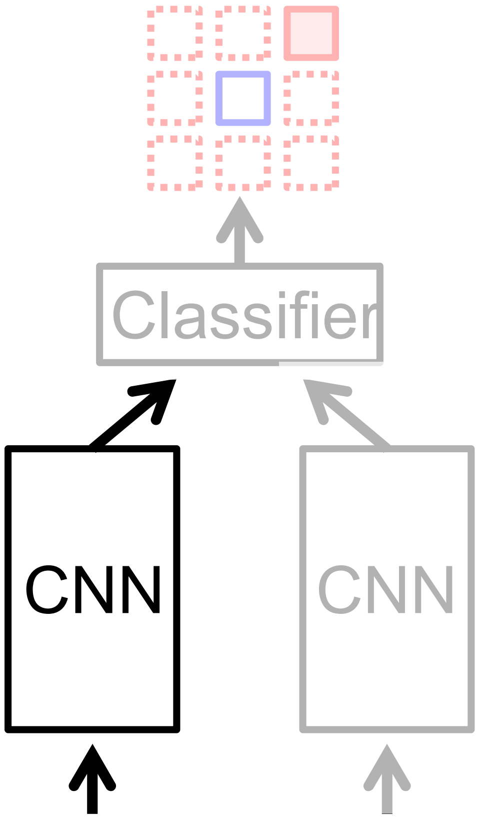
Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

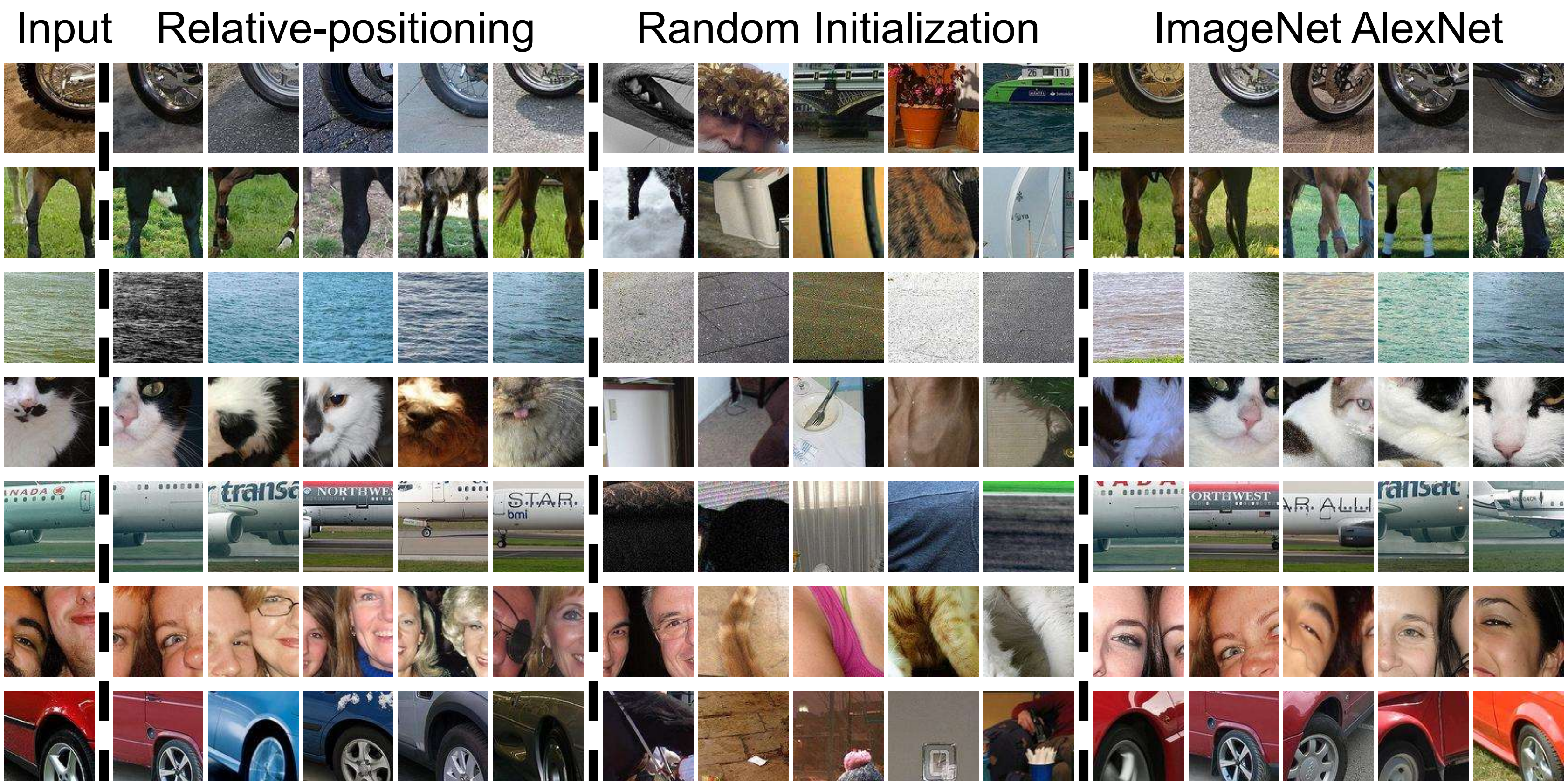
Semantics from a non-semantic task



Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

What is learned?





Outline

Self-supervised learning in three parts:

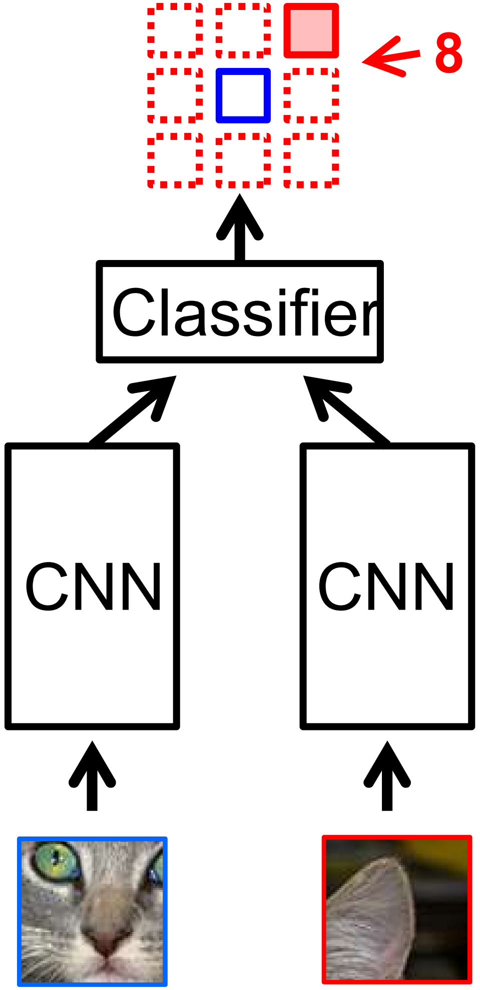
1. from images
2. from videos
3. from videos with sound

Part I

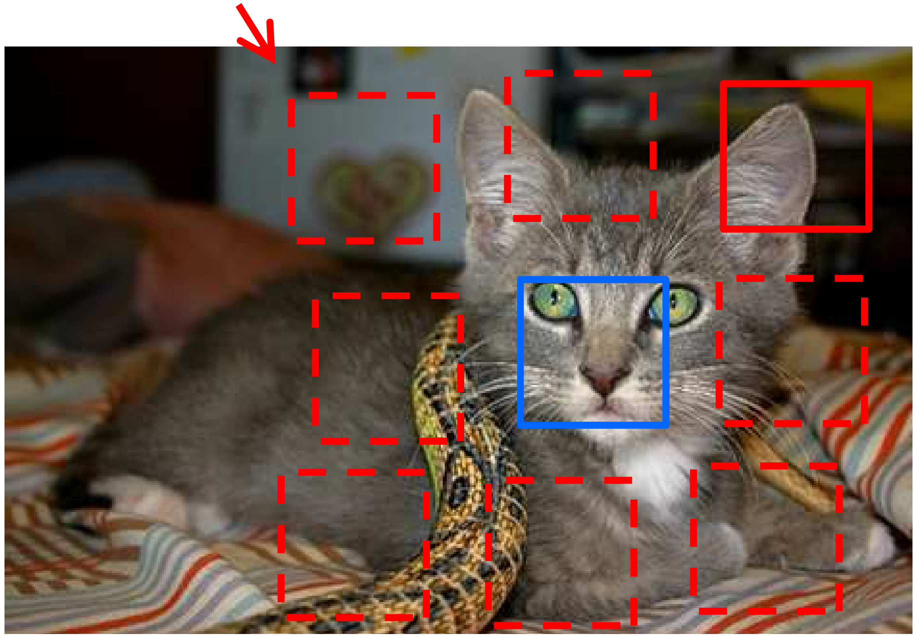
Self-Supervised Learning from Images

Recap: relative positioning

Train network to predict relative position of two regions in the same image



8 possible locations



Randomly Sample Patch

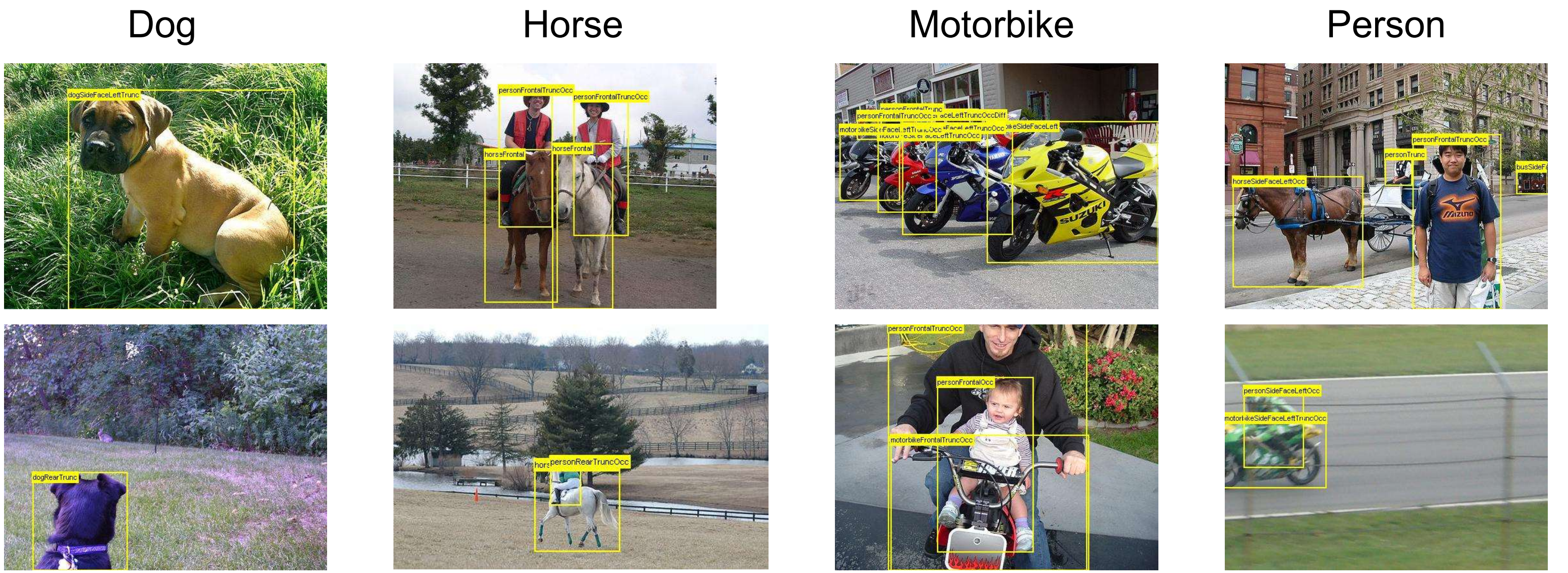
Sample Second Patch

Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Evaluation: PASCAL VOC Detection

- 20 object classes (car, bicycle, person, horse …)

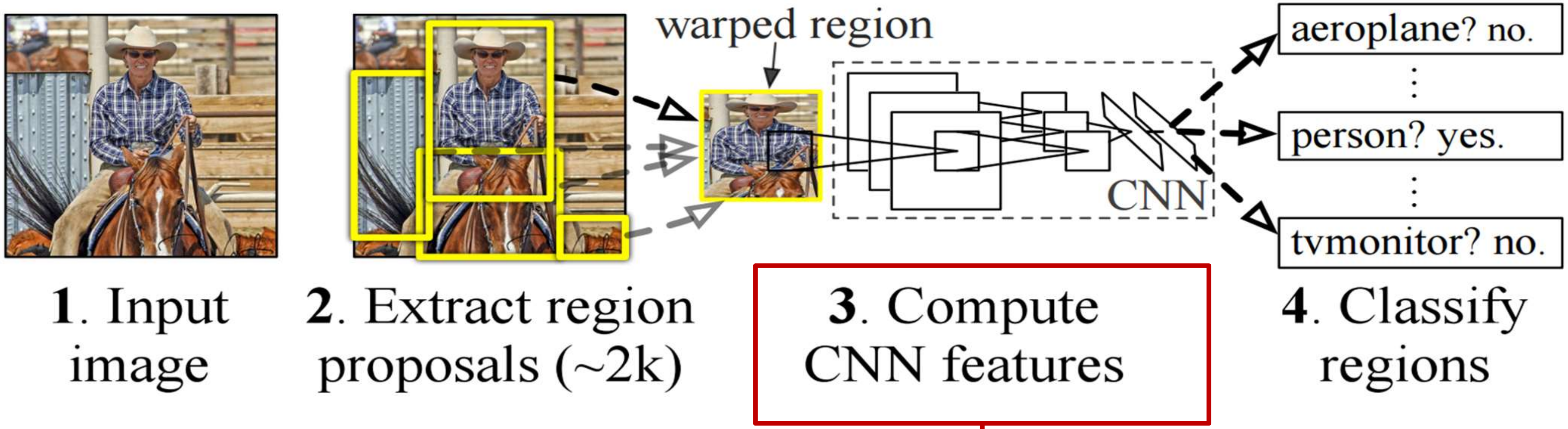
- Predict the bounding boxes of all objects of a given class in an image (if any)



Evaluation: PASCAL VOC Detection

- Pre-train CNN using self-supervision (no labels)
- Train CNN for detection in R-CNN object category detection pipeline

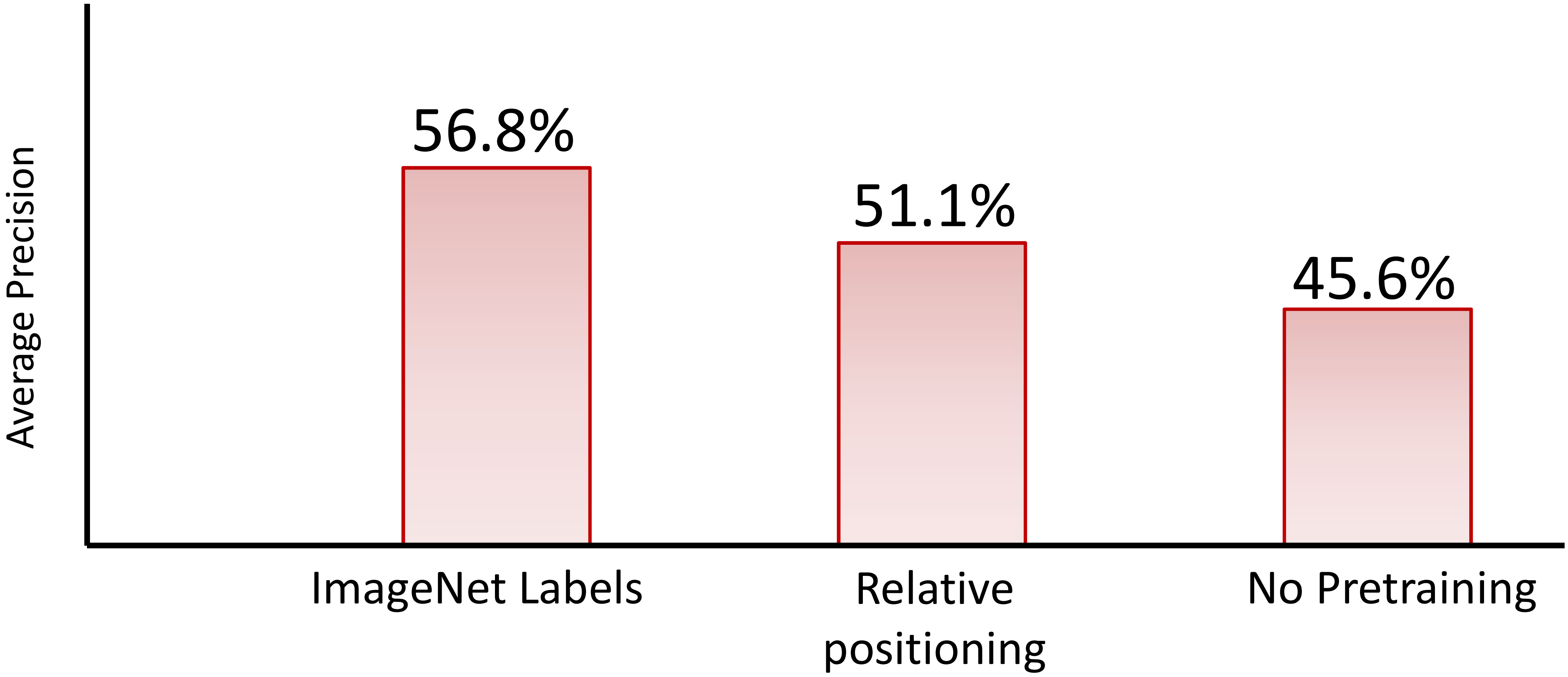
R-CNN



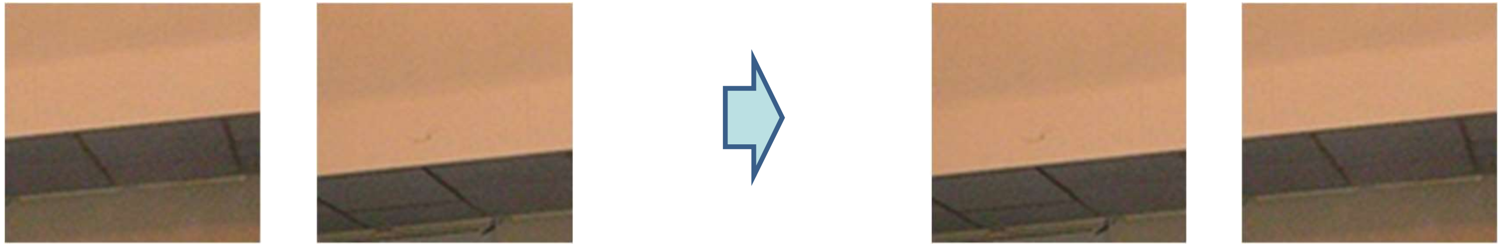
Pre-train on relative-position task, w/o labels

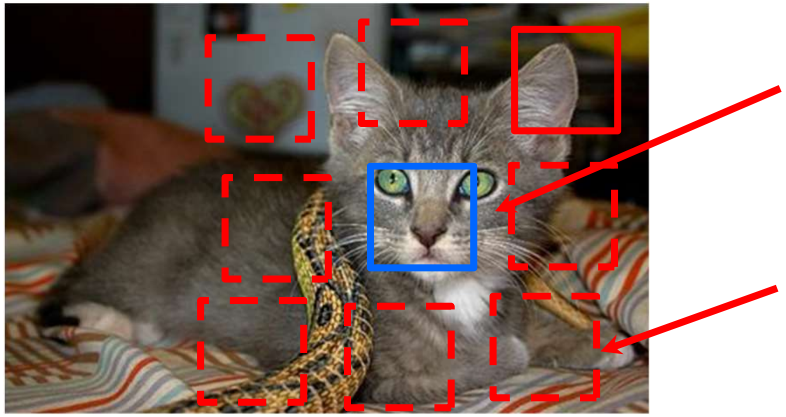
[Girshick et al. 2014]

Evaluation: PASCAL VOC Detection



Avoiding Trivial Shortcuts

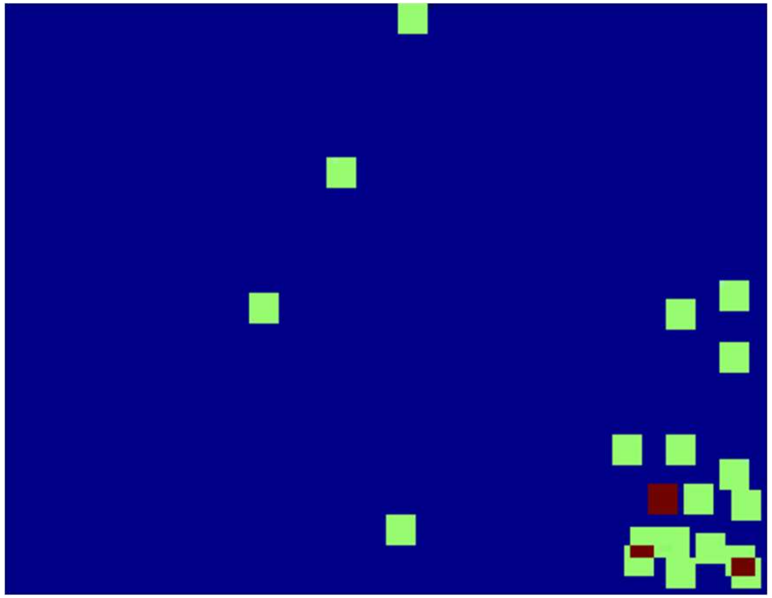




Include a gap

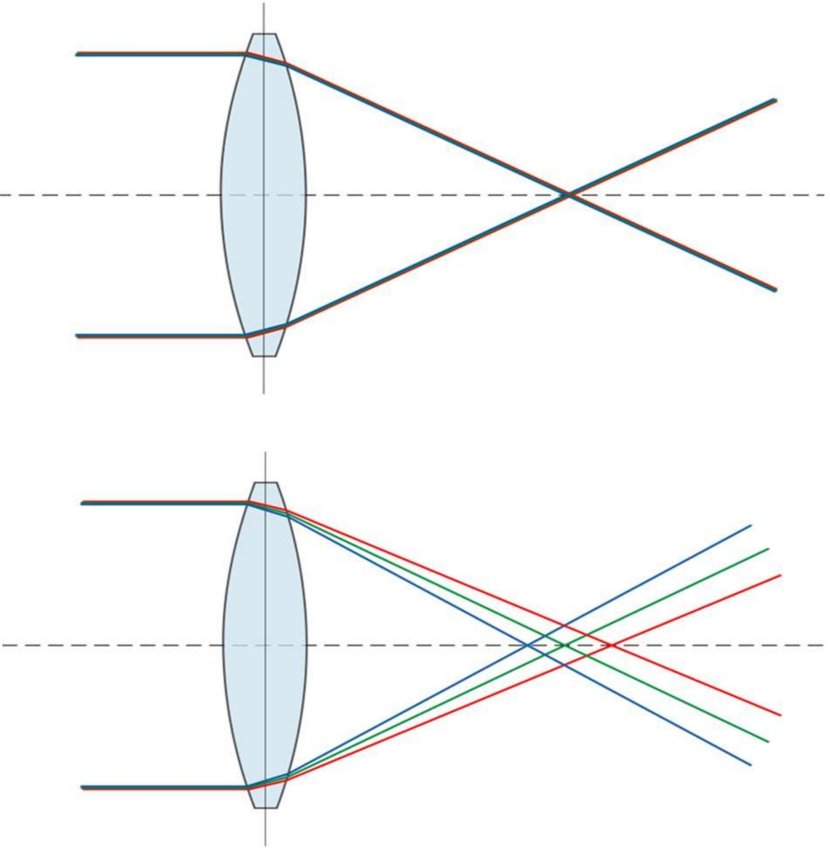
Jitter the patch locations

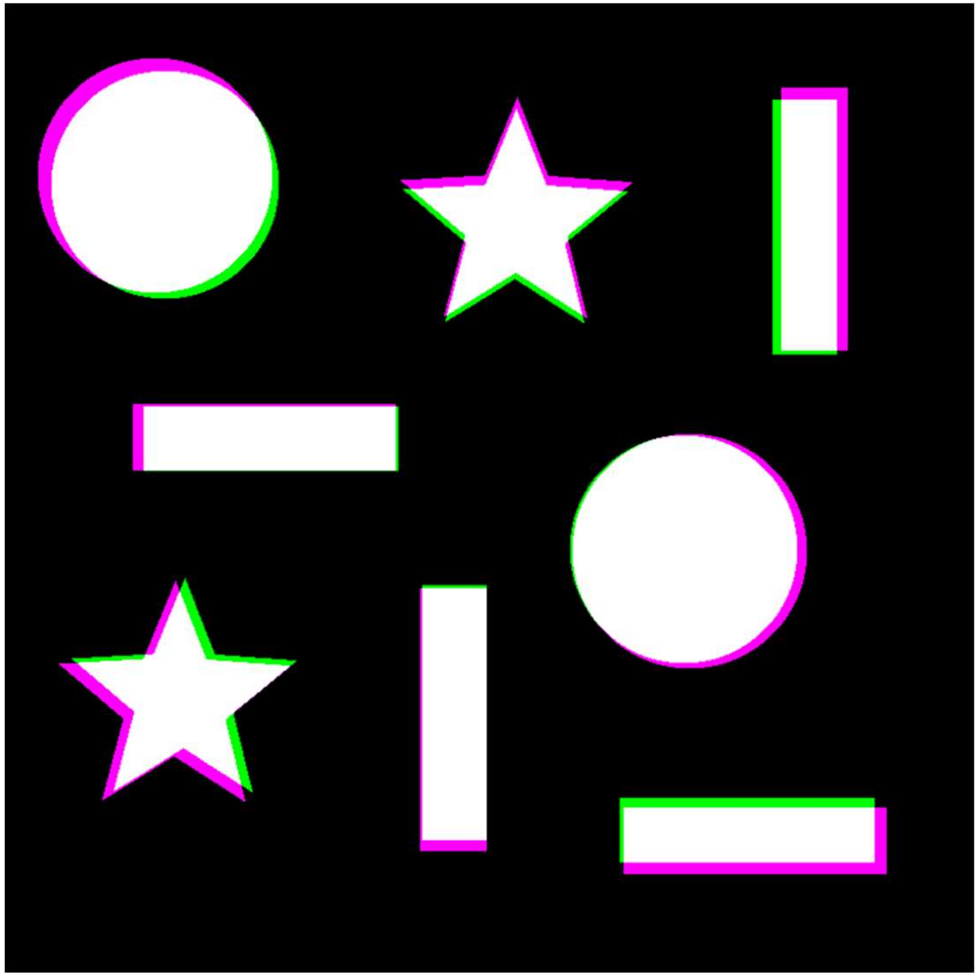
A Not-So "Trivial" Shortcut



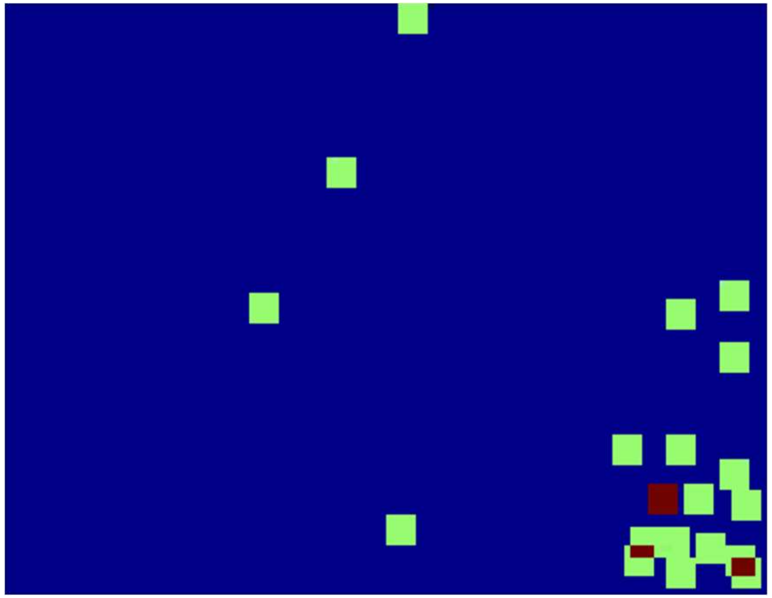
Position in Image

Chromatic Aberration





A Not-So “Trivial” Shortcut



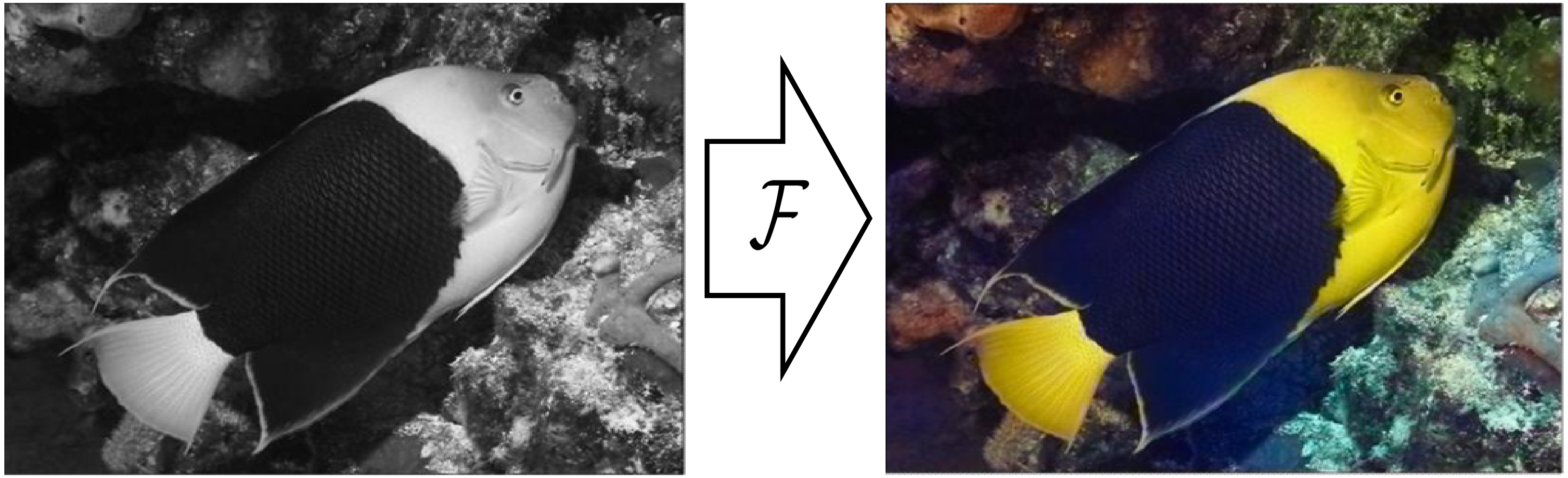
Position in Image

Solution?

Only use one of the colour channels

Image example II: colourization

Train network to predict pixel colour from a monochrome input



Grayscale image: *L* channel

$$\mathbf{X} \in \mathbb{R}^{H\times W\times 1}$$

Concatenate (L,ab)

$$(\mathbf{X}, \widehat{\mathbf{Y}})$$

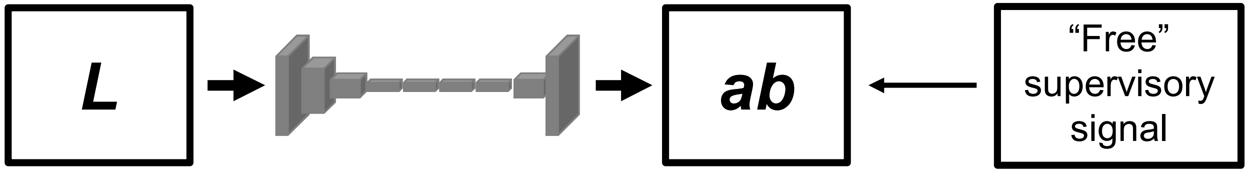
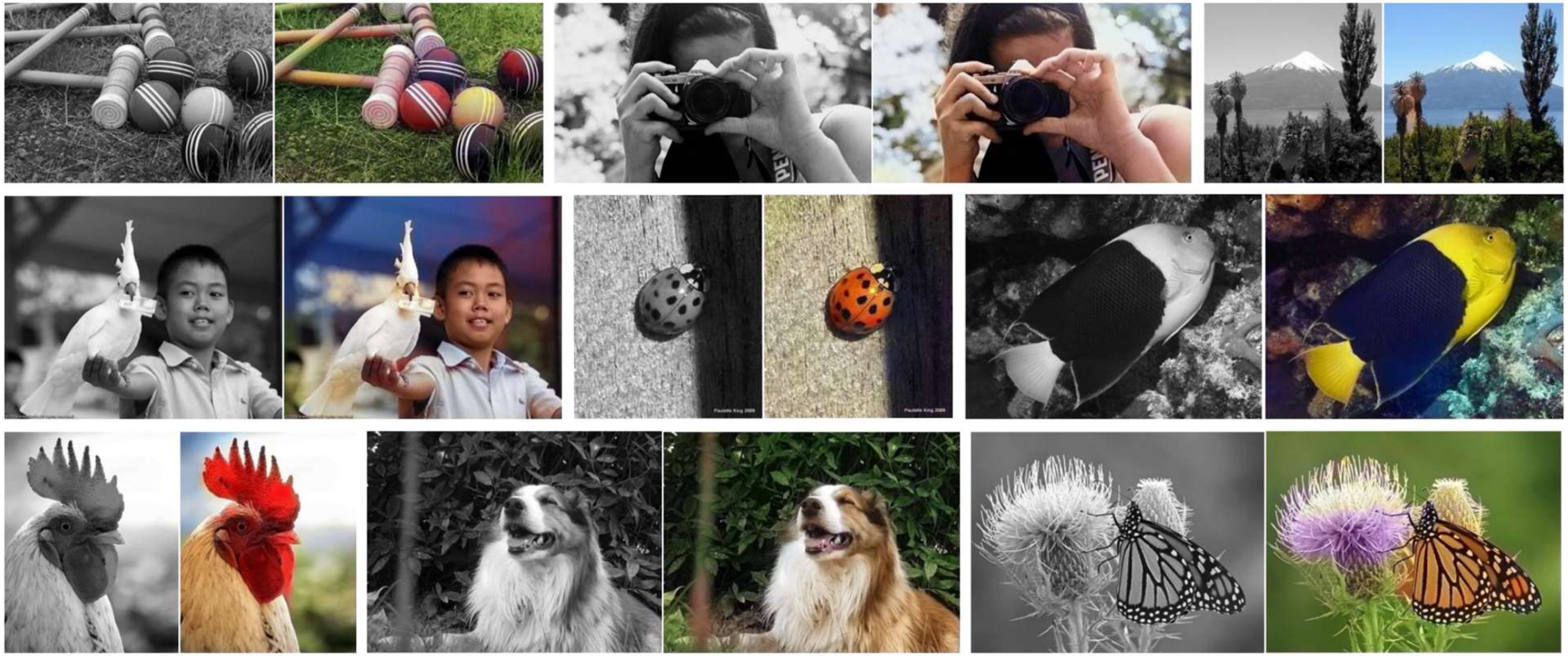


Image example II: colourization

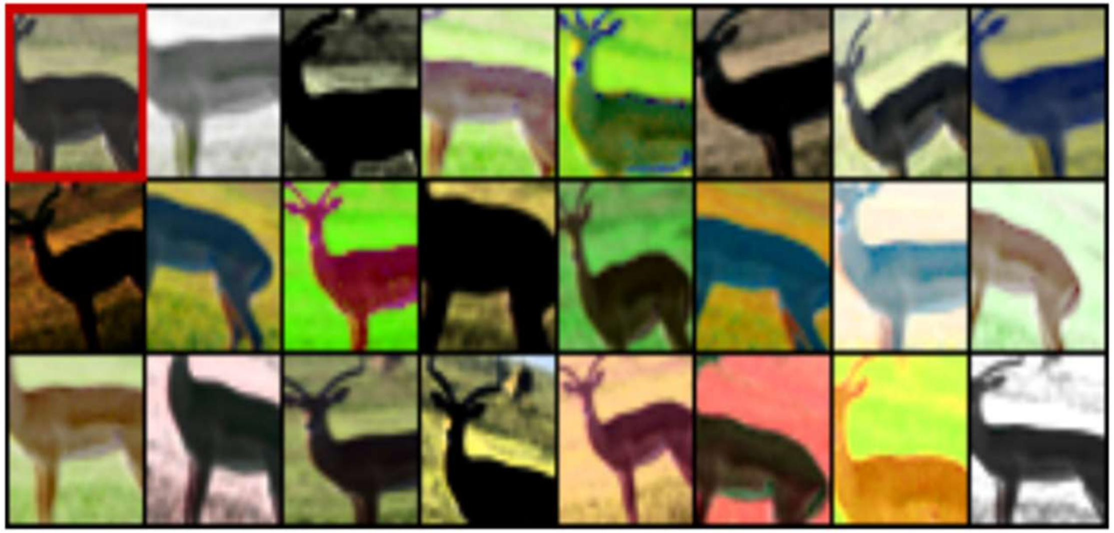
Train network to predict pixel colour from a monochrome input



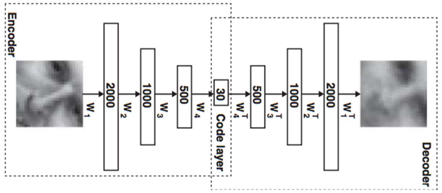
Colorful Image Colorization, Zhang *et al*., ECCV 2016

Image example III: exemplar networks

- Exemplar Networks (Dosovitskiy *et al*., 2014)
- Perturb/distort image patches, e.g. by cropping and affine transformations
- Train to classify these exemplars as same class

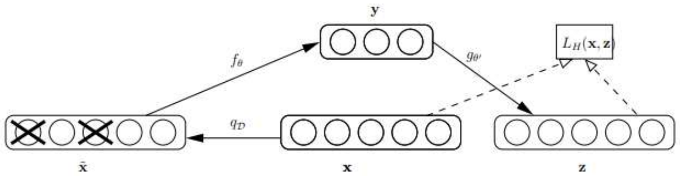


Autoencoders



Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders

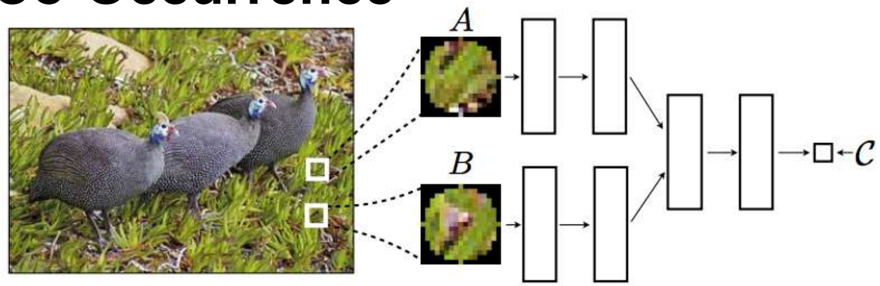


Vincent *et al.* ICML 2008.

Exemplar networks

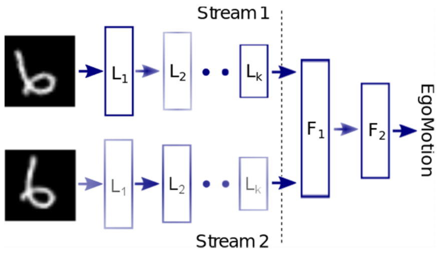
Dosovitskiy *et al*., NIPS 2014

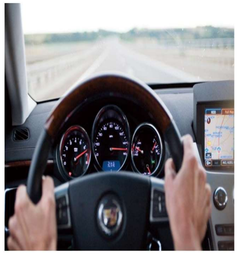
Co-Occurrence

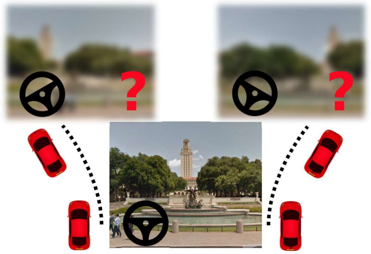


Isola *et al.* ICLR Workshop 2016.

Egomotion

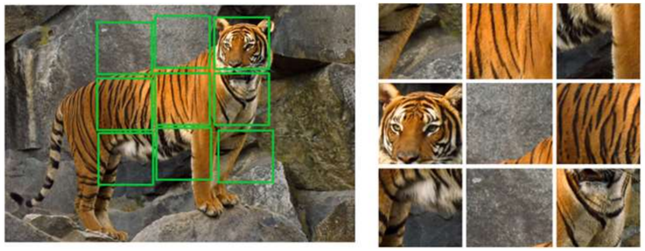




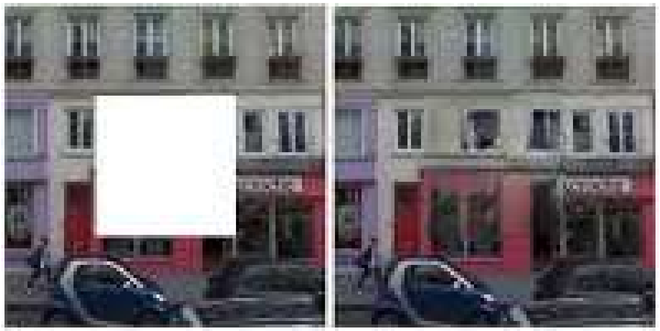


Agrawal *et al.* ICCV 2015   Jayaraman *et al.* ICCV 2015

Context

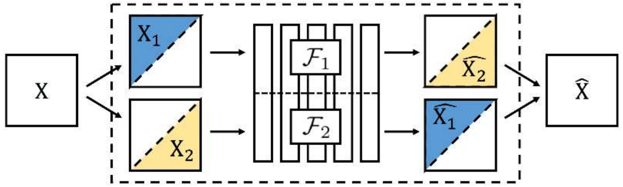


Noroozi et al 2016



Pathak *et al.* CVPR 2016

Split-brain auto-encoders



Zhang *et al.* CVPR 2017

Multi-Task Self-Supervised Learning

Procedure:

- ImageNet-frozen: self-supervised training, network fixed, classifier trained on features
- PASCAL: self-supervised pre-training, then train Faster-RCNN
- ImageNet labels: strong supervision

NB: all methods re-implemented on same backbone network (ResNet-101)

Self-supervision task	ImageNet Classification top-5 accuracy	PASCAL VOC Detection mAP
Rel. Pos	59.21	66.75
Colour	62.48	65.47
Exemplar	53.08	60.94
Rel. Pos + colour	66.64	68.75
Rel. Pos + Exemplar	65.24	69.44
Rel. Pos + colour + Exemplar	68.65	69.48
ImageNet labels	85.10	74.17

Multi-task self-supervised visual learning, C Doersch, A Zisserman, ICCV 2017

Multi-Task Self-Supervised Learning

Findings:

- Deeper network improves performance (ResNet vs AlexNet)
- Colour and Rel-Pos superior to Exemplar
- Gap between self-supervision and strong supervision closing

Self-supervision task	ImageNet Classification top-5 accuracy	PASCAL VOC Detection mAP
Rel. Pos	59.21	66.75
Colour	62.48	65.47
Exemplar	53.08	60.94
Rel. Pos + colour	66.64	68.75
Rel. Pos + Exemplar	65.24	69.44
Rel. Pos + colour + Exemplar	68.65	69.48
ImageNet labels	85.10	74.17

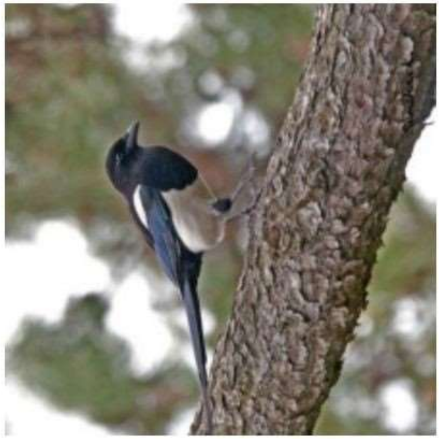
Procedure:

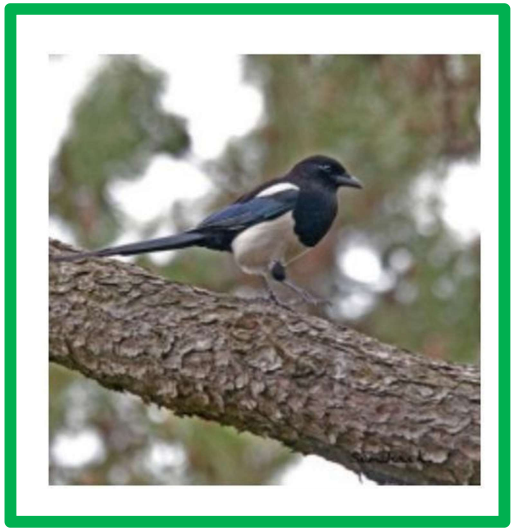
- ImageNet-frozen: self-supervised training, network fixed, classifier trained on features
- PASCAL: self-supervised pre-training, then train Faster-RCNN
- ImageNet labels: strong supervision

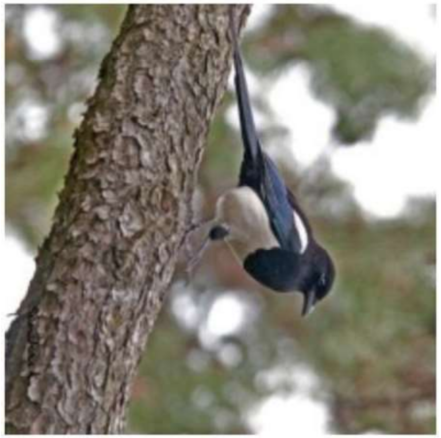
Multi-task self-supervised visual learning, C Doersch, A Zisserman, ICCV 2017

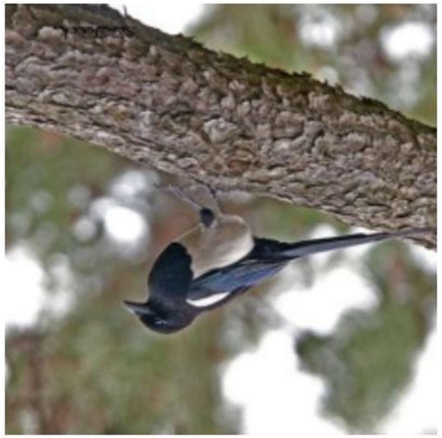
Image Transformations – 2018

Which image has the correct rotation?



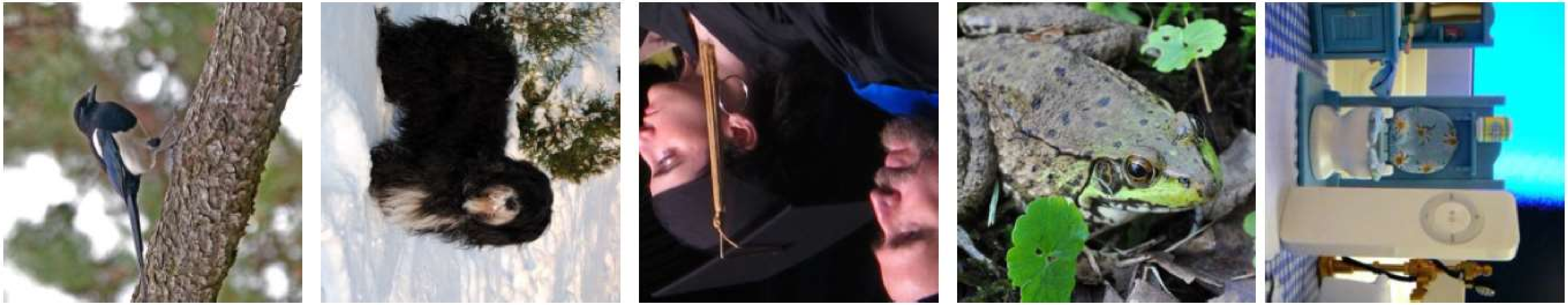






Unsupervised representation learning by predicting image rotations,
Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

Image Transformations – 2018

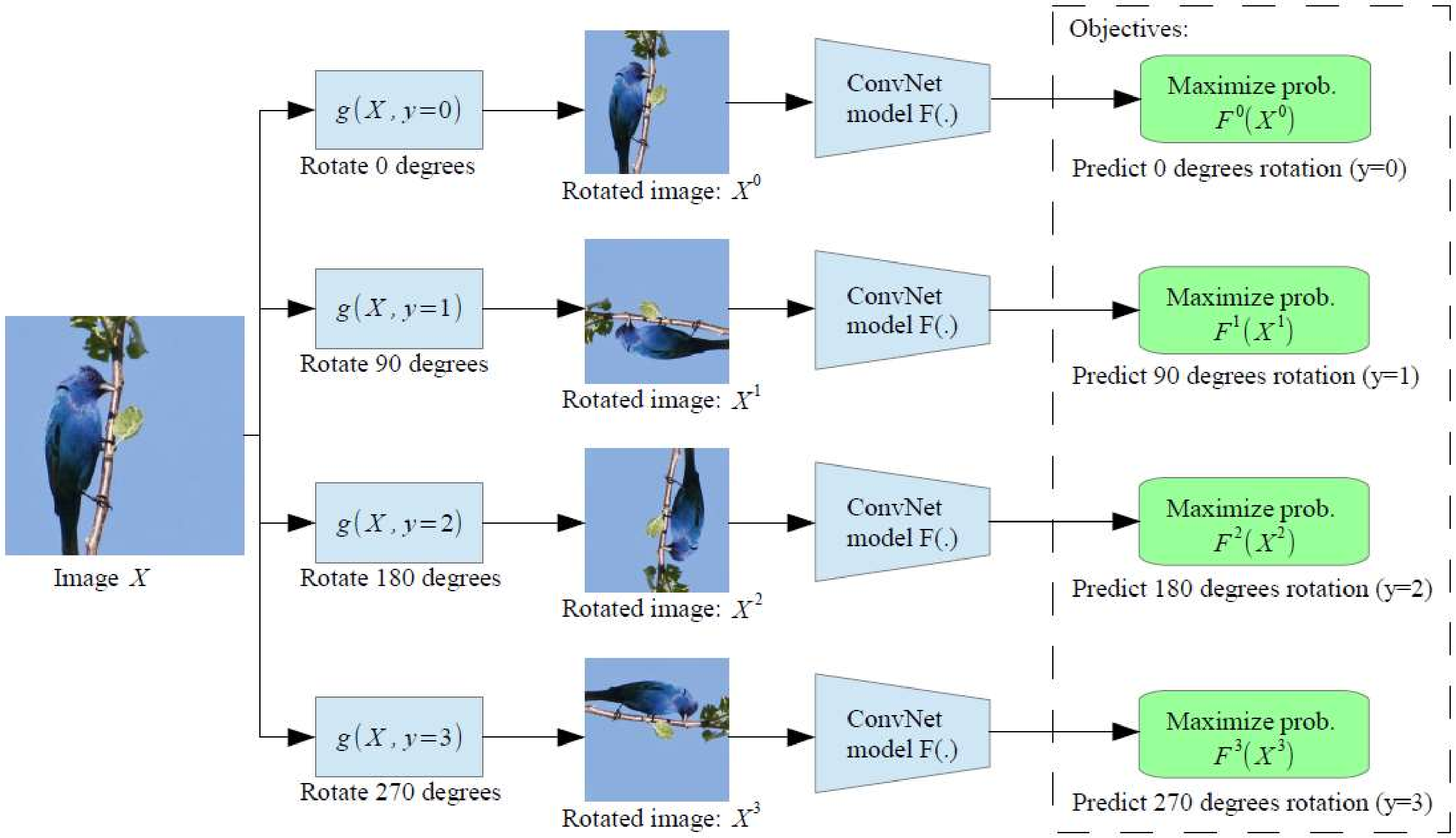


90° rotation 270° rotation 180° rotation 0° rotation 270° rotation

Figure 1: Images rotated by random multiples of 90 degrees (e.g., 0, 90, 180, or 270 degrees). The core intuition of our self-supervised feature learning approach is that if someone is not aware of the concepts of the objects depicted in the images, he cannot recognize the rotation that was applied to them.

Unsupervised representation learning by predicting image rotations,
Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

Image Transformations – 2018



Unsupervised representation learning by predicting image rotations,
Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

Image Transformations – 2018

- Uses AlexNet
- Closes gap between ImageNet and self-supervision

	PASCAL VOC Detection mAP
Random	43.4
Rel. Pos.	51.1
Colour	46.9
Rotation	54.4
ImageNet Labels	56.8

Unsupervised representation learning by predicting image rotations,
Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

Summary Point

- Self-Supervision:
  - A form of unsupervised learning where the data provides the **supervision**
  - In general, withhold some information about the data, and task the network with predicting it
  - The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to solve it
- Many self-supervised tasks for images
- Often complementary, and combining improves performance
- Closing gap with strong supervision from ImageNet label training
  - ImageNet image classification, PASCAL VOC detection
- Deeper networks improve performance

Part II

Self-Supervised Learning from Videos

Video

A temporal sequence of frames

What can we use to define a proxy loss?

- Nearby (in time) frames are strongly correlated, further away may not be
- Temporal order of the frames
- Motion of objects (via optical flow)
- …

Outline

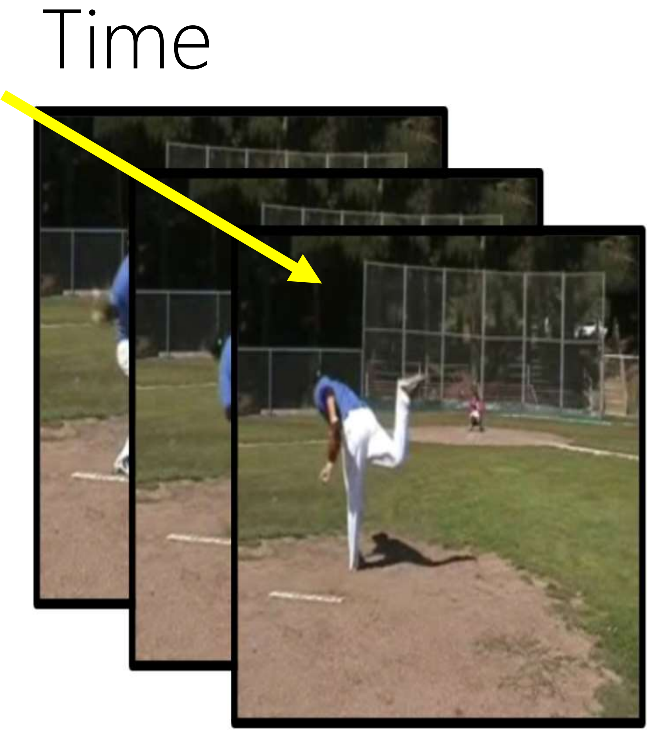
Three example tasks:

- Video sequence order
- Video direction
- Video tracking

Temporal structure in videos

Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra, C. Lawrence Zitnick and Martial Hebert
ECCV 2016

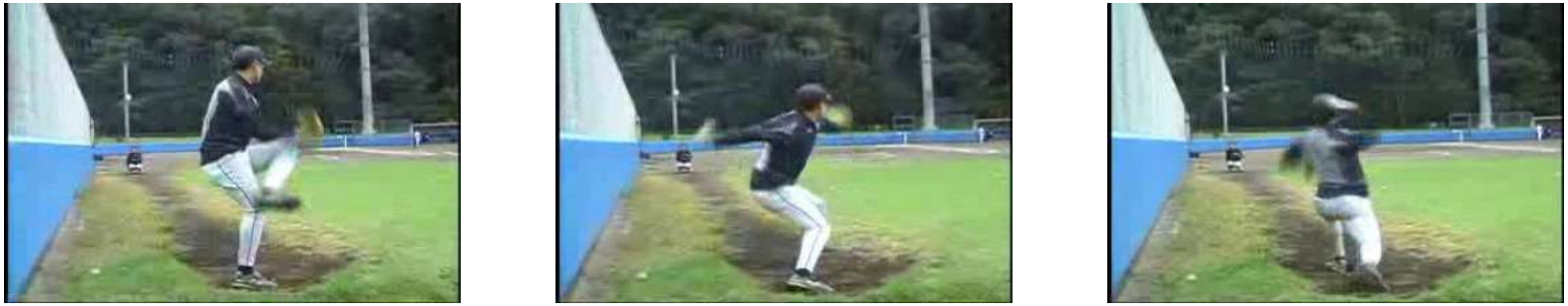


"Sequence" of data

Slide credit: Ishan Misra

Sequential Verification

- Is this a valid sequence?



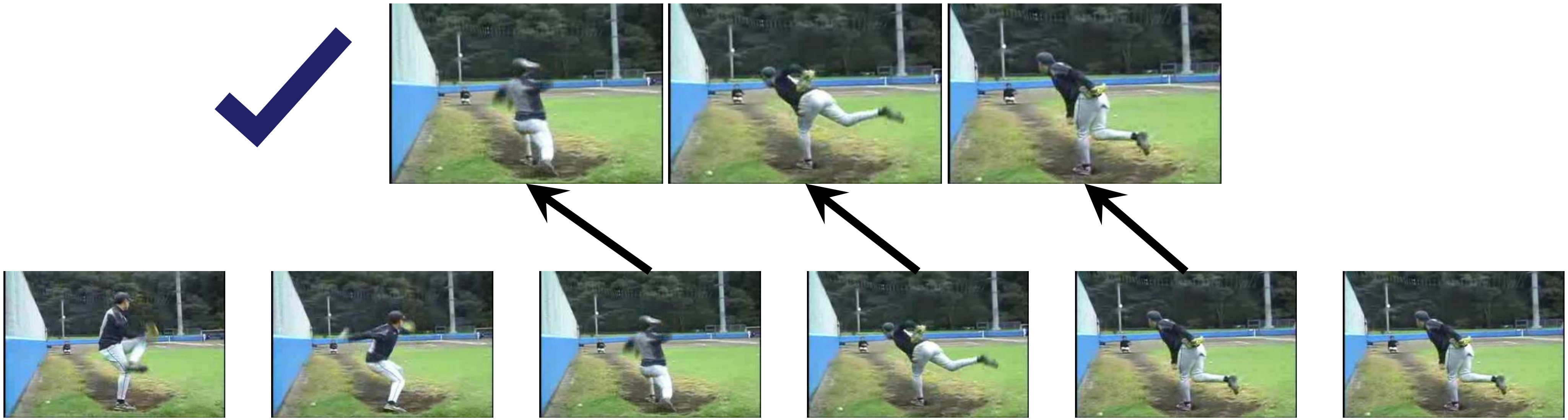
Sun and Giles, 2001; Sun et al., 2001; Cleermans 1993; Reber 1989
Arrow of Time - Pickup et al., 2014

Slide credit: Ishan Misra

Original video

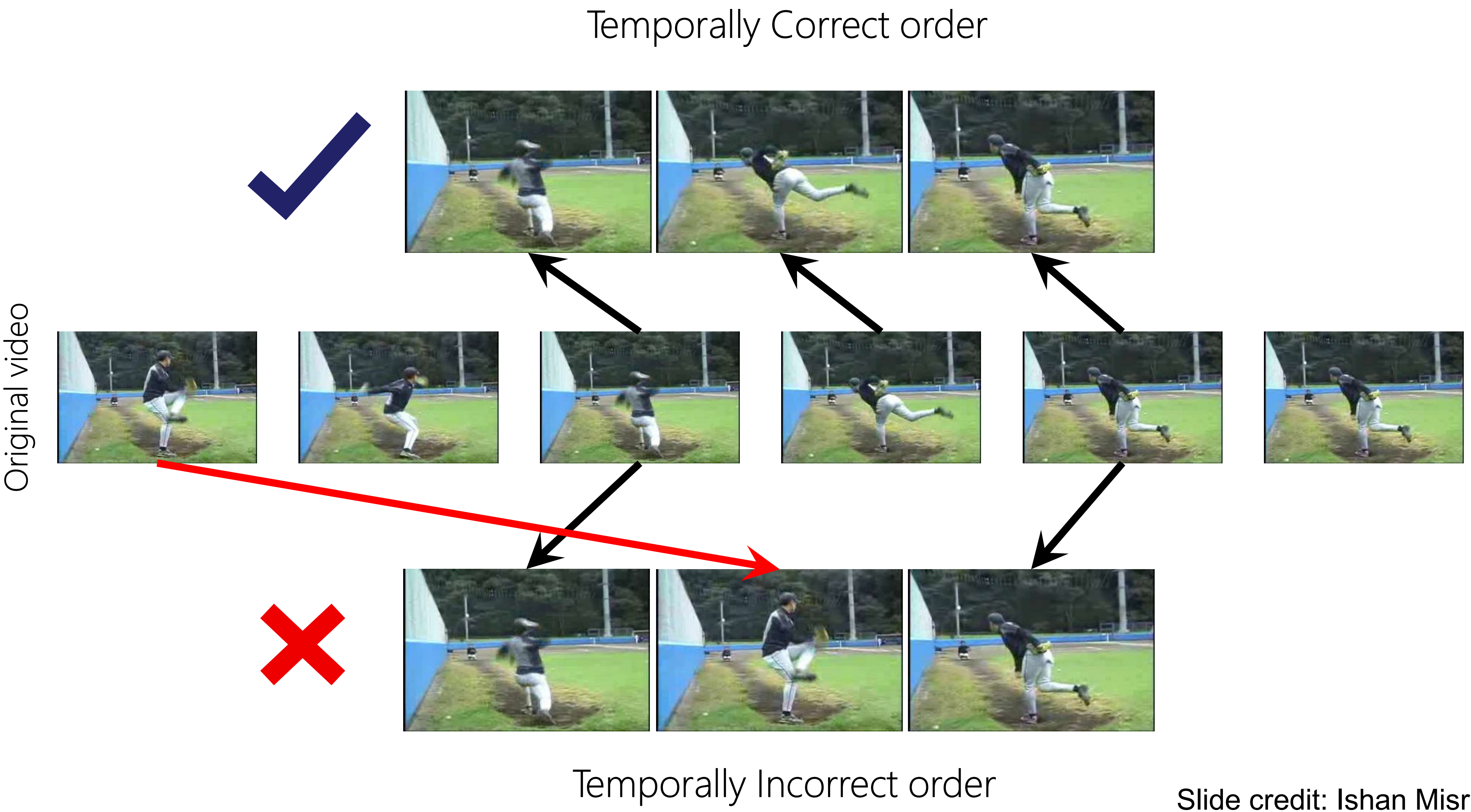
Slide credit: Ishan Misra

Temporally Correct order



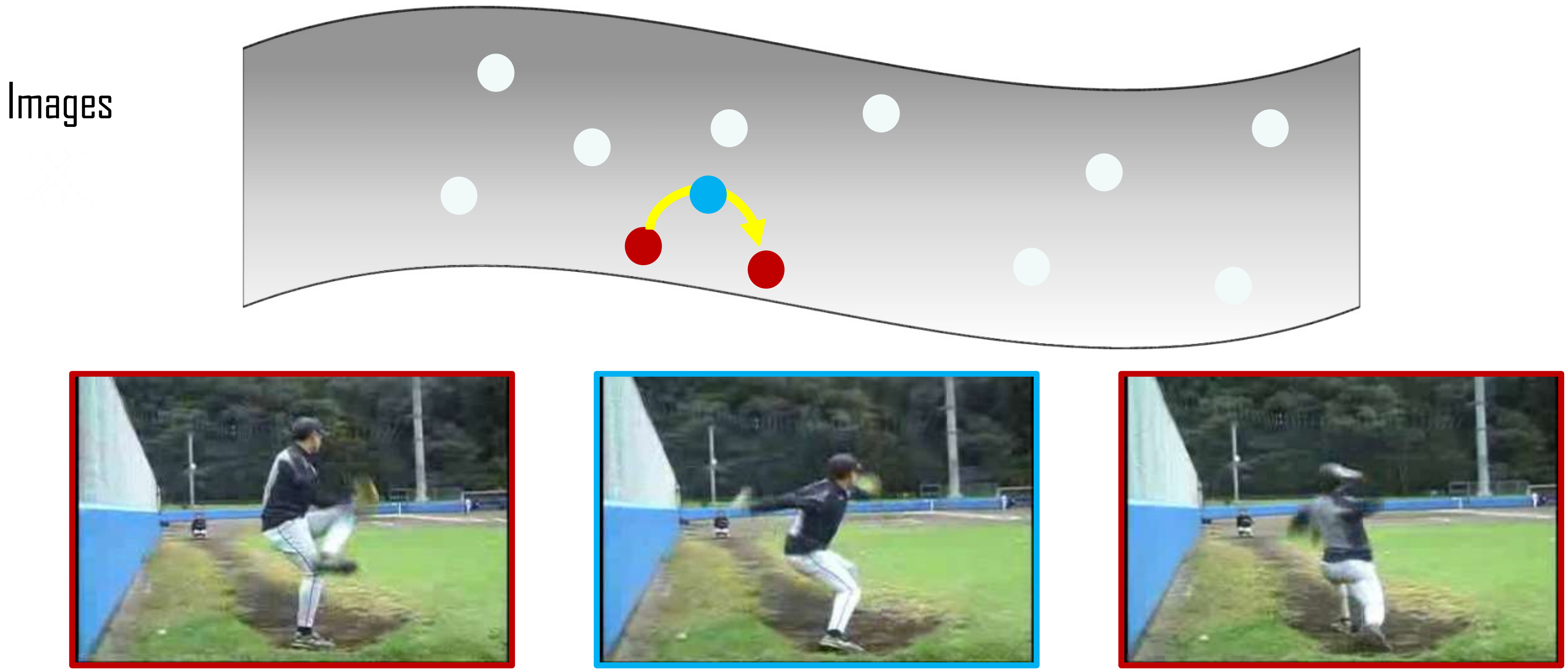
Original video

Slide credit: Ishan Misra



Slide credit: Ishan Misra

Geometric View



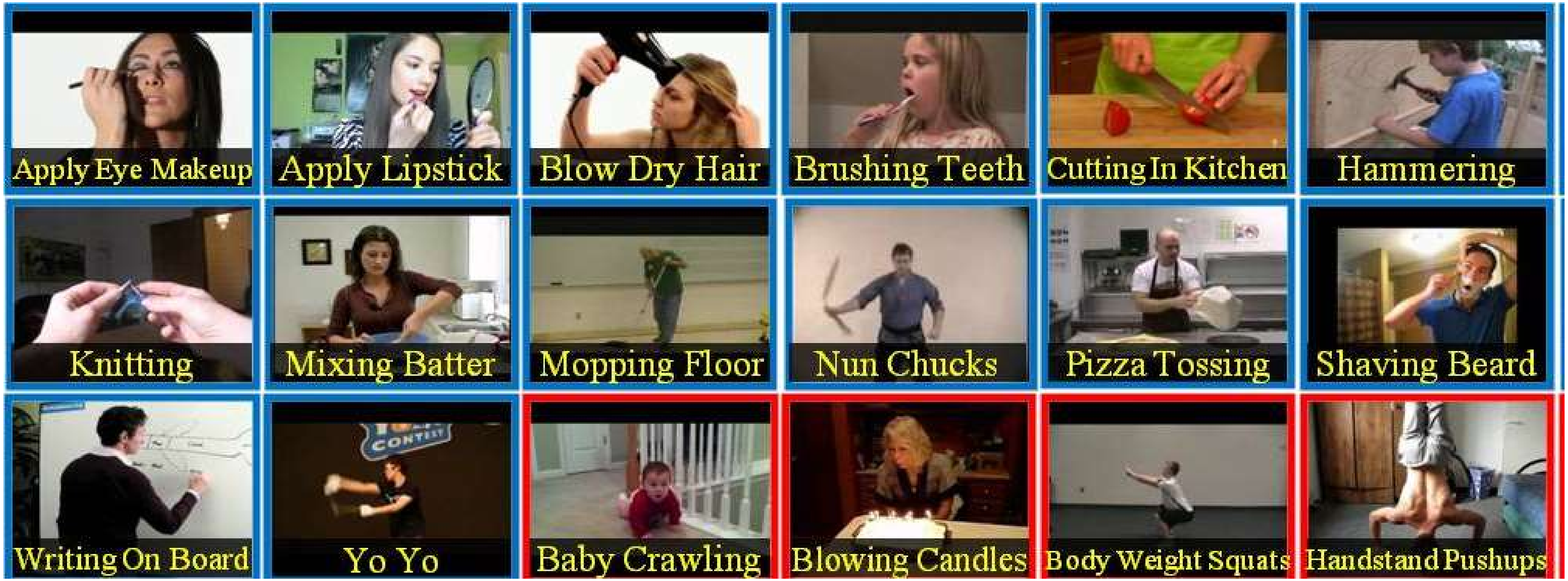
Images

Given a start and an end, can this point lie in between?

Shuffle and Learn – I. Misra, L. Zitnick, M. Hebert – ECCV 2016

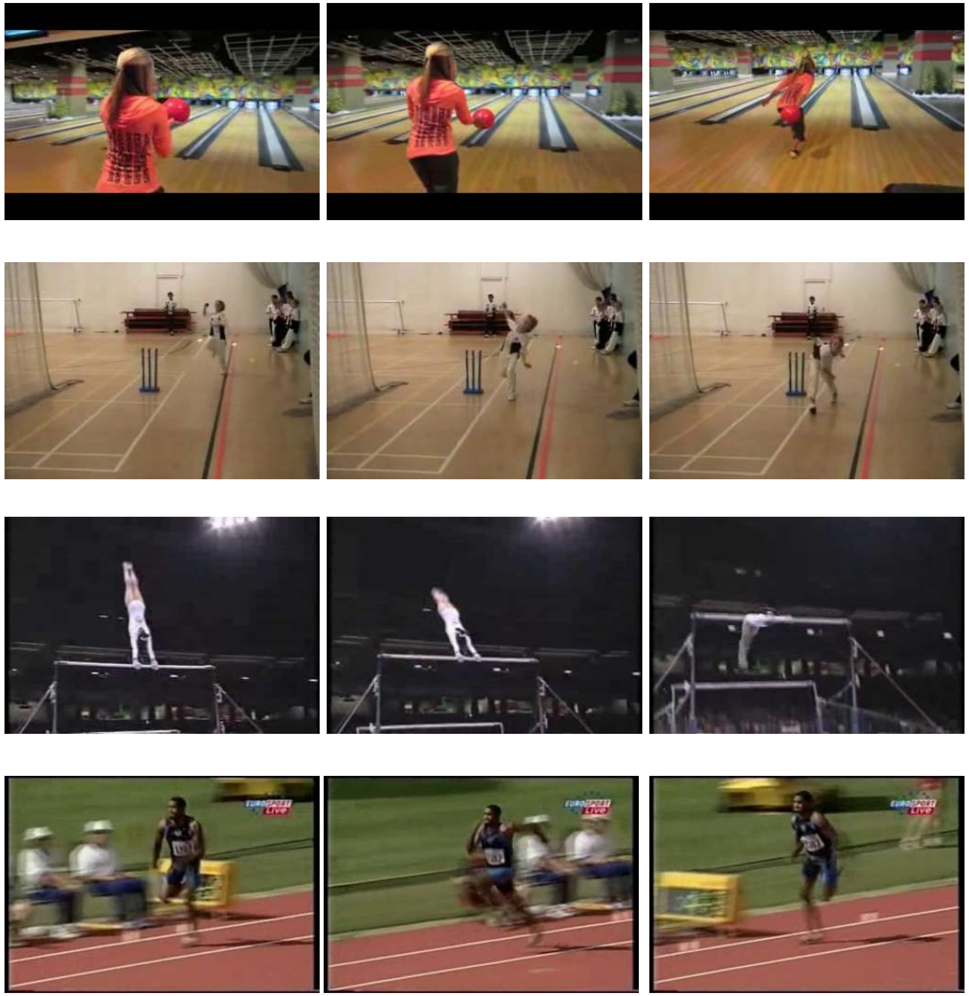
Slide credit: Ishan Misra

Dataset: UCF-101 Action Recognition

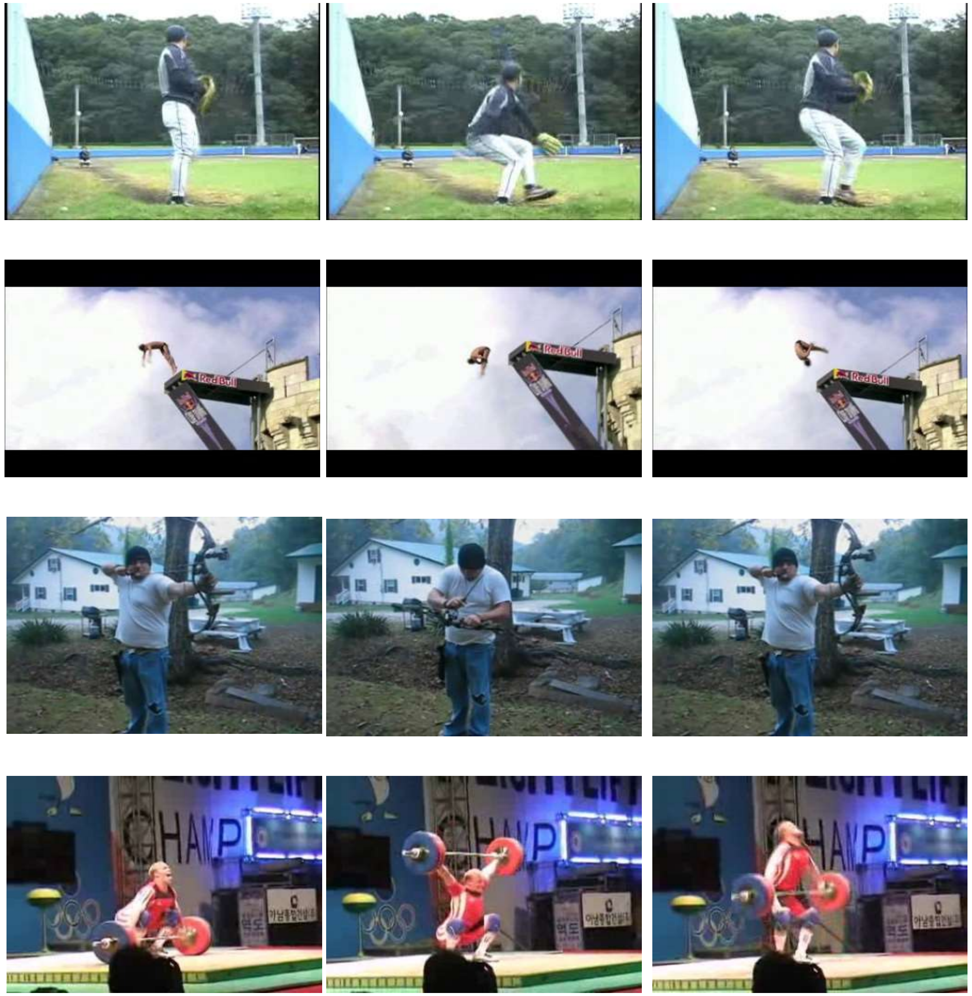


UCF101 - Soomro et al., 2012

Positive Tuples



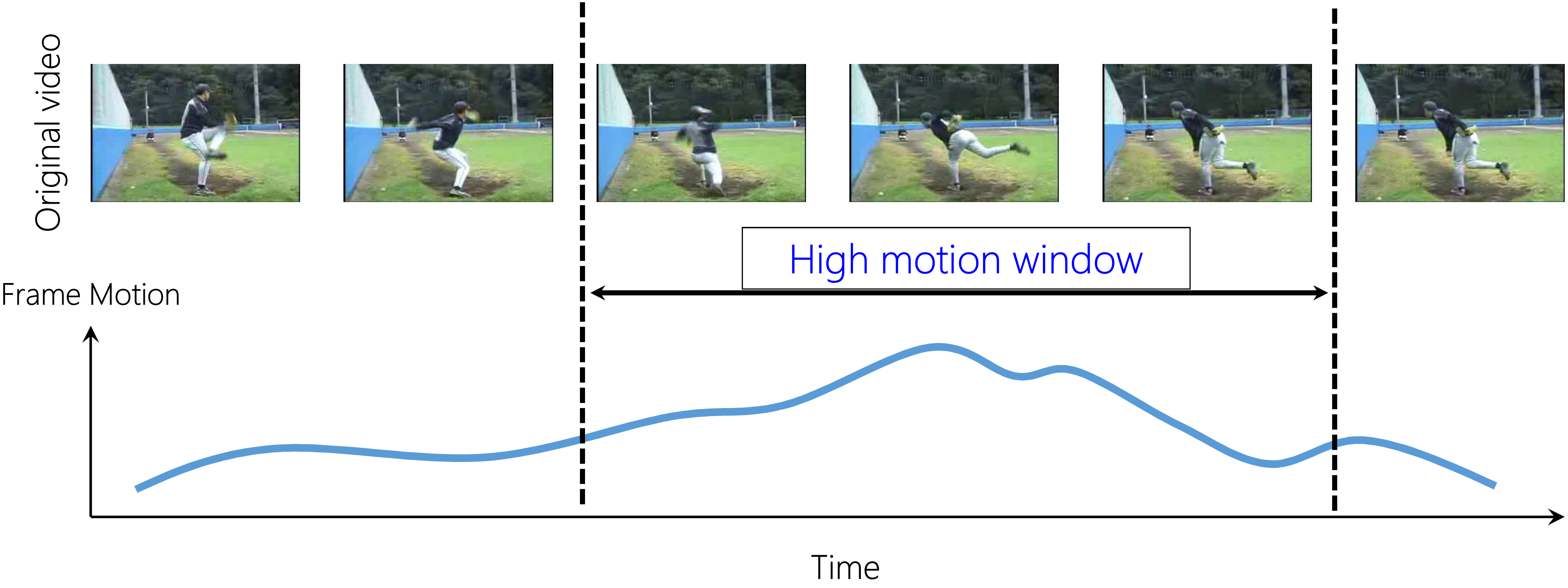
Negative Tuples



~900k tuples from UCF-101 dataset (Soomro et al., 2012)

Slide credit: Ishan Misra

Informative training tuples



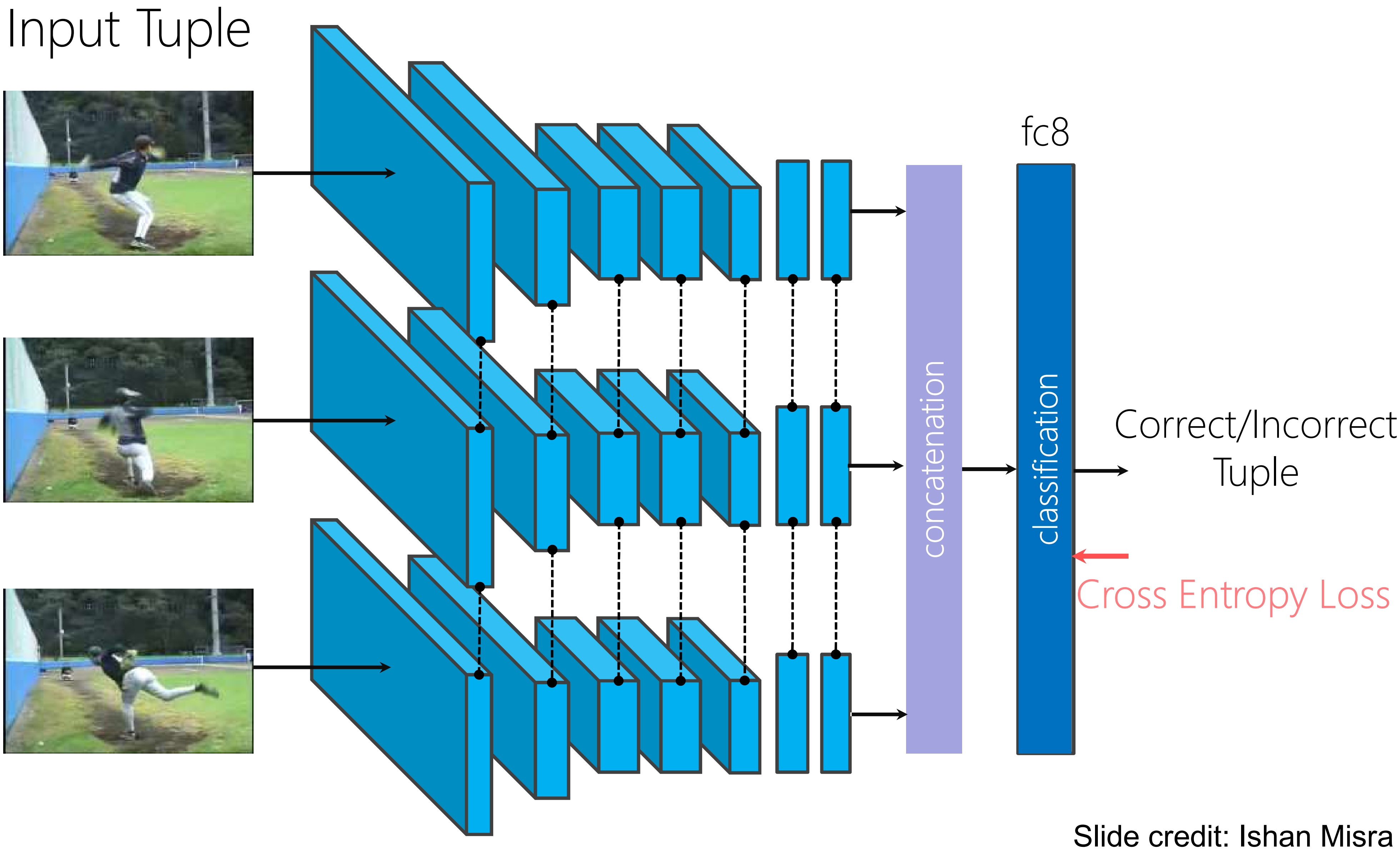
Original video

High motion window

Frame Motion

Time

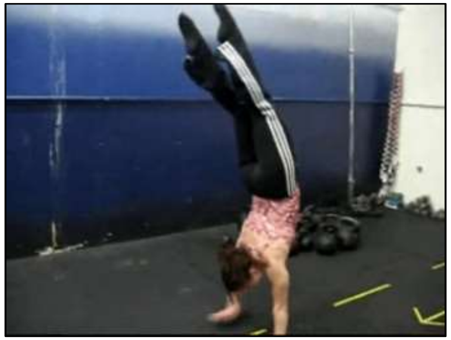
Slide credit: Ishan Misra

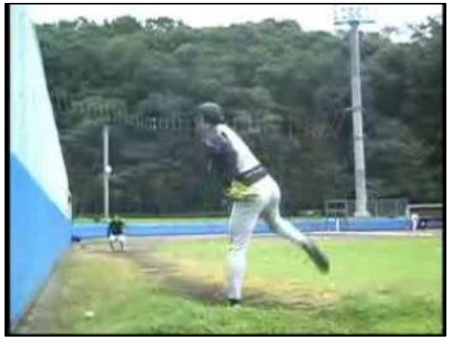


Slide credit: Ishan Misra

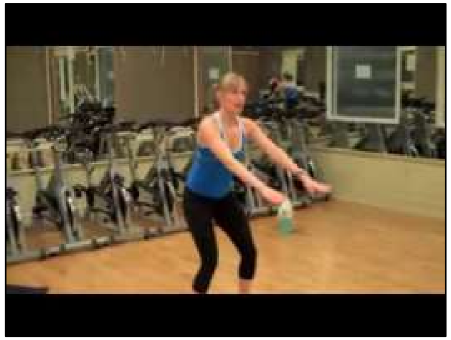
Nearest Neighbors of Query Frame (fc7 features)

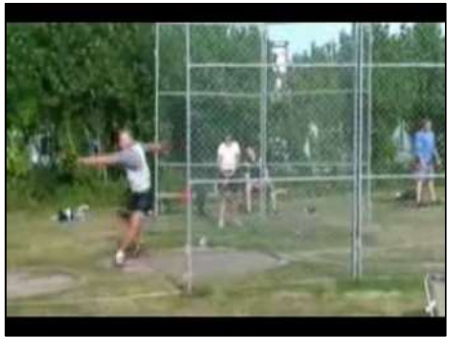
Query



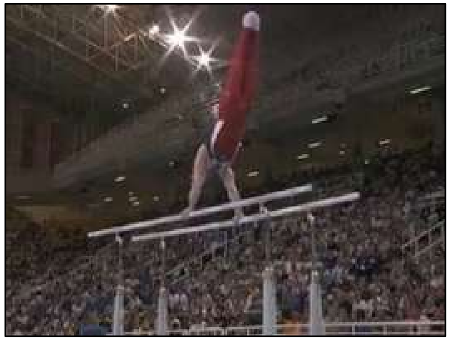


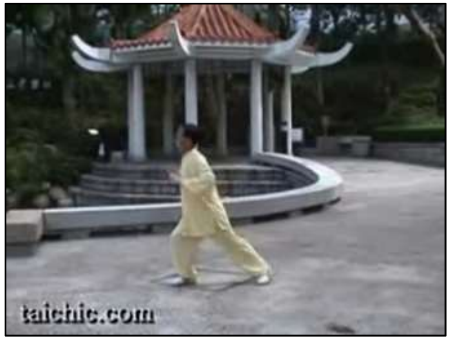
ImageNet



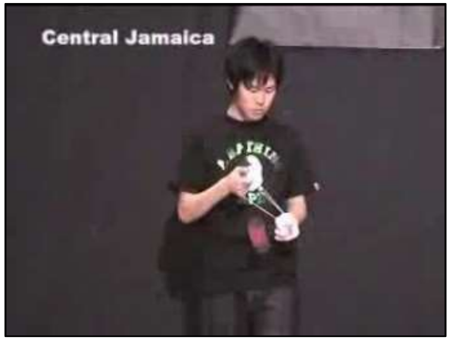


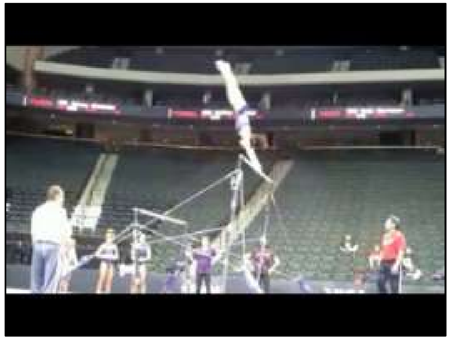
Shuffle & Learn





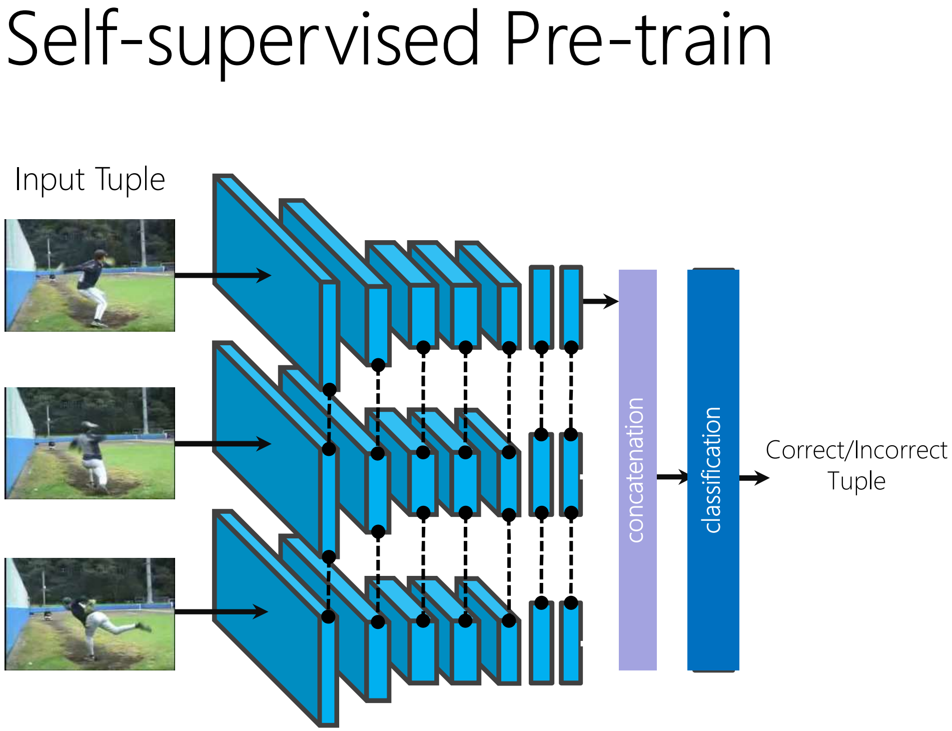
Random

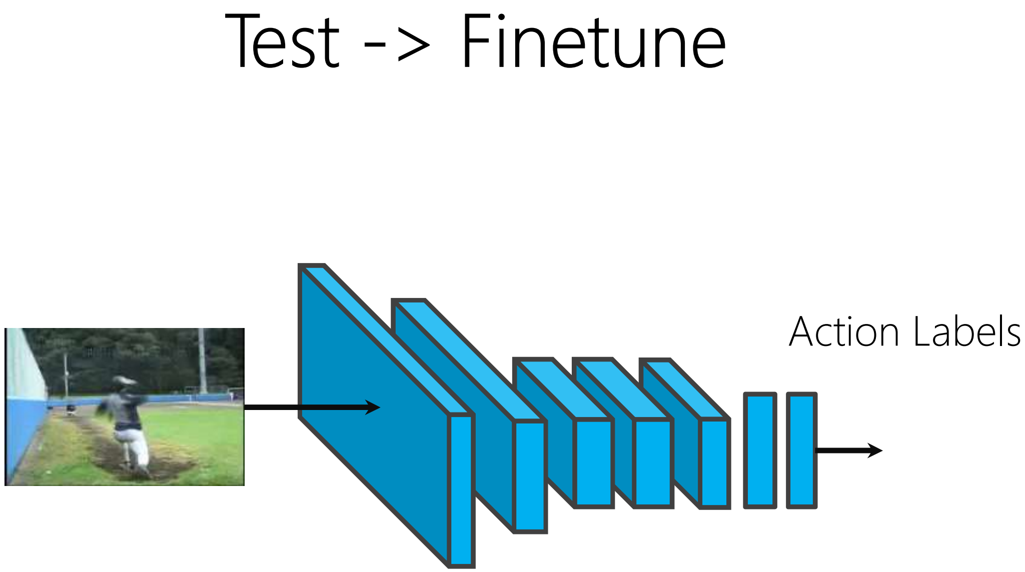




Slide credit: Ishan Misra

Finetuning setup





Slide credit: Ishan Misra

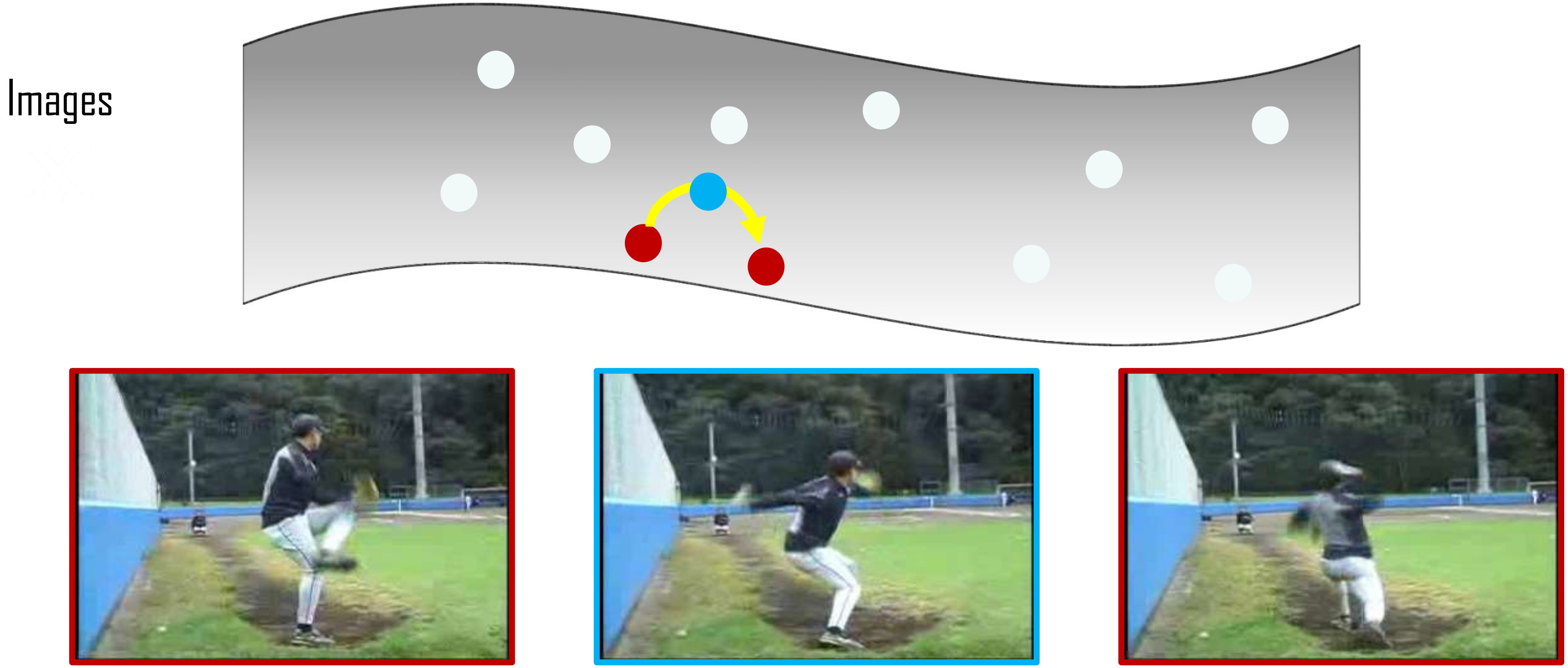
Results: Finetune on Action Recognition

Dataset	Initialization	Mean Classification Accuracy
UCF101	Random	38.6
	Shuffle & Learn	50.2
	ImageNet pre-trained	**67.1**

Setup from - Simonyan & Zisserman, 2014

Slide credit: Ishan Misra

What does the network learn?



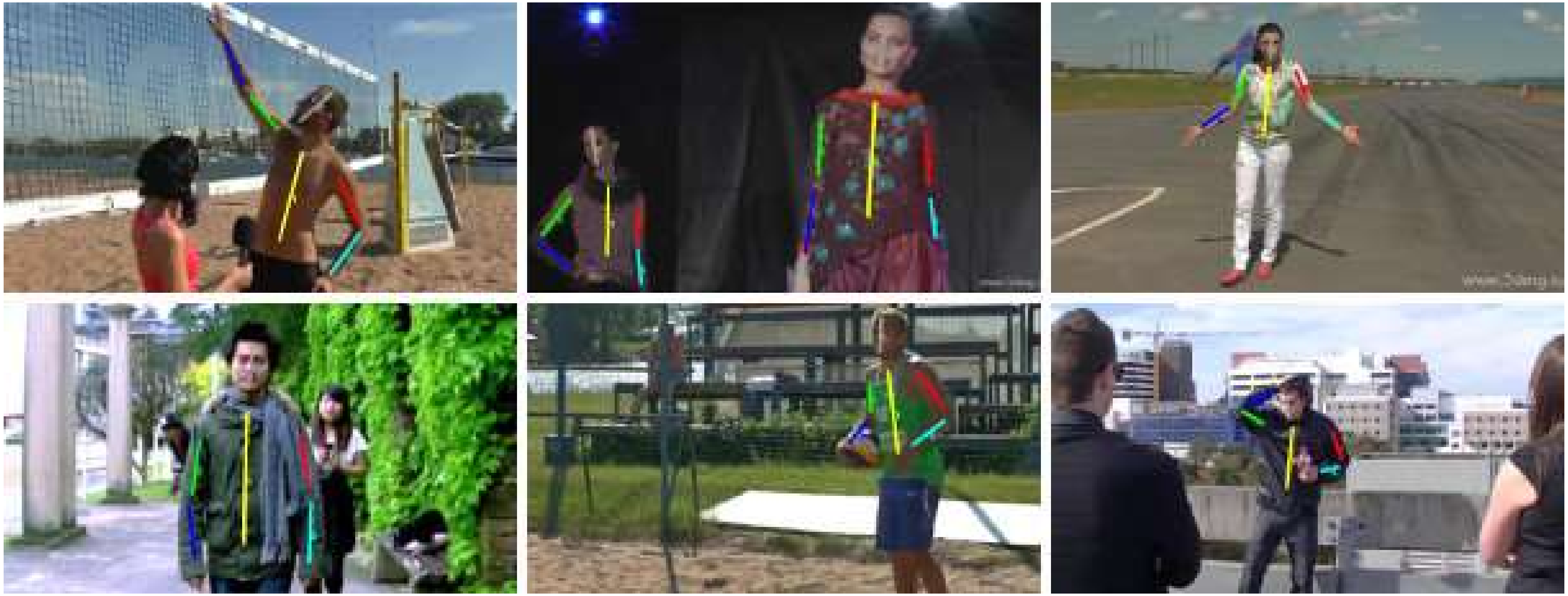
Given a start and an end, can this point lie in between?

Shuffle and Learn – I. Misra, L. Zitnick, M. Hebert – ECCV 2016

Slide credit: Ishan Misra

Human Pose Estimation

- Keypoint estimation using FLIC and MPII Datasets



Slide credit: Ishan Misra

Human Pose Estimation

- Keypoint estimation using FLIC and MPII Datasets

	FLIC Dataset		**MPII Dataset**	
Initialization	Mean PCK	AUC PCK	Mean PCKh@0.5	AUC PCKh@0.5
Shuffle & Learn	84.9	49.6	**87.7**	**47.6**
ImageNet pre-train	**85.8**	**51.3**	85.1	47.2

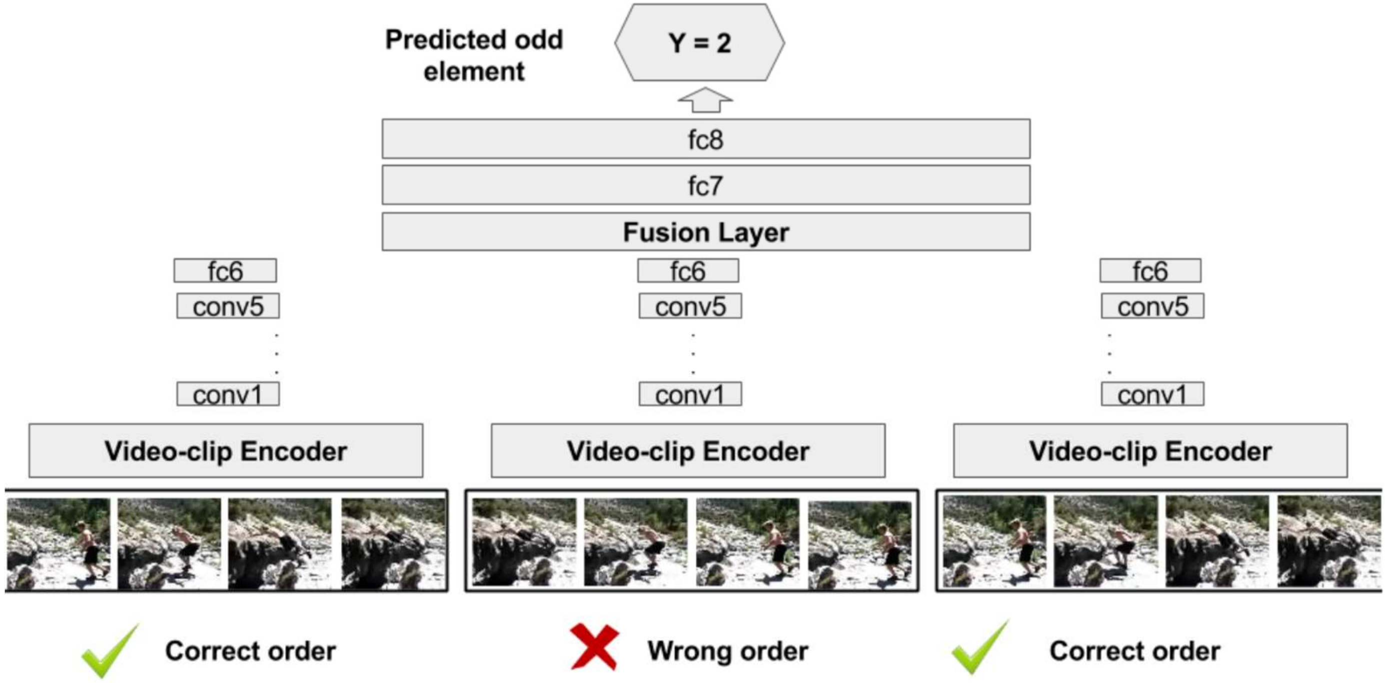
FLIC - Sapp & Taskar, 2013
MPII - Andriluka et al., 2014
Setup fom – Toshev et al., 2013

Slide credit: Ishan Misra

More temporal structure in videos

Self-Supervised Video Representation Learning With **Odd-One-Out Networks**

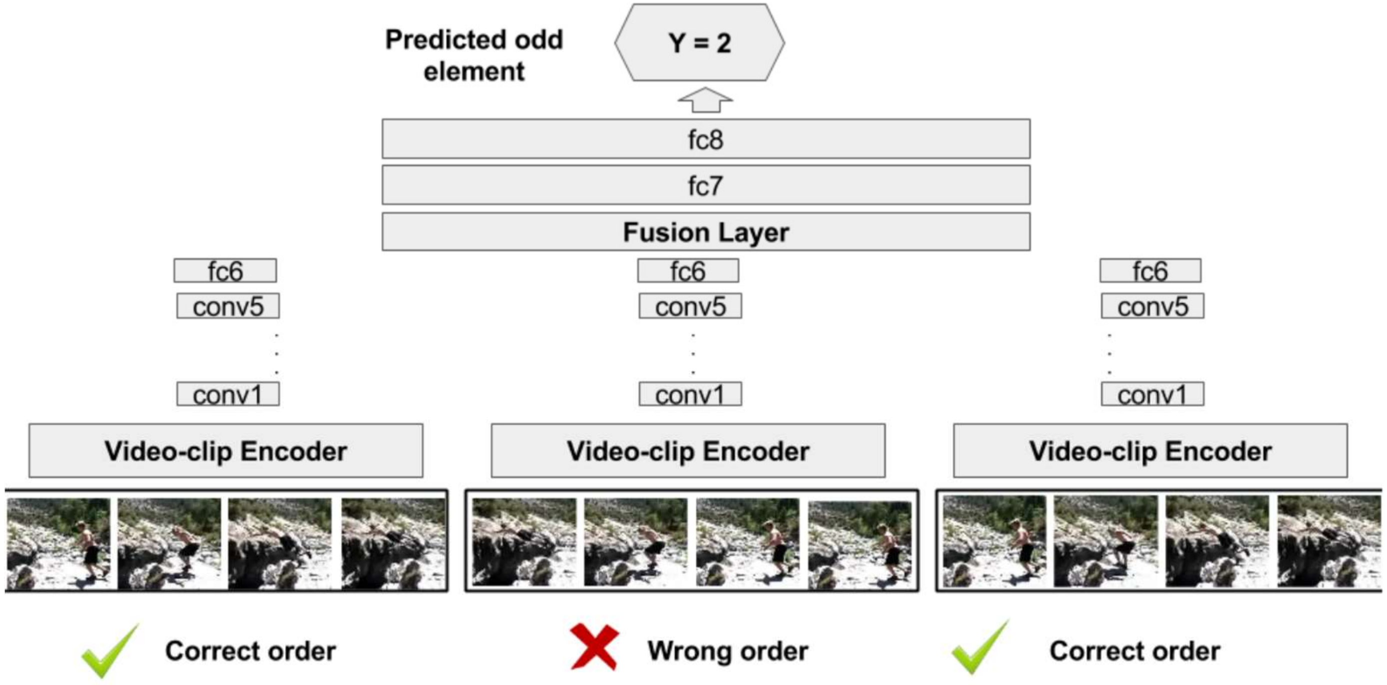
Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould, ICCV 2017



More temporal structure in videos

Self-Supervised Video Representation Learning With **Odd-One-Out Networks**

Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould, ICCV 2017



Initialization	Mean Classification Accuracy
Random	38.6
Shuffle and Learn	50.2
Odd-One-Out	60.3
ImageNet pre-trained	**67.1**

Summary: lessons so far

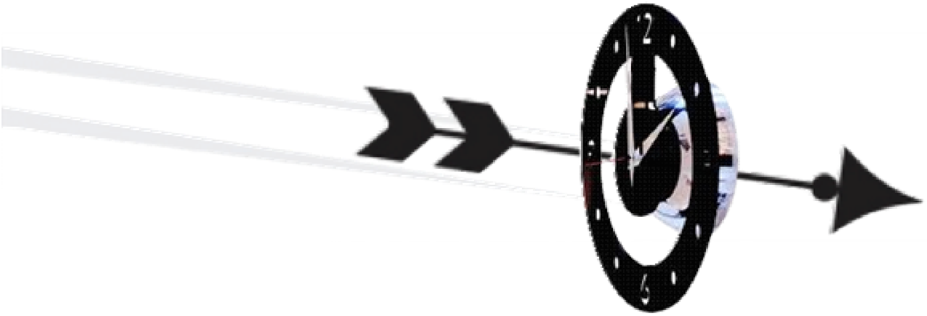
- Important to select informative data in training
  - Hard negatives and positives
  - Otherwise, most data is too easy or has no information and the network will not learn
  - Often use heuristics for this, e.g. motion energy
- Consider how the network can possibly solve the task (without cheating)
  - This determines what it must learn, e.g. human keypoints in `shuffle and learn’
- Choose the proxy task to encourage learning the features of interest

Self-Supervision using the Arrow of Time

Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

Learning the arrow of time

Task: predict if video playing forwards or backwards



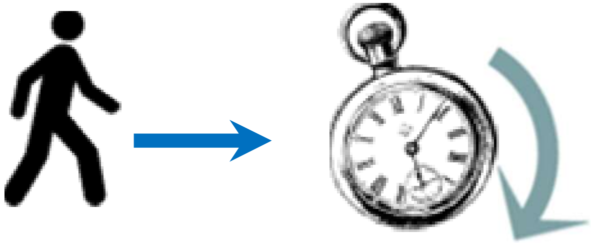
Supervision:

Positive training samples: video clips playing forwards

Negative training samples: video clips playing backwards

Strong cues

Semantic, face motion direction, ordering



Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

Strong cues

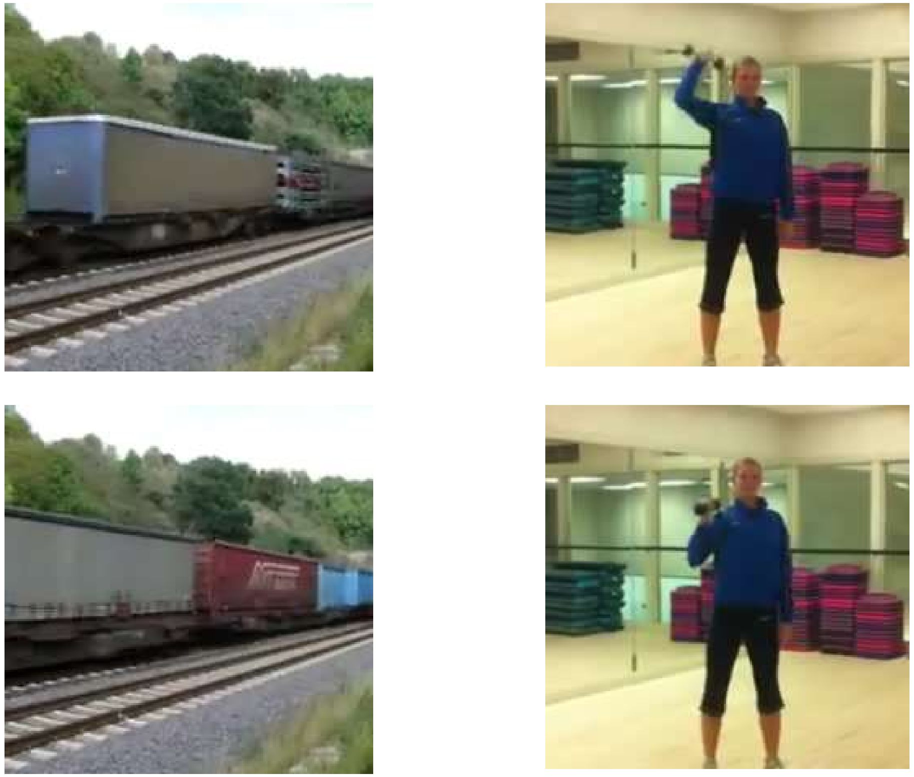
`Simple' physics:

- gravity
- entropy
- friction
- causality

Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

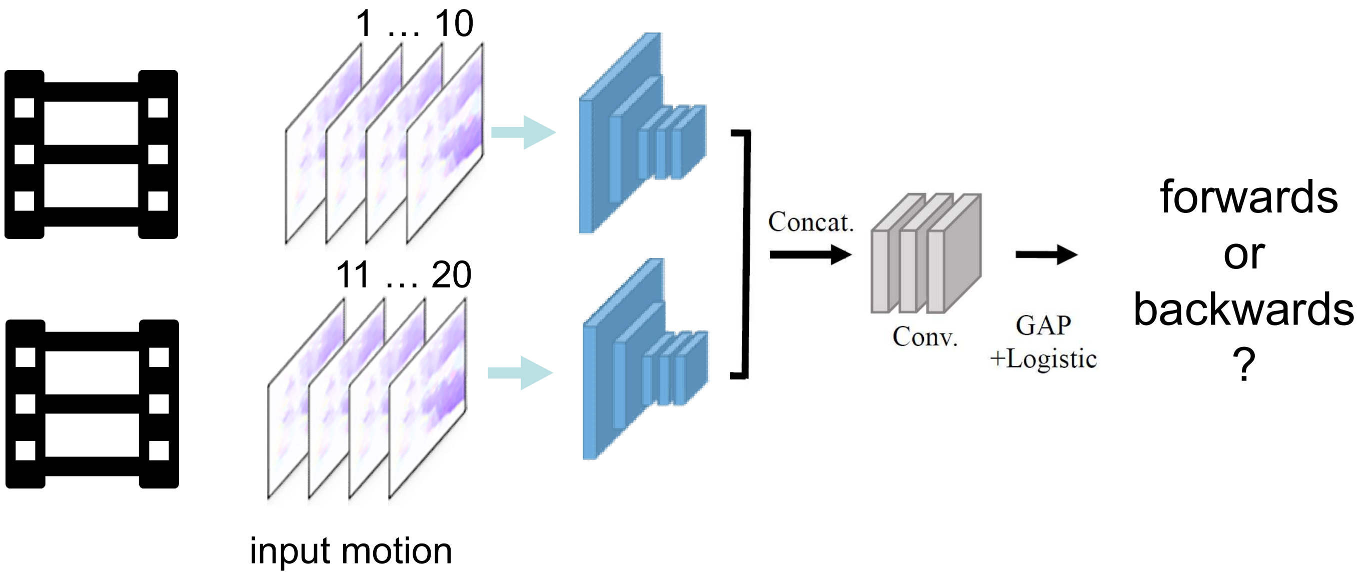
Weak or no cues

Symmetric in time, constant motion, repetitions



Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

Temporal Class-Activation Map Network



T-CAM Model:

Input: optical flow in two chunks

Final layer: global average pooling to allow class activation map (CAM)

The inevitable cheating …

Cautionary tale:

Chromatic aberration used as shortcut in Doersch C, Gupta A, Efros AA,
Unsupervised visual representation learning by context prediction.
ICCV 2015

Dataset: UCF-101 actions

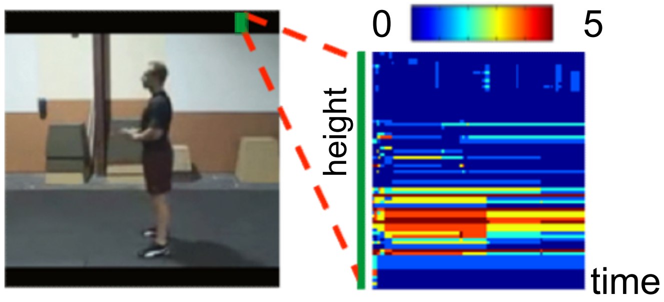
Train/Test: 70%/30%

AoT Test accuracy: 98%

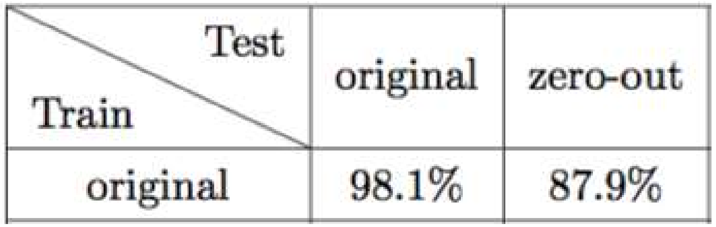
Chance accuracy: 50%

Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

Cue I: black framing



black stripes are not “purely black”



Train \ Test	original	zero-out
original	98.1%	87.9%

when black stripe signals are zeroed-out, test accuracy ***drops ~10%***

46% of videos have black framing

Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

Cue II: cinematic conventions

K-means clustering on test clips with top scores

cluster A
(camera zoom-in)

cluster B
(camera tilt-down)

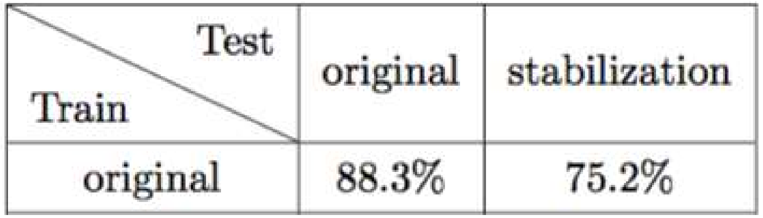
73% of videos have camera motion

Stabilize to remove camera motion/zoom

original

camera stabilized

(black stripe removed)



Test Train	original	stabilization
original	88.3%	75.2%

when camera motion is stabilized, test accuracy ***drops ~10%***

Datasets and Performance

Flickr 150K shots

- Obtained from 1.74M shots used in Thomee et al (2016) & Vondrick et al (2016), after black stripe removal and stabilization

- Split 70:30 for train:test

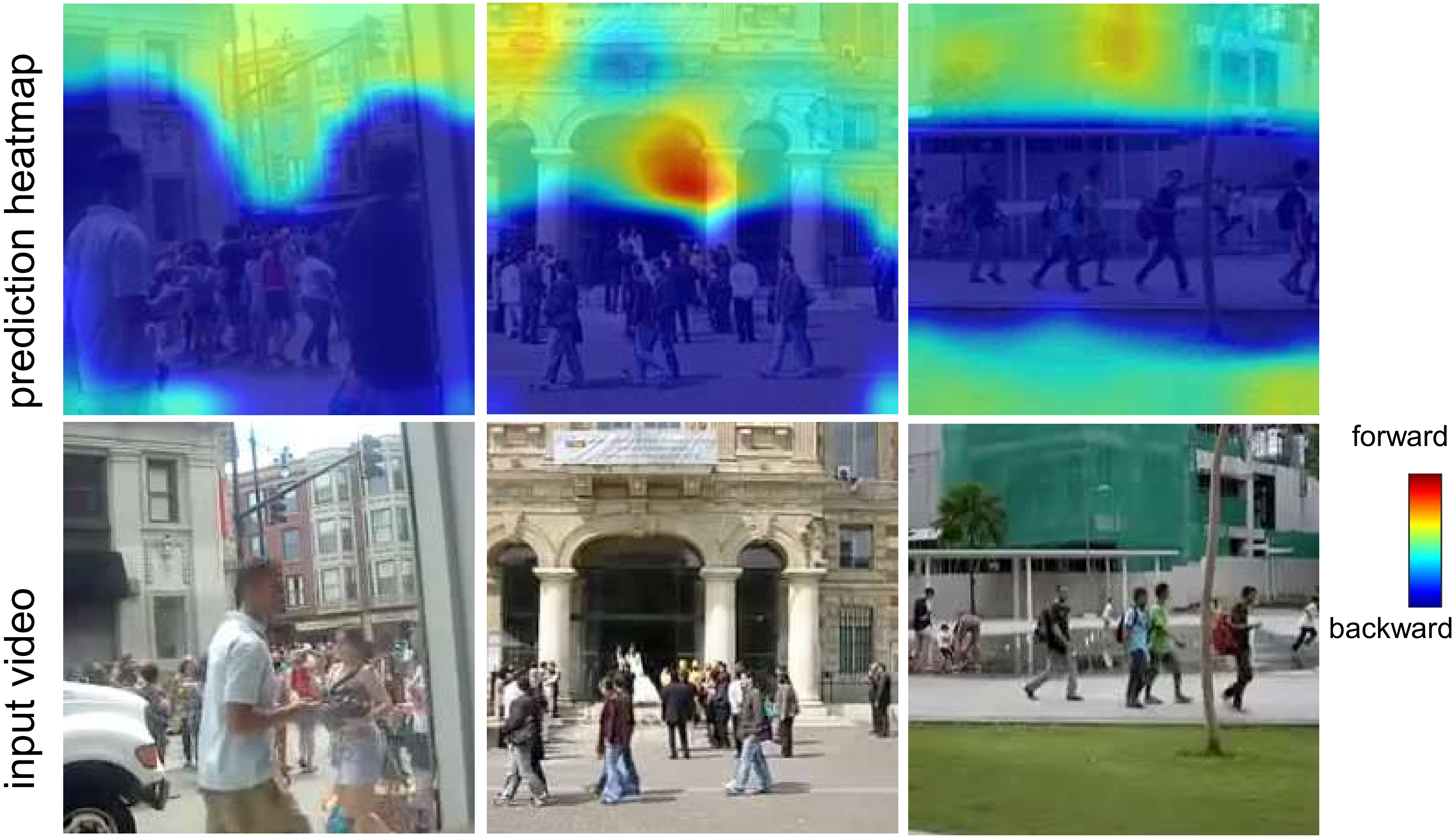
Model accuracy on test set: 81%

Human accuracy on test set: 81%

Chance: 50%

Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

“Semantic” motions



Evaluation: Action Classification

Procedure:

- Pre-train network
- Fine tune & test network on UCF101 human action classification benchmark

Pre-train	Performance
T-CAM on AoT on Flickr 150k shots	84.1
T-CAM on AoT on UCF-101	**86.3**
Flow network on ImageNet*	85.7

- * = Wang et al, Temporal Segment Networks, 2016 (also VGG-16 and flow, pre-trained on ImageNet)

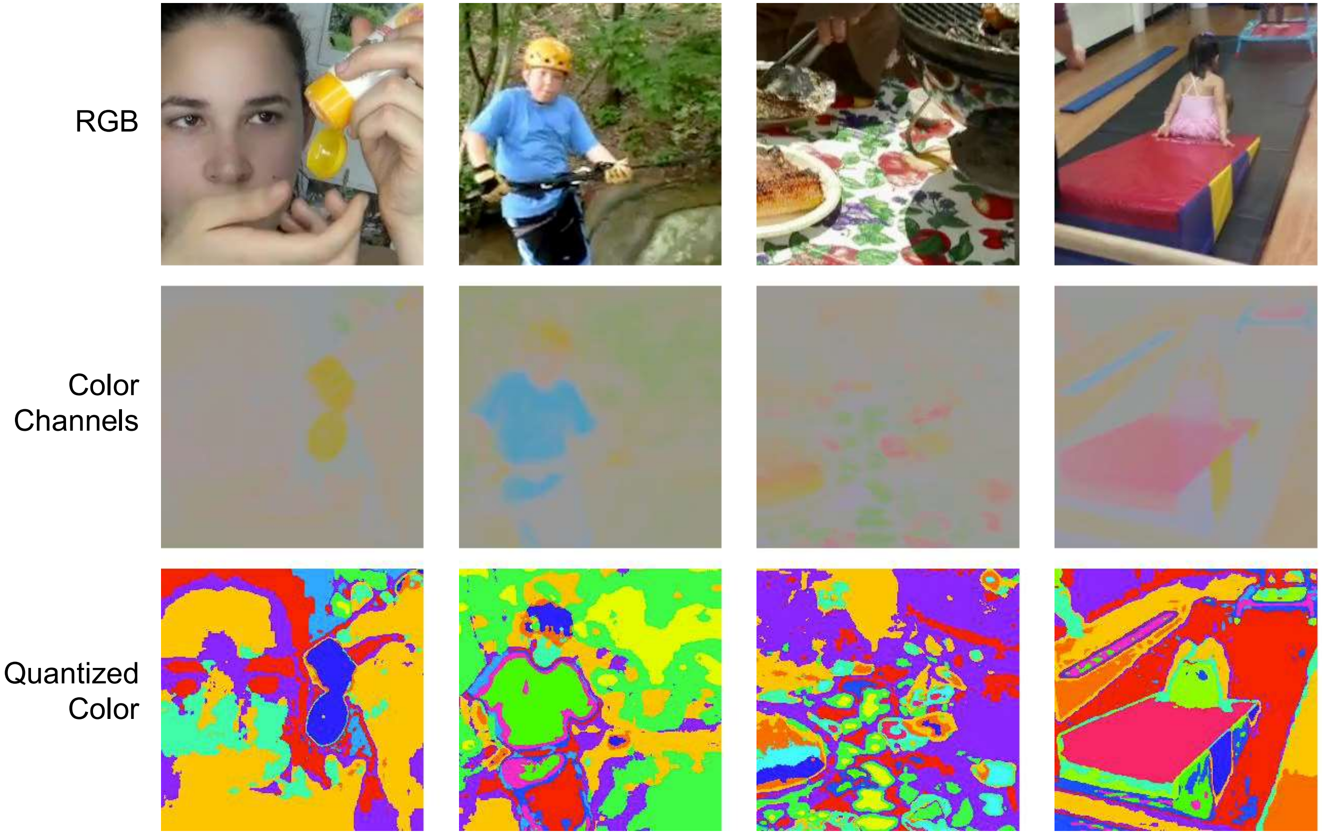
Donglai Wei, Joseph Lim, Bill Freeman, Andrew Zisserman CVPR 2018

Tracking Emerges by Colorizing Videos

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy, ECCV 2018

Color is <u>mostly</u> temporally coherent

Temporal Coherence of Color



Self-supervised Tracking

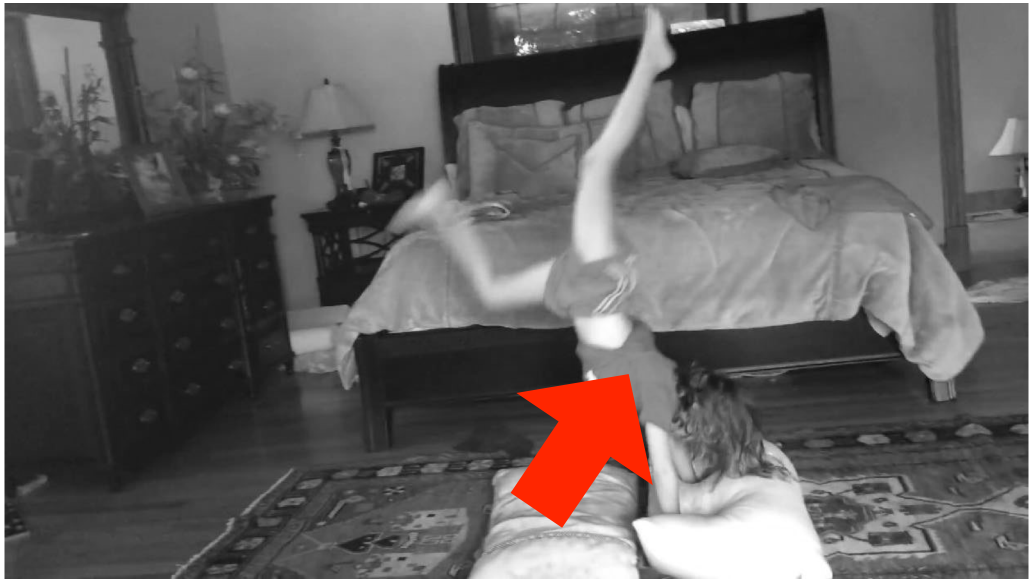
Task: given a color video …
Colorize all frames of a gray scale version using a reference frame

Reference Frame

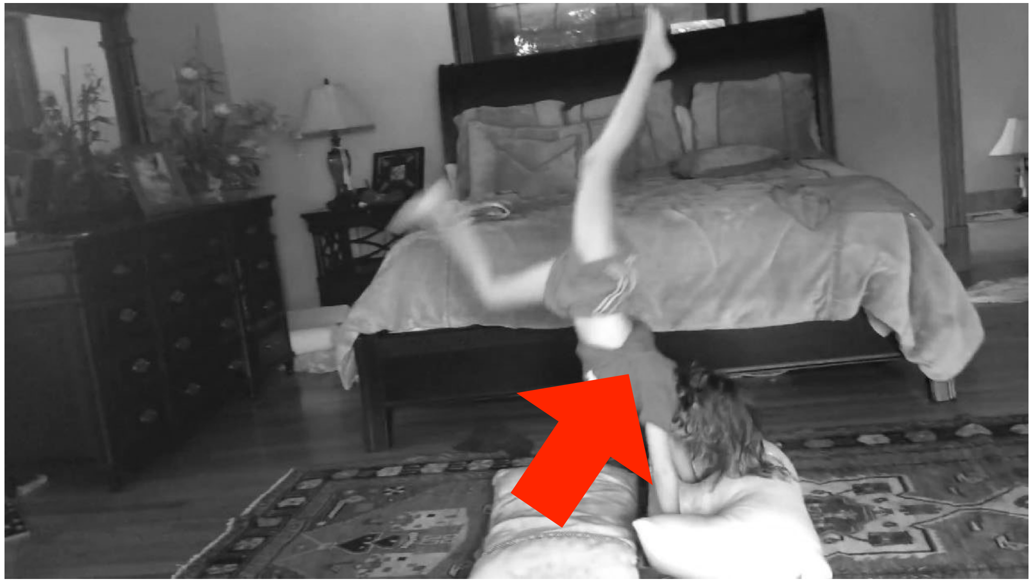
Gray-scale Video

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

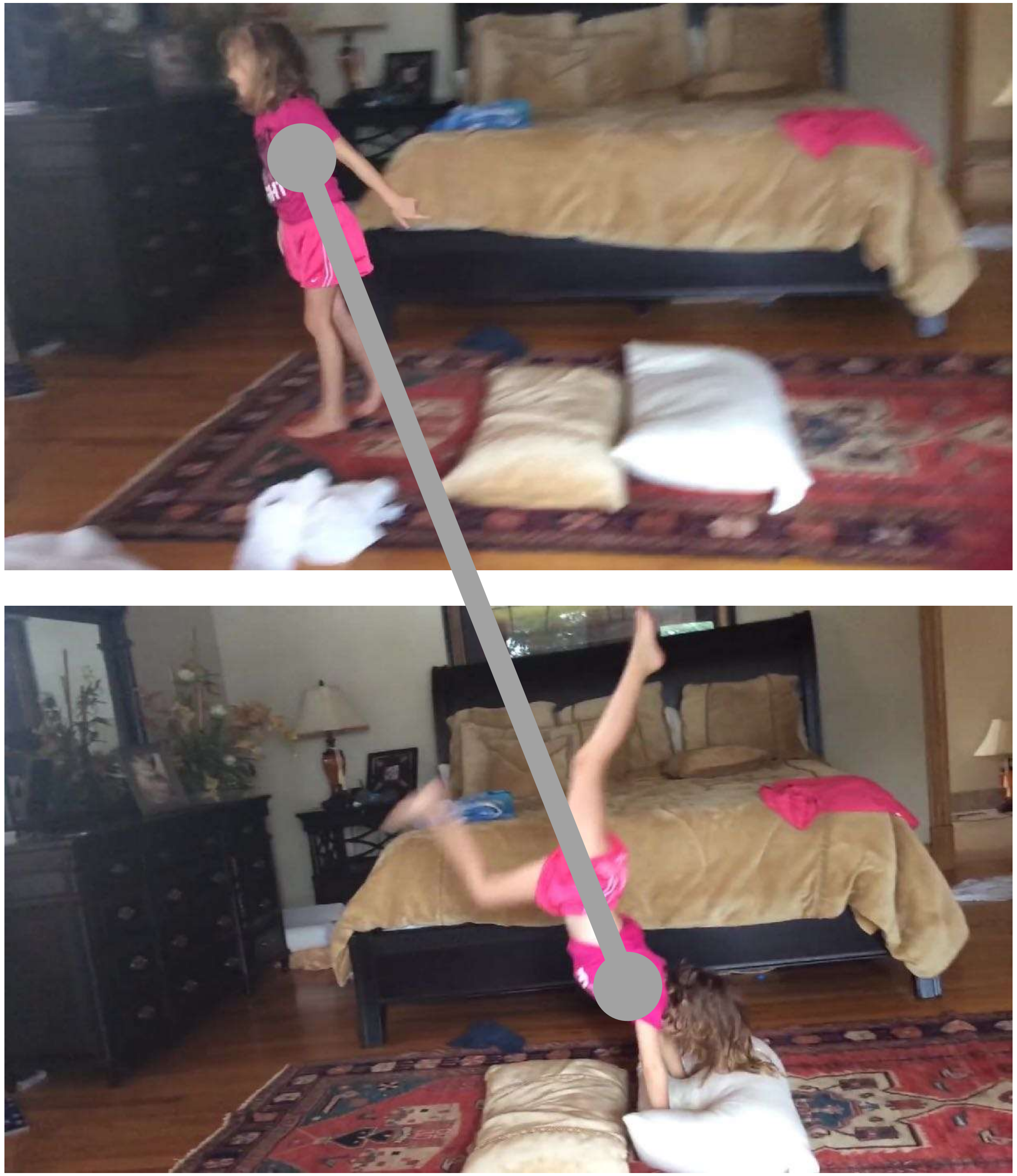
What color is this?



Where to copy color from?



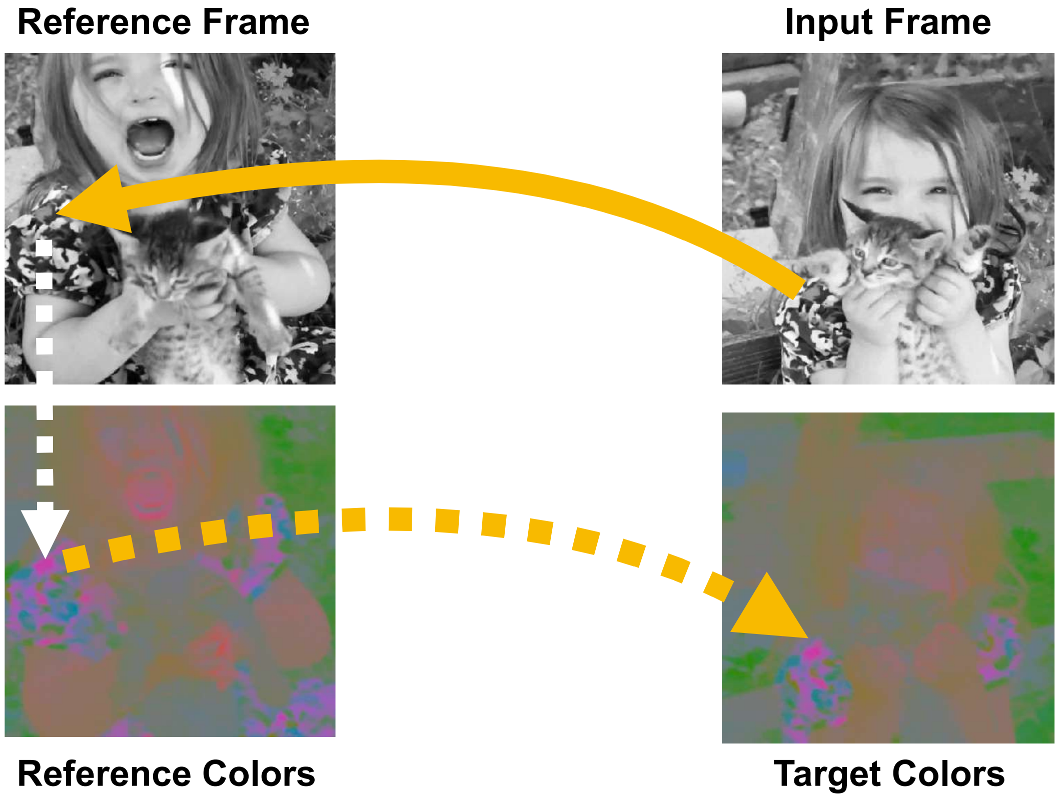
Semantic correspondence



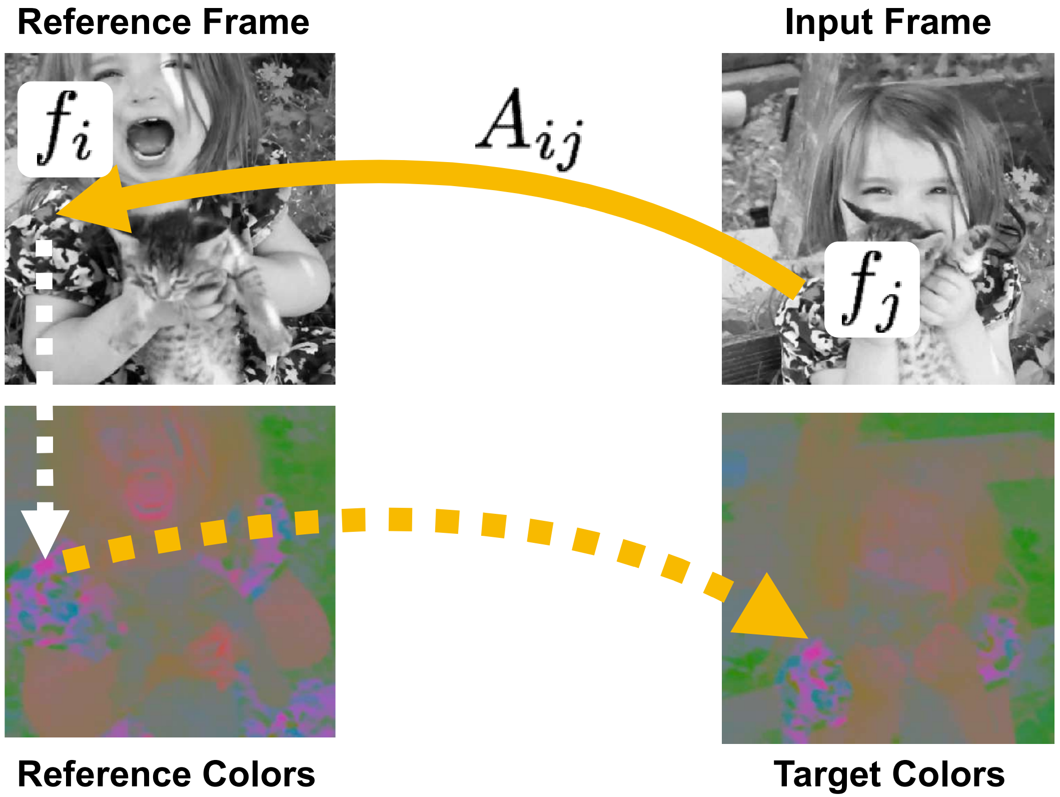
Input Frame

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

Colorize by Pointing

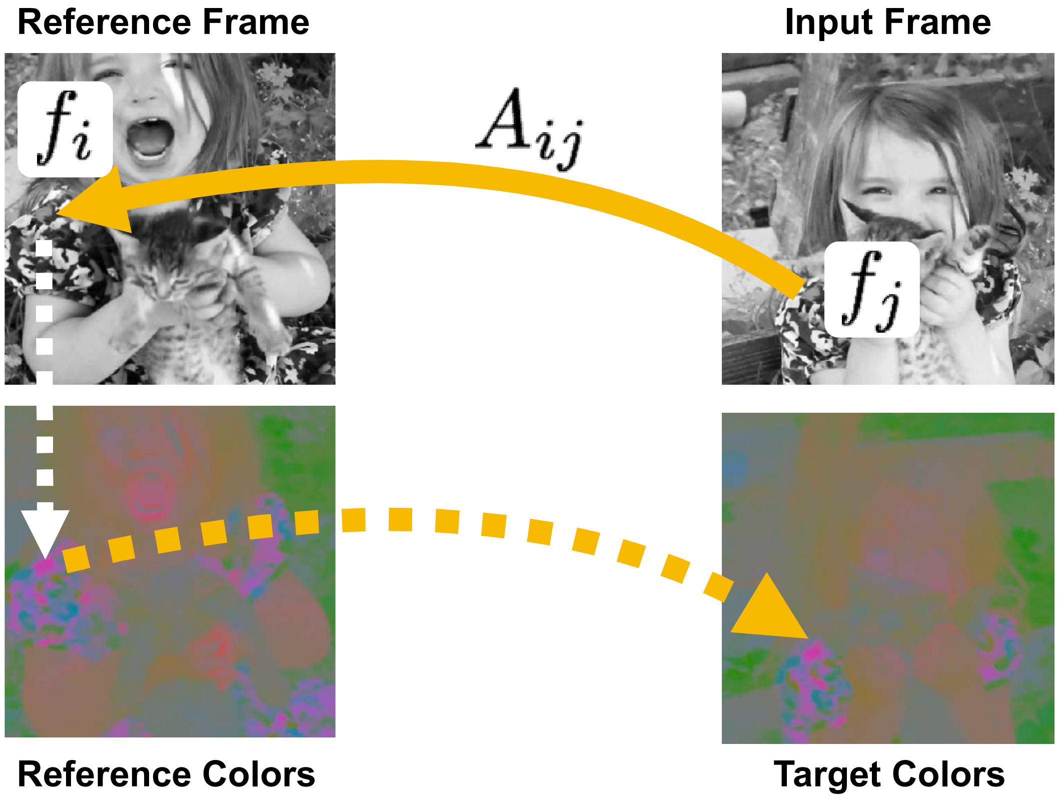


Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.



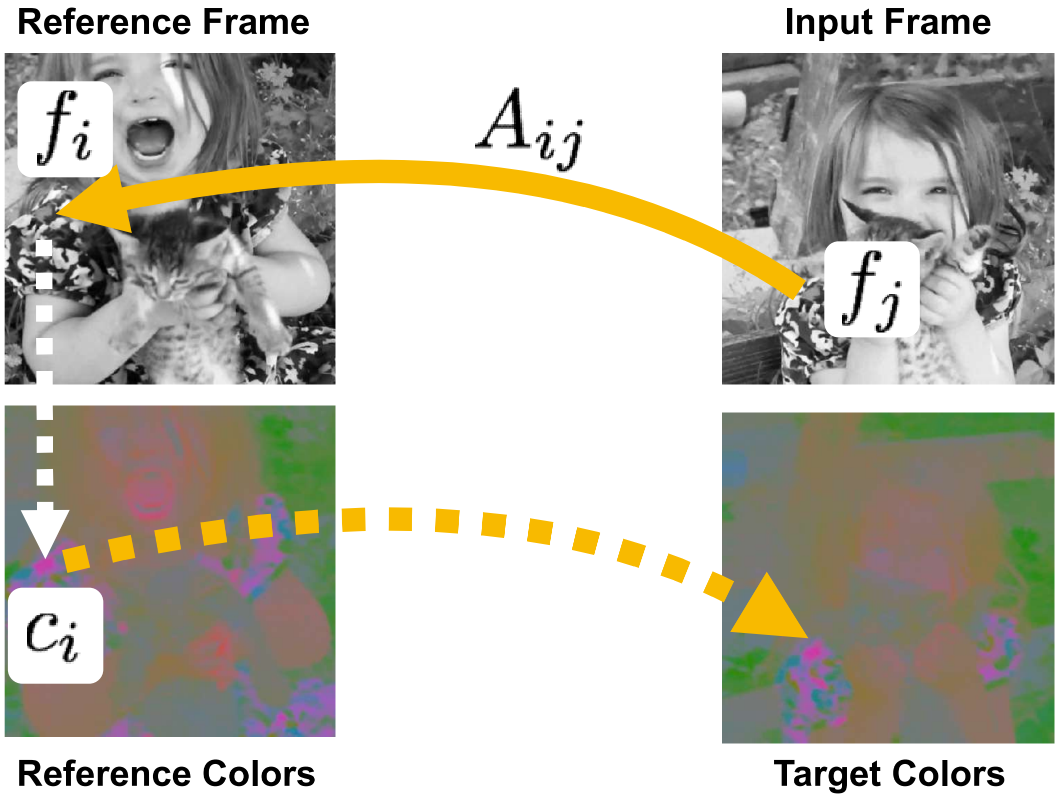
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

$$A_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_k^T f_j\right)}$$



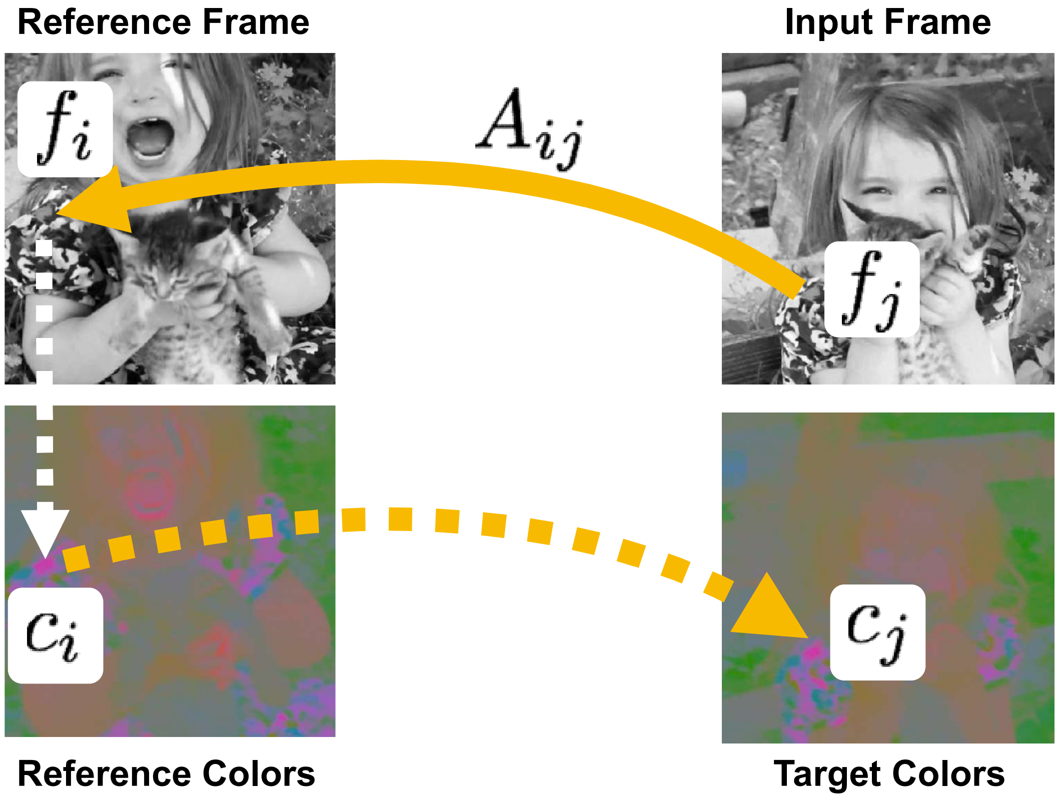
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

$$\hat{c}_j = \sum_i A_{ij} c_i \quad \text{where } A_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_k^T f_j\right)}$$



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

$$\min_f \mathcal{L}\left(c_j, \sum_i A_{ij} c_i\right) \text{ where } A_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_k^T f_j\right)}$$

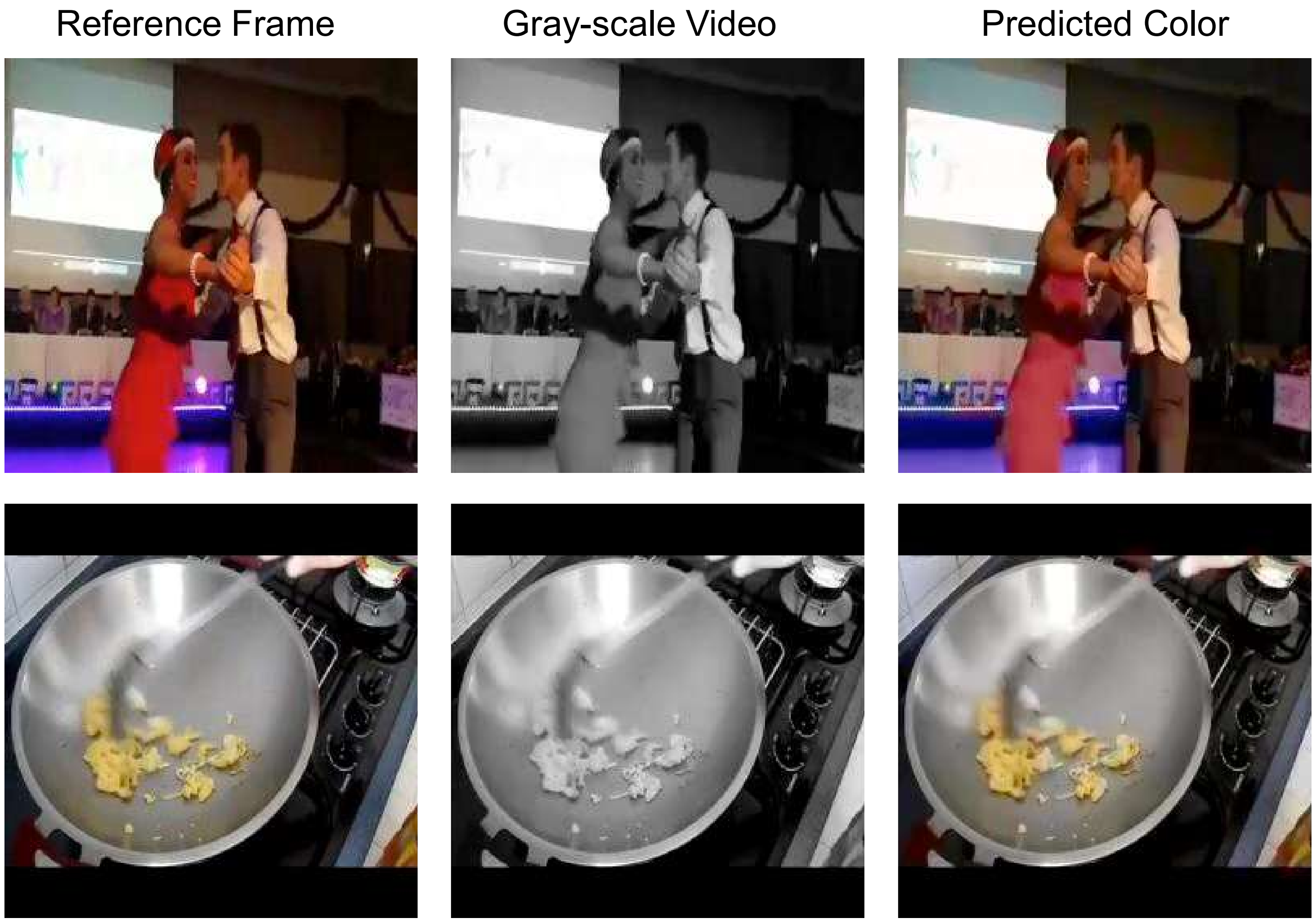


Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

Video Colorization

Train: Kinetics

Evaluate: DAVIS



Reference Frame

Gray-scale Video

Predicted Color

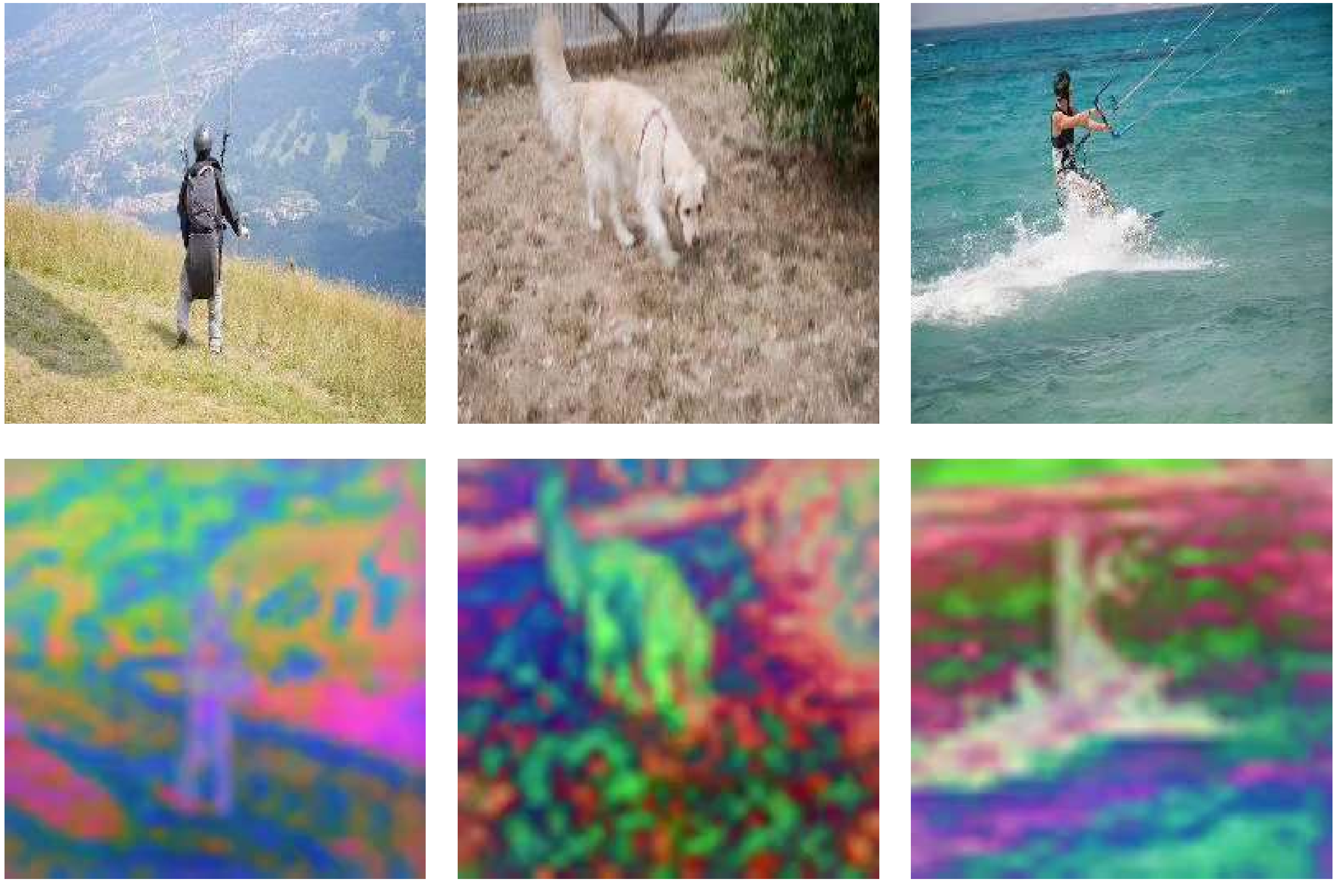
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

Visualizing Embeddings

Project embedding to 3 dimensions and visualize as RGB

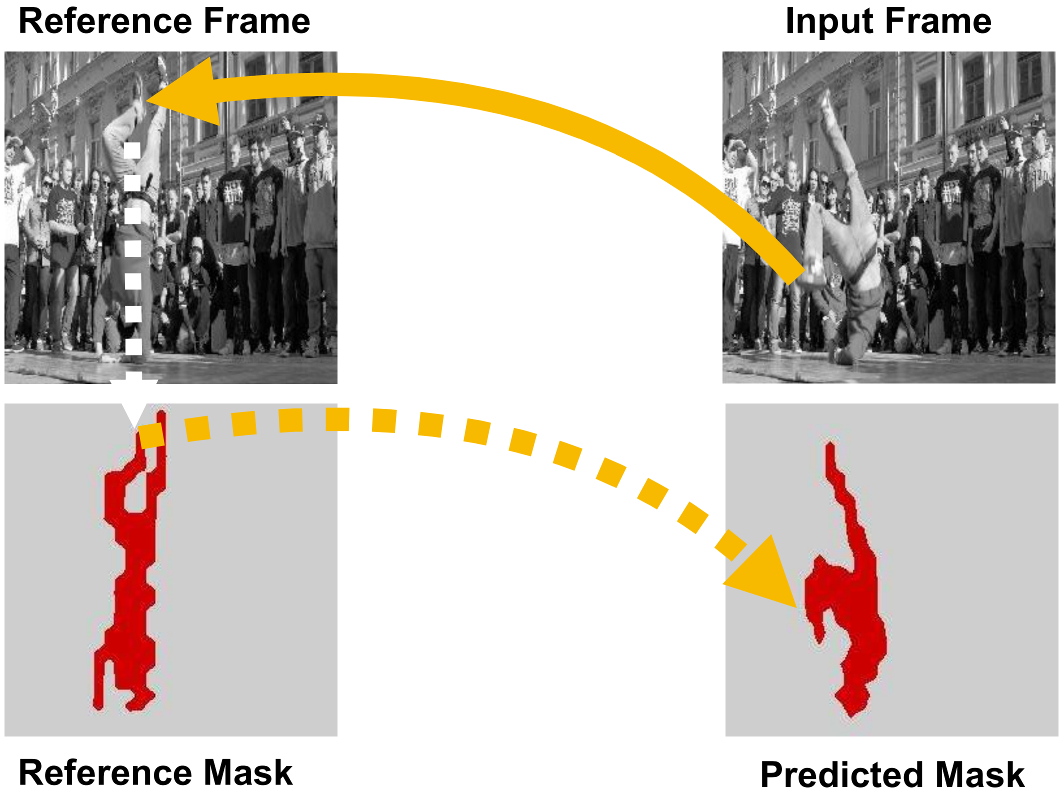
Train: Kinetics

Evaluate: DAVIS



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

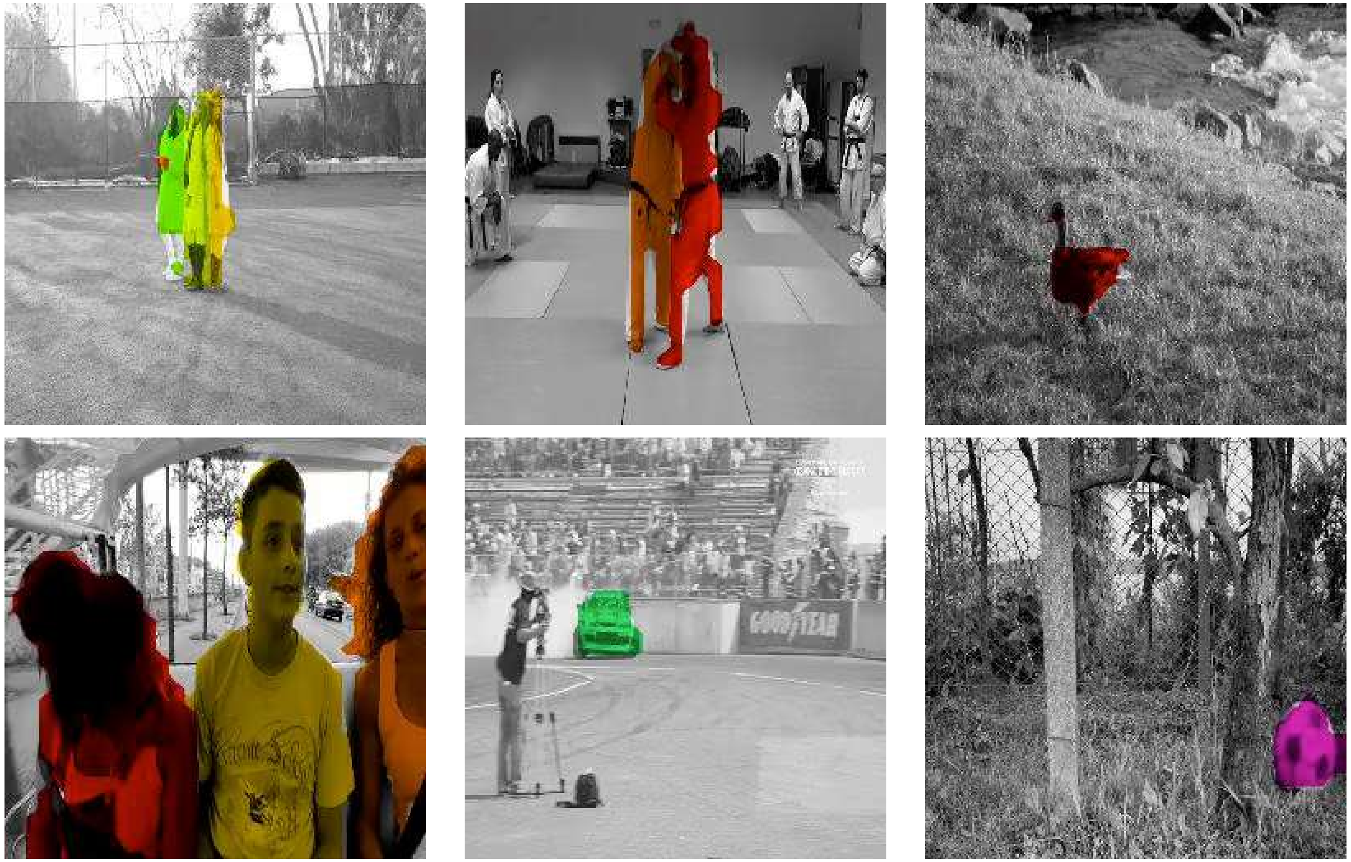
Tracking Emerges!



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

Segment Tracking Results

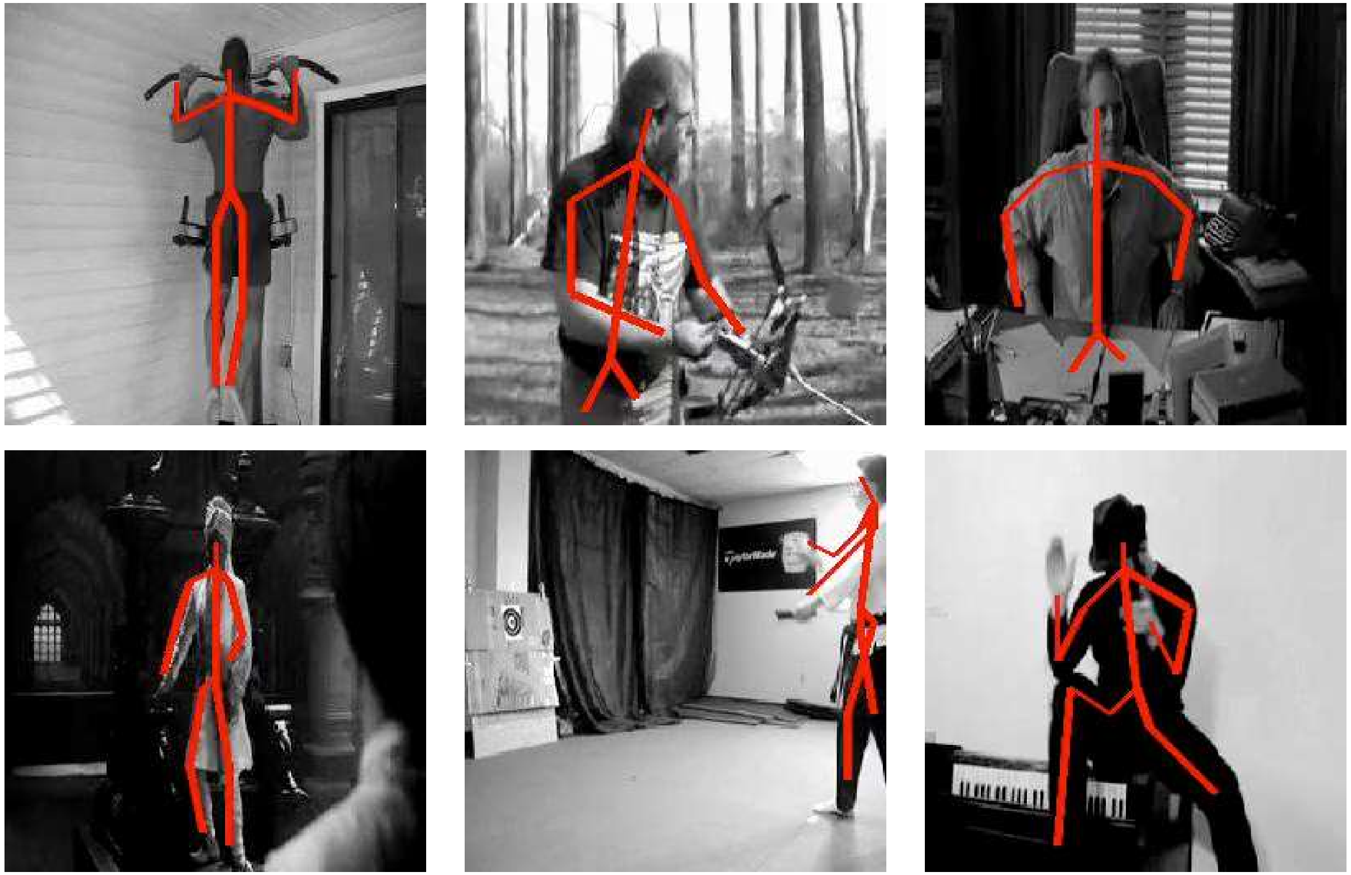
Only the first frame is given. Colors indicate different instances.



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

Pose Tracking Results

Only the skeleton in the first frame is given.

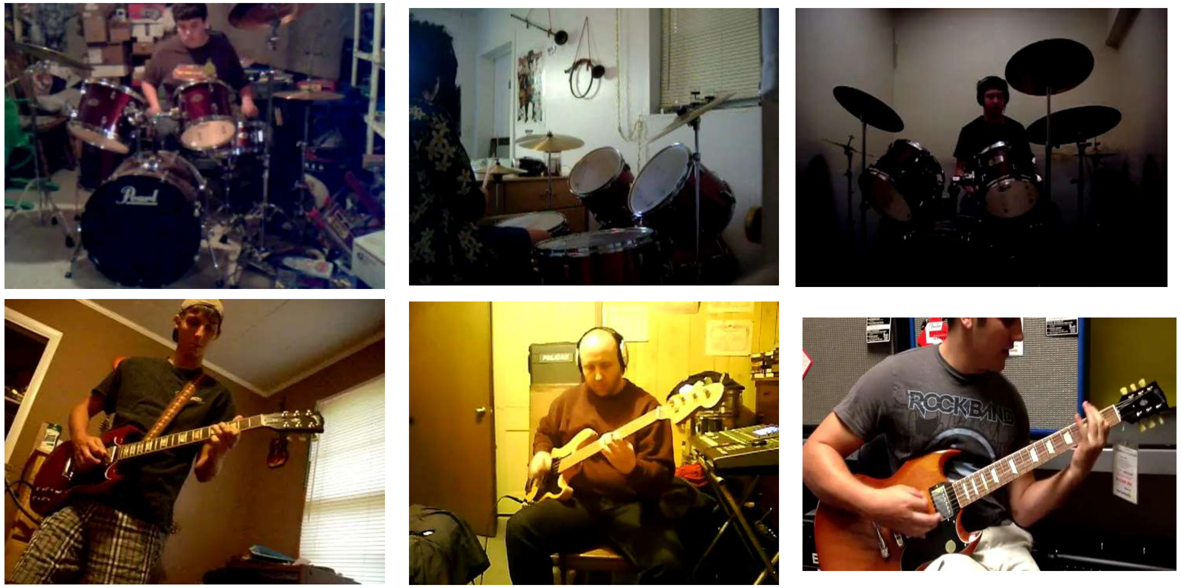


Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. ECCV 2018.

Part III

Self-Supervised Learning from Videos with Sound

Audio-Visual Co-supervision



Sound and frames are:

- Semantically consistent
- Synchronized

Audio-Visual Co-supervision

Objective: use vision and sound to learn from each other

- Two types of proxy task:
  1. Predict audio-visual **correspondence**
  2. Predict audio-visual **synchronization**

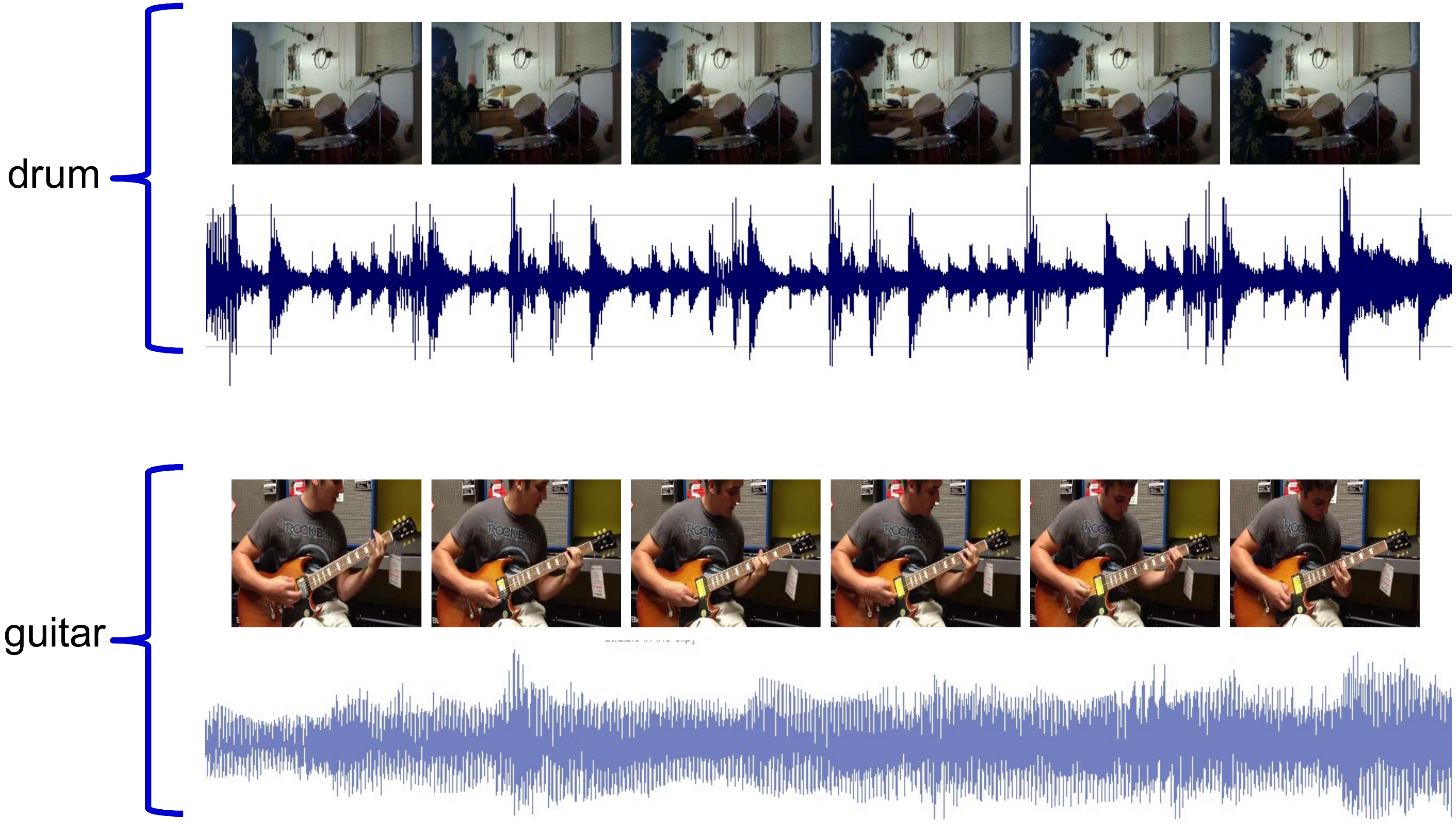
Audio-Visual Co-supervision

Train a network to predict if **image** and audio clip correspond

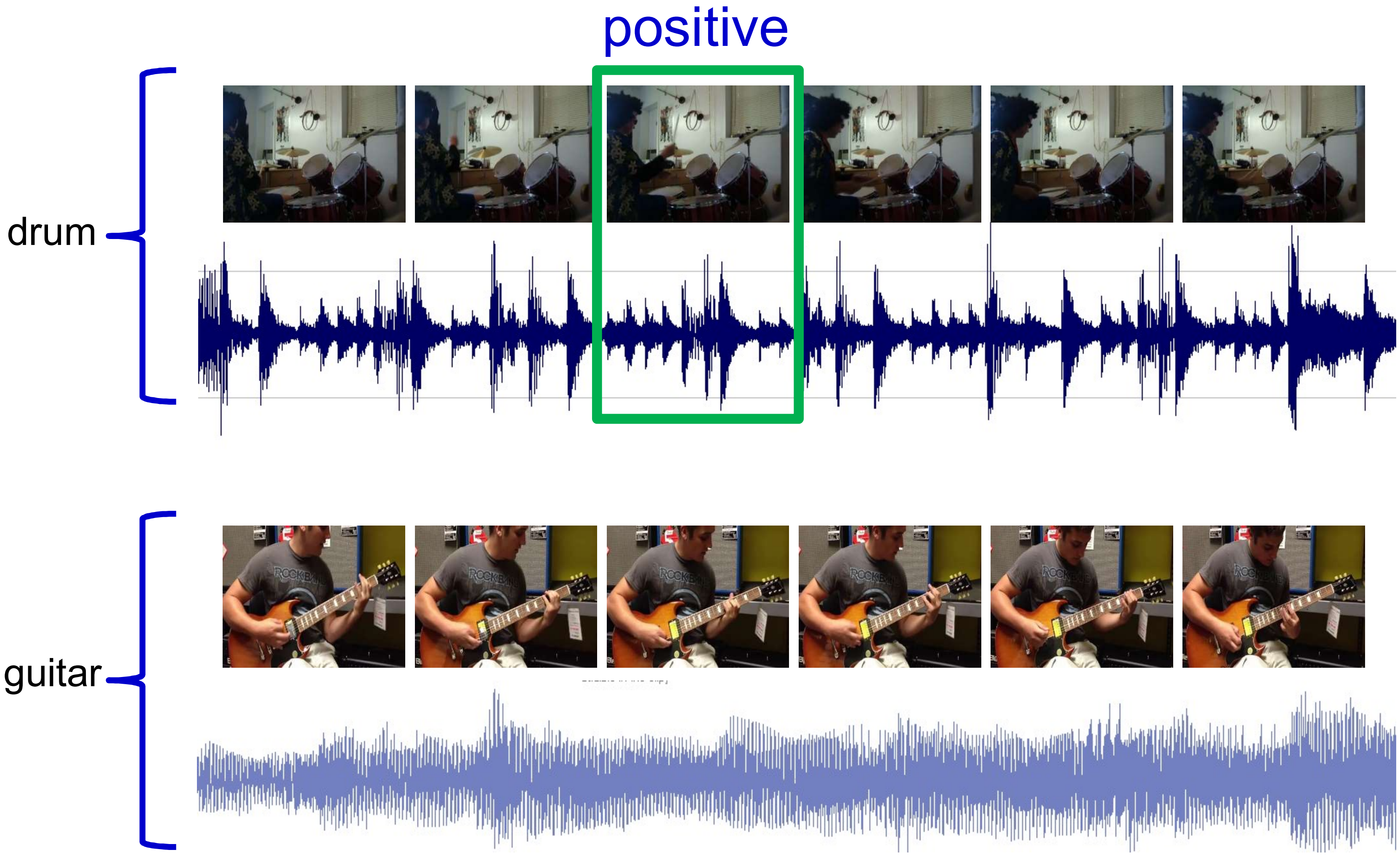
Correspond?

“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

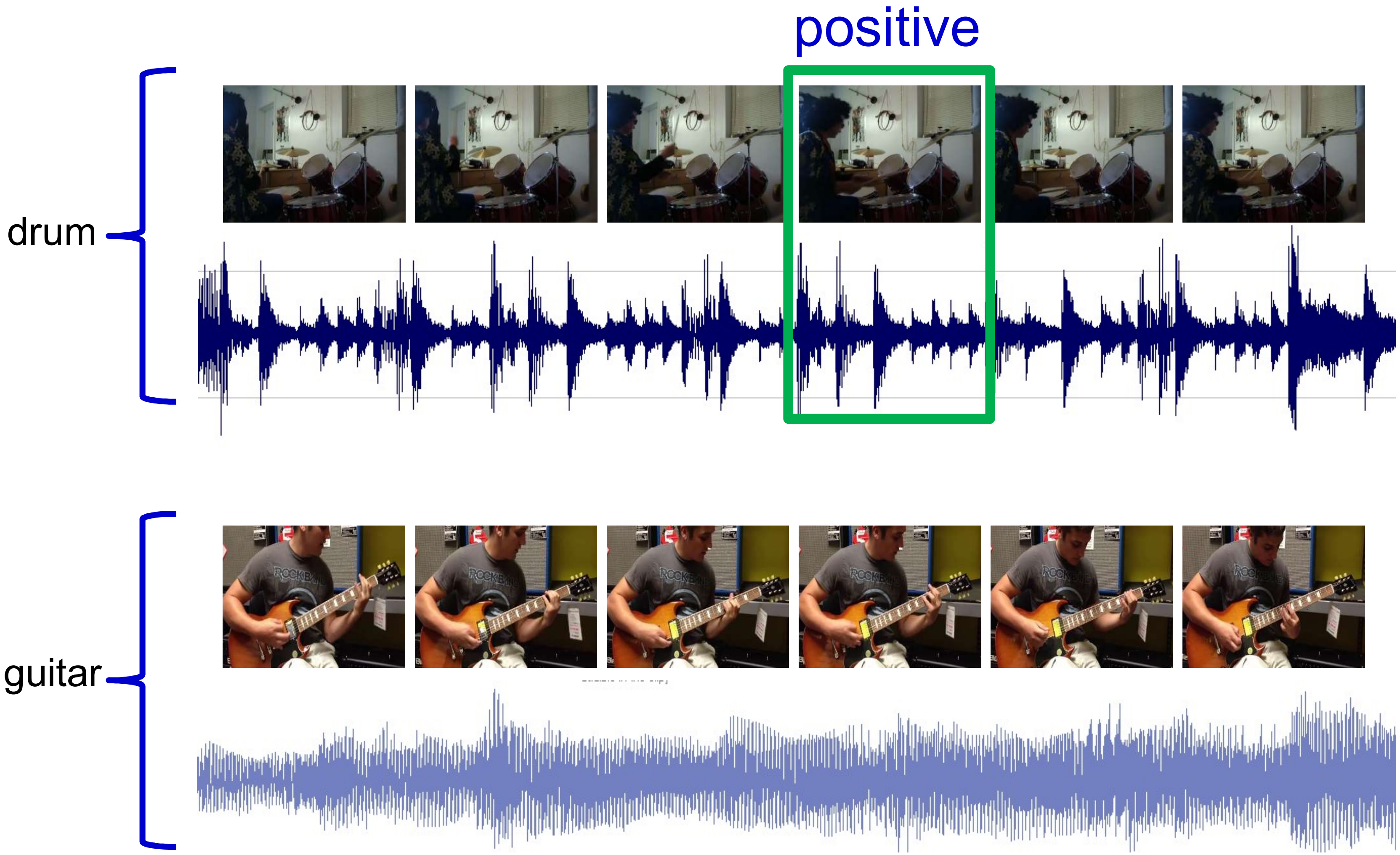
Audio-Visual Correspondence



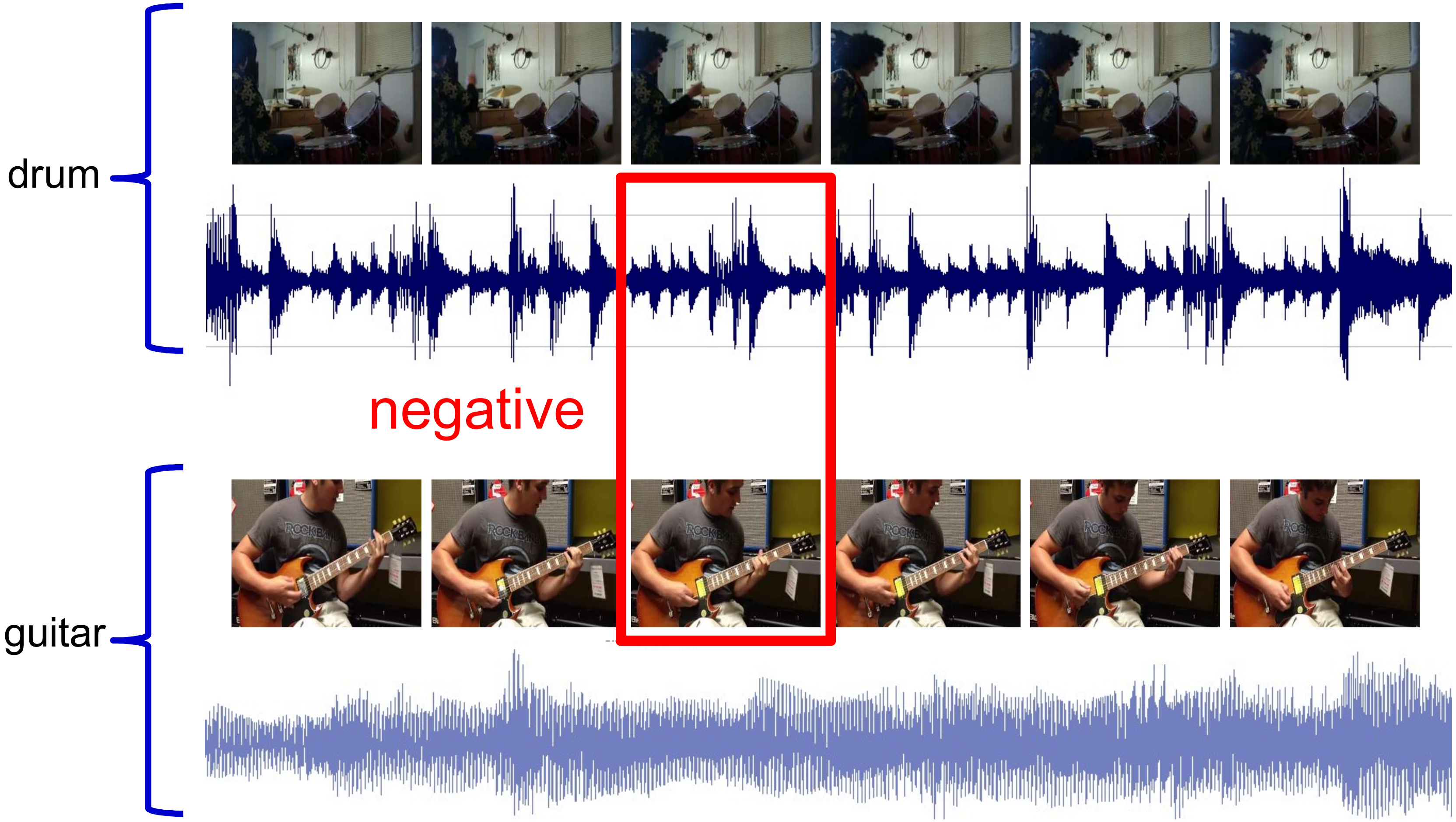
Audio-Visual Correspondence



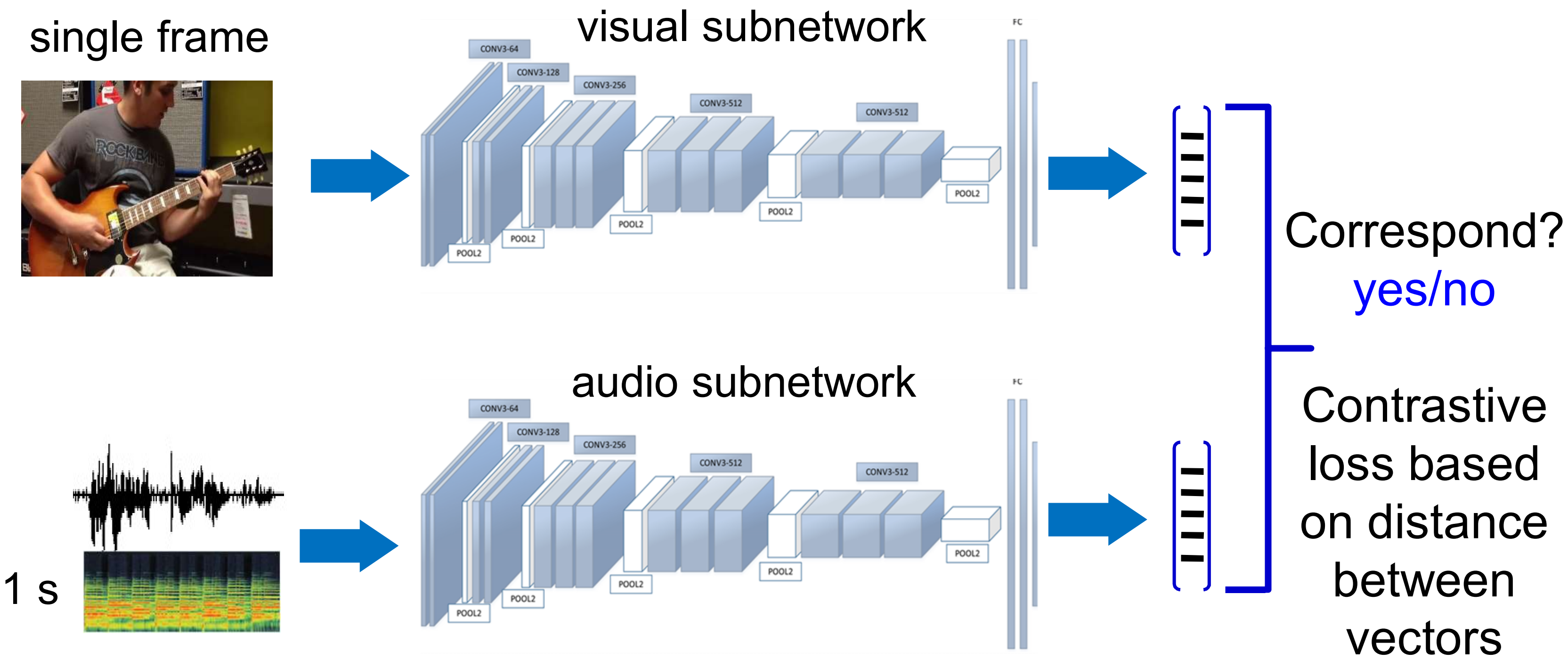
Audio-Visual Correspondence



Audio-Visual Correspondence



Audio-Visual Embedding (AVE-Net)



Correspond?
yes/no

Contrastive loss based on distance between vectors

Distance between audio and visual vectors:

- **Small:** AV from the same place in a video (**Positives**)
- **Large:** AV from different videos (**Negatives**)

Train network from scratch

Overview

What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- “What is making the sound?”
  - Learn to localize objects that sound

“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Background: Audio-Visual

- Andrew Owens ….
  - Owens, A., Jiajun, W., McDermott, J., Freeman, W., Torralba, A.: Ambient sound provides supervision for visual learning. ECCV 2016
  - Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E., Freeman,W.: Visually indicated sounds. CVPR 2016
- Other MIT work:
  - Aytar, Y., Vondrick, C., Torralba, A.: SoundNet: Learning sound representations from unlabeled video. NIPS 2016
- From the past:
  - Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. CVPR 2005
  - De Sa, V.: Learning classification from unlabelled data, NIPS 1994

Dataset

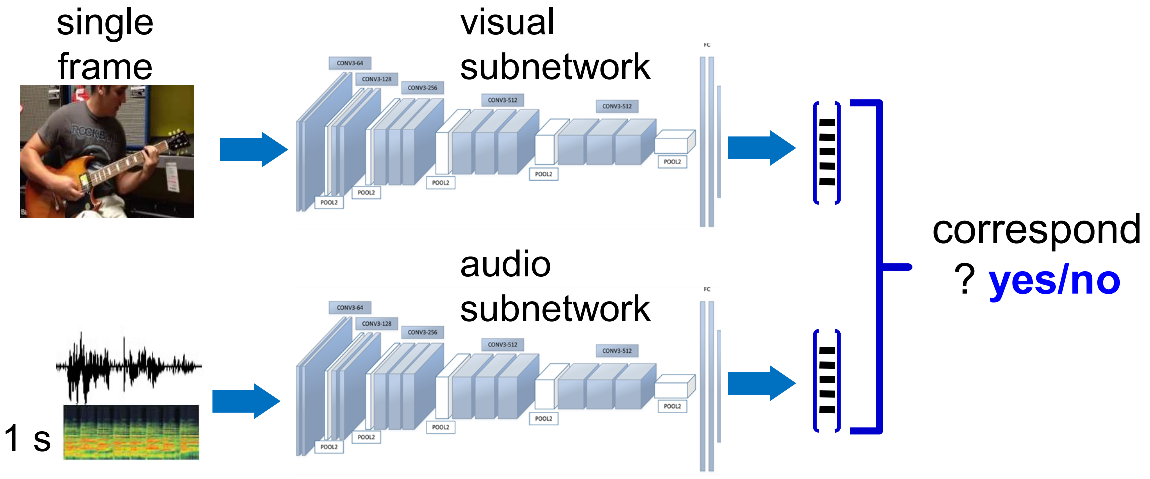
- AudioSet (from YouTube), has labels
  - 200k x 10s clips
  - use musical instruments classes
- Correspondence accuracy on test set: 82% (chance: 50%)

“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Use audio and visual features

What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- "What is making the sound?"
  - Learn to localize objects that sound



single frame

visual subnetwork

audio subnetwork

1 s

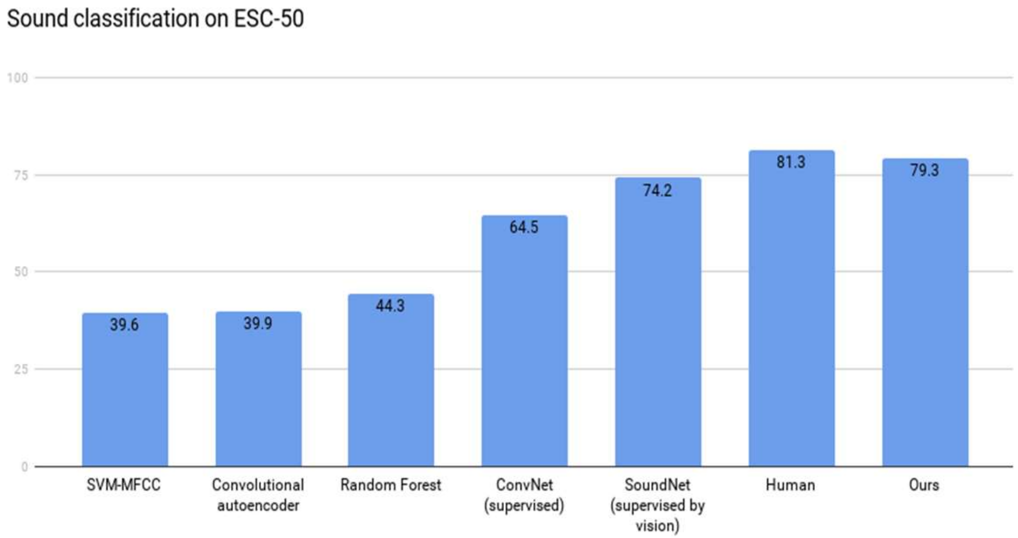
correspond ? **yes/no**

"Objects that Sound", Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Results: Audio features

Sound classification

- ESC-50 dataset
  - Environmental sound classification
  - Use the net to extract features
  - Train linear SVM

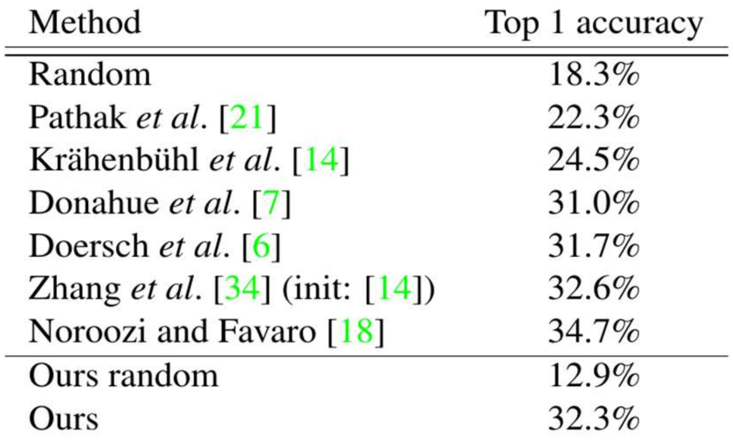


“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Results: Vision features

ImageNet classification

- Standard evaluation procedure for unsupervised / self-supervised setting
  - Use the net to extract visual features
  - Linear classification on ImageNet



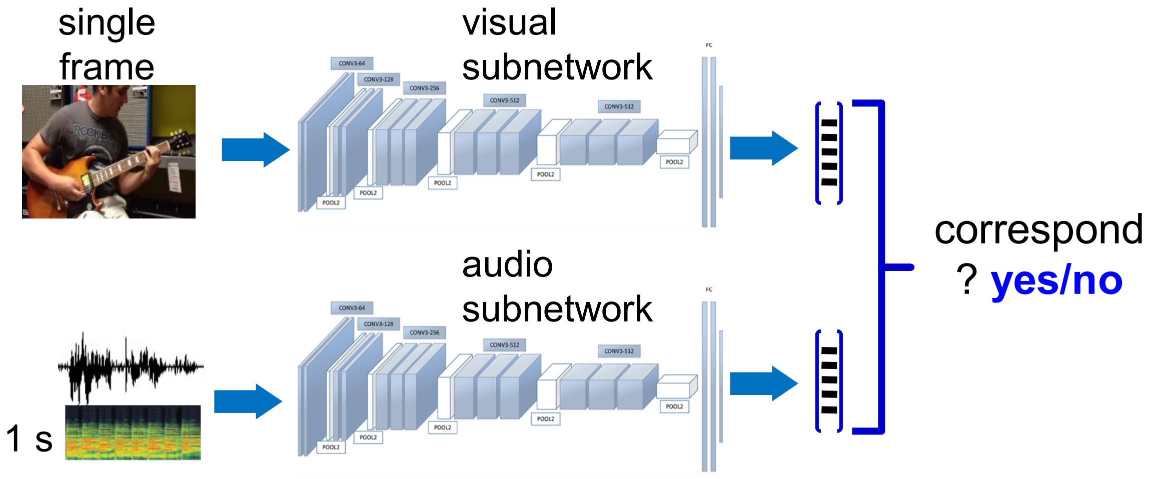
Method	Top 1 accuracy
Random	18.3%
Pathak *et al.* [21]	22.3%
Krähenbühl *et al.* [14]	24.5%
Donahue *et al.* [7]	31.0%
Doersch *et al.* [6]	31.7%
Zhang *et al.* [34] (init: [14])	32.6%
Noroozi and Favaro [18]	34.7%
Ours random	12.9%
Ours	32.3%

- On par with state-of-the-art self-supervised approaches
- The only method whose features haven't seen ImageNet images
  - Probably never seen 'Tibetan terrier'
  - Video frames are quite different from images

Use audio and visual features

What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- “What is making the sound?”
  - Learn to localize objects that sound



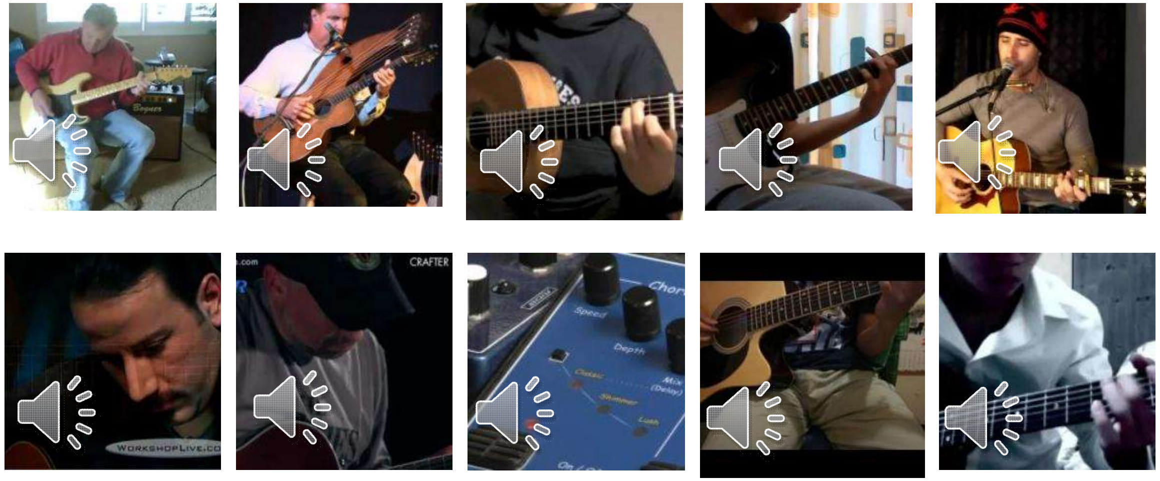
“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Query on image, retrieve audio

Search in 200k video clips of AudioSet

Query frame

Top 10 ranked audio clips

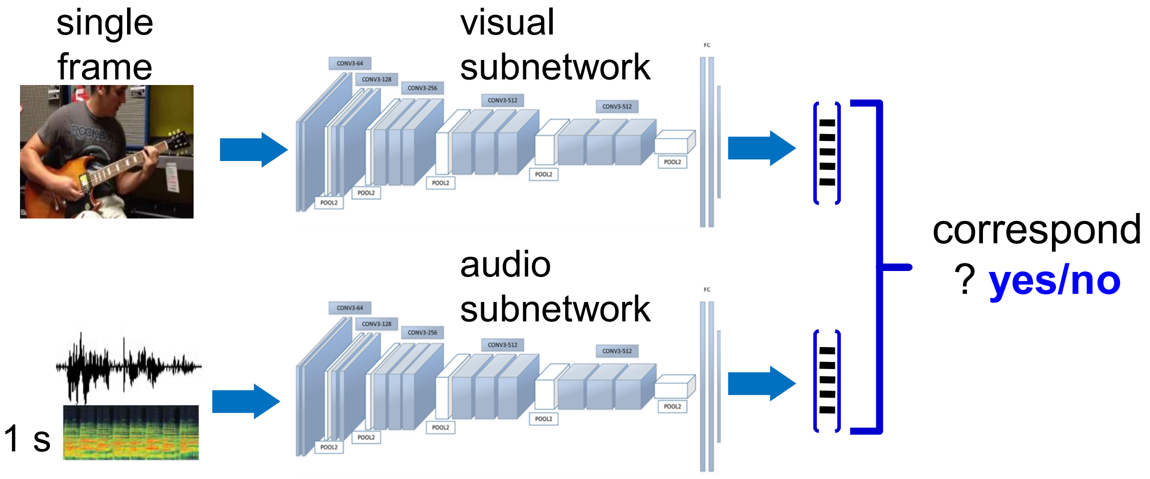


“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Use audio and visual features

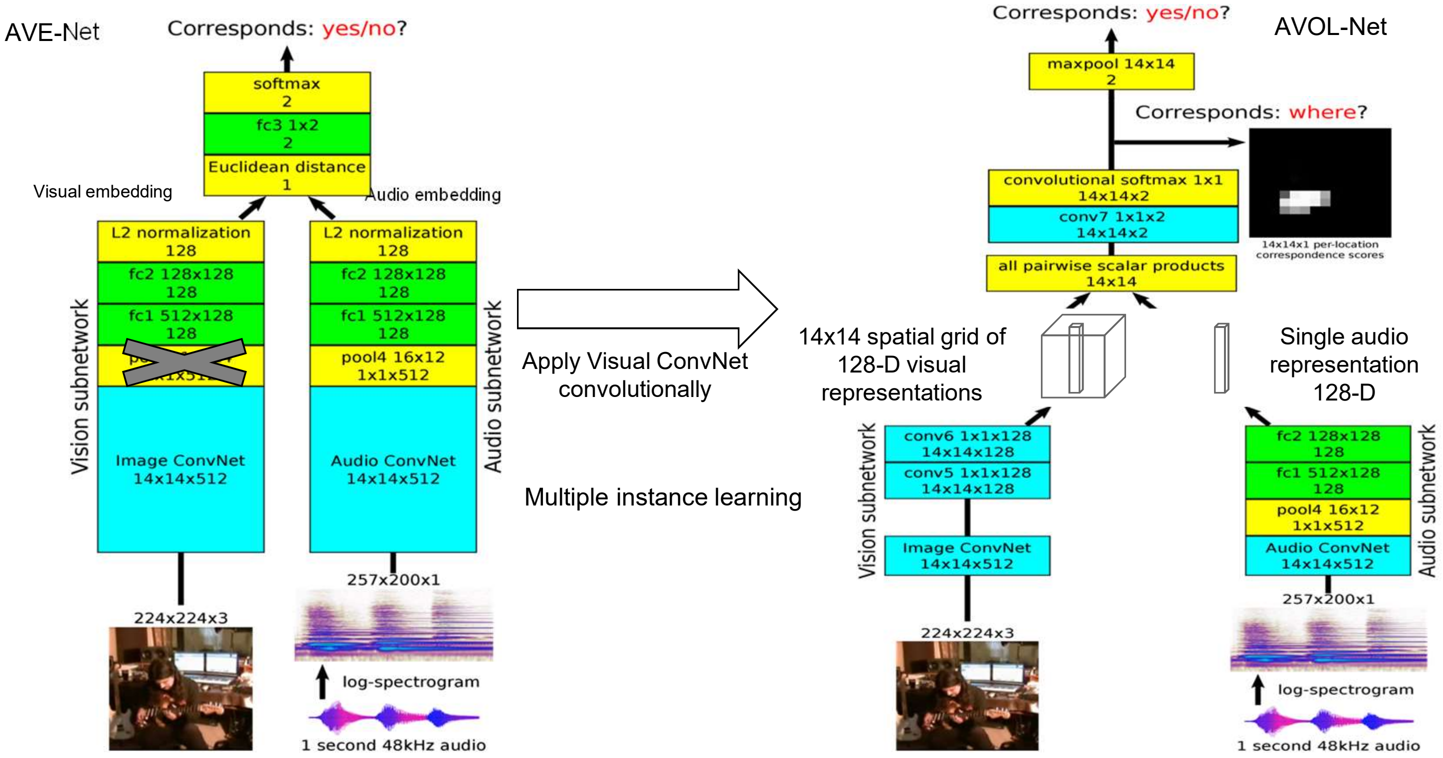
What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- “What is making the sound?”
  - Learn to localize objects that sound



“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Objects that Sound



Apply Visual ConvNet convolutionally

Multiple instance learning

14x14 spatial grid of 128-D visual representations

Single audio representation 128-D

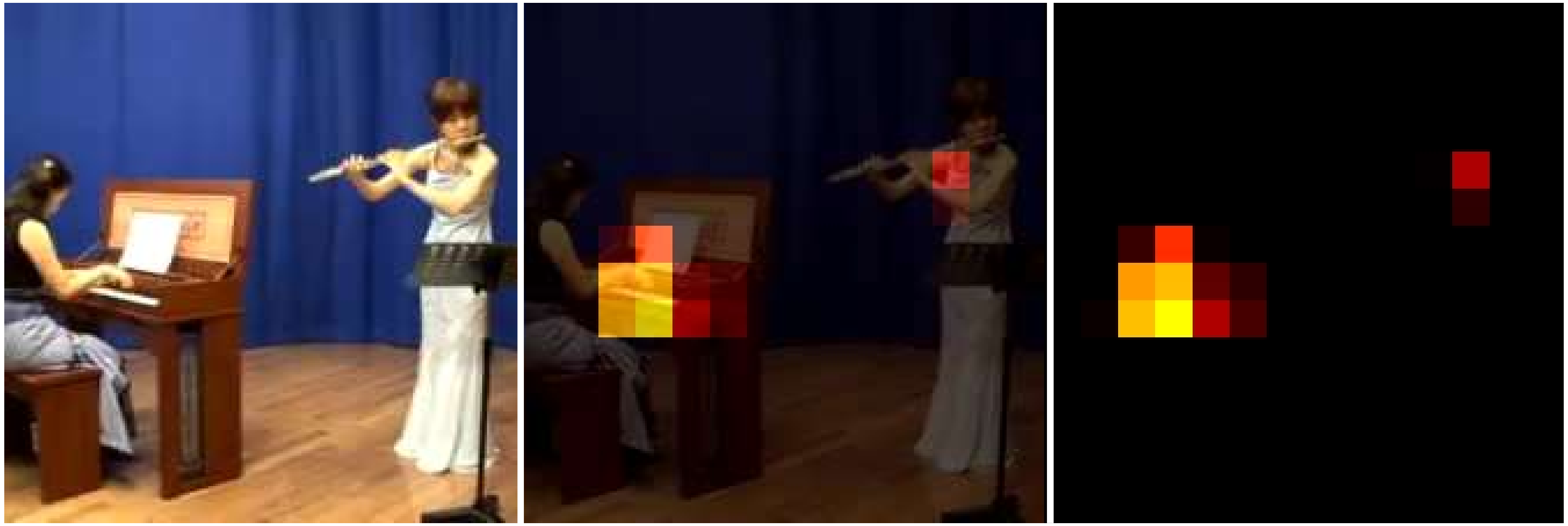
“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Localizing objects with sound

Input: audio and video frame

Output: localization heatmap on frame

What would make this sound?

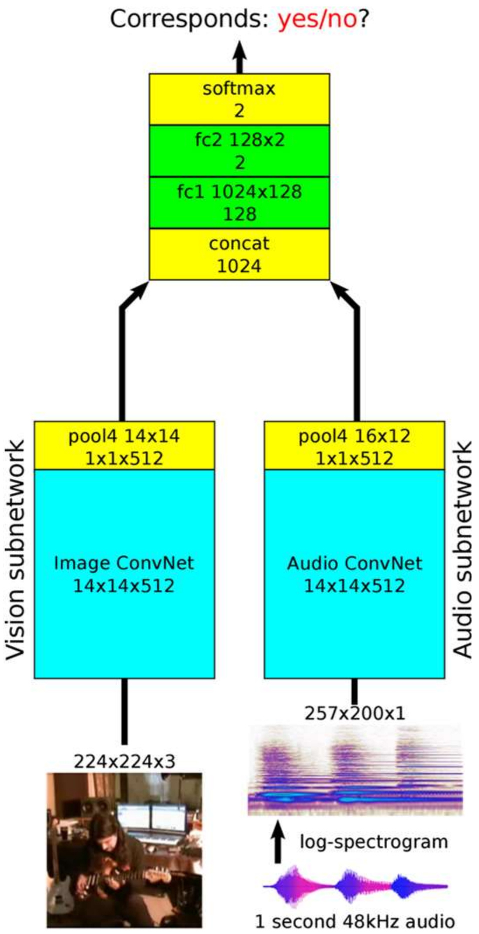


Note, no video (motion) information is used

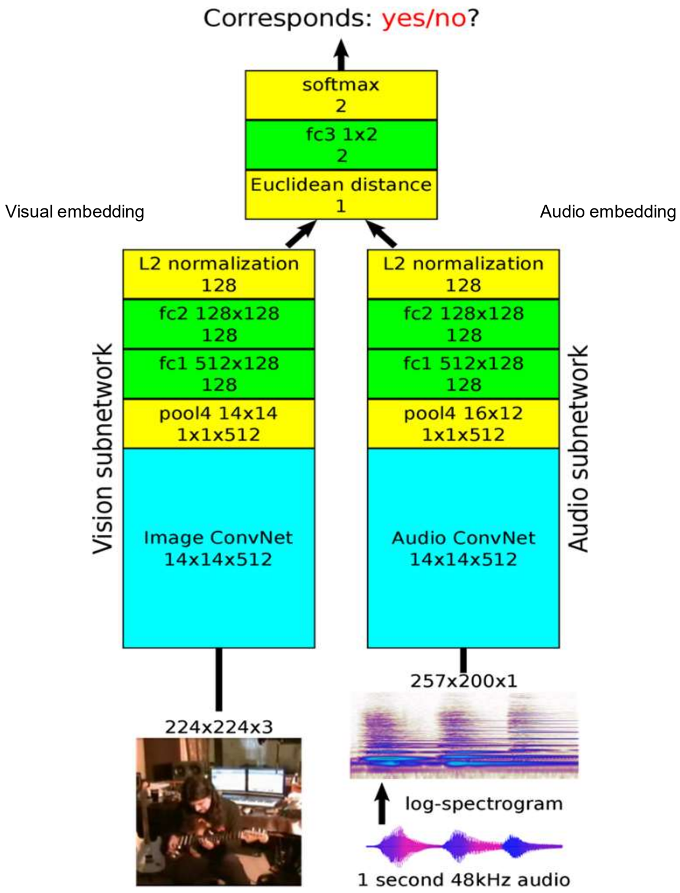
“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

To embed or not to embed?

Concatenation



Embedding

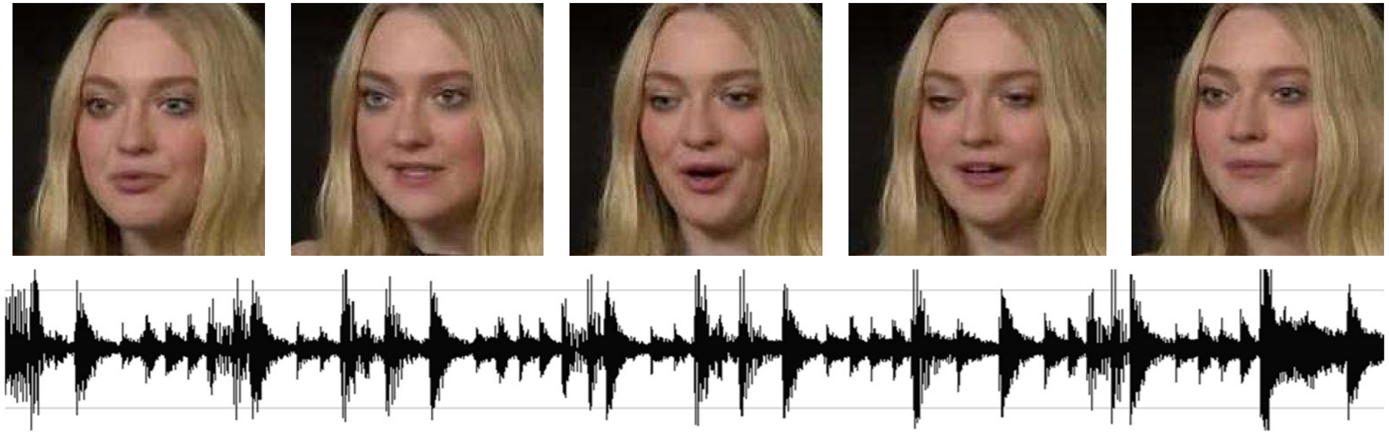


Features available

Cross-modal alignment in embedding

Specialize to talking heads …

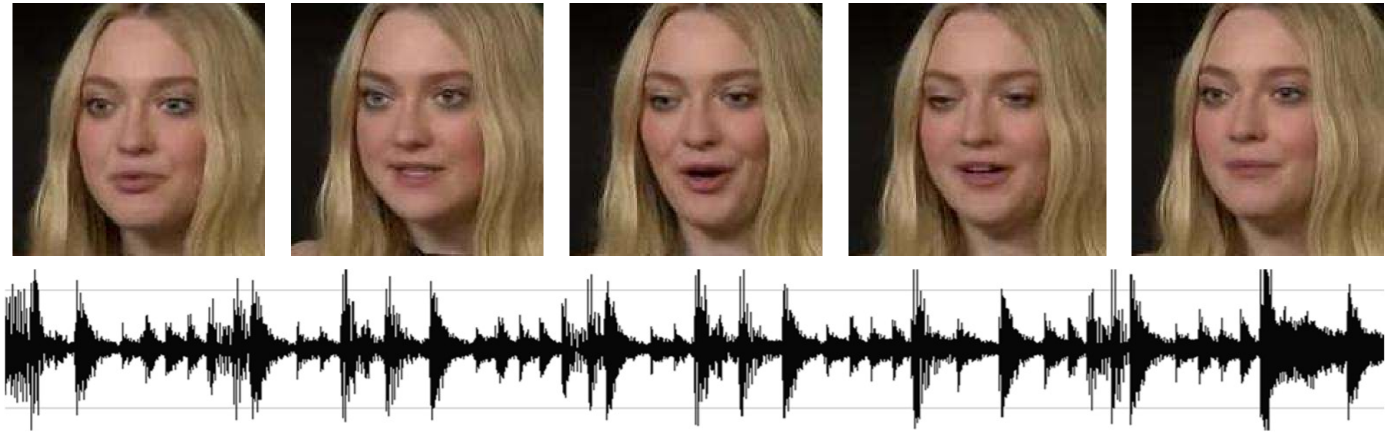
Objective: use faces and voice to learn from each other



- Two types of proxy task:
  1. Predict audio-visual **correspondence**
  2. Predict audio-visual **synchronization**

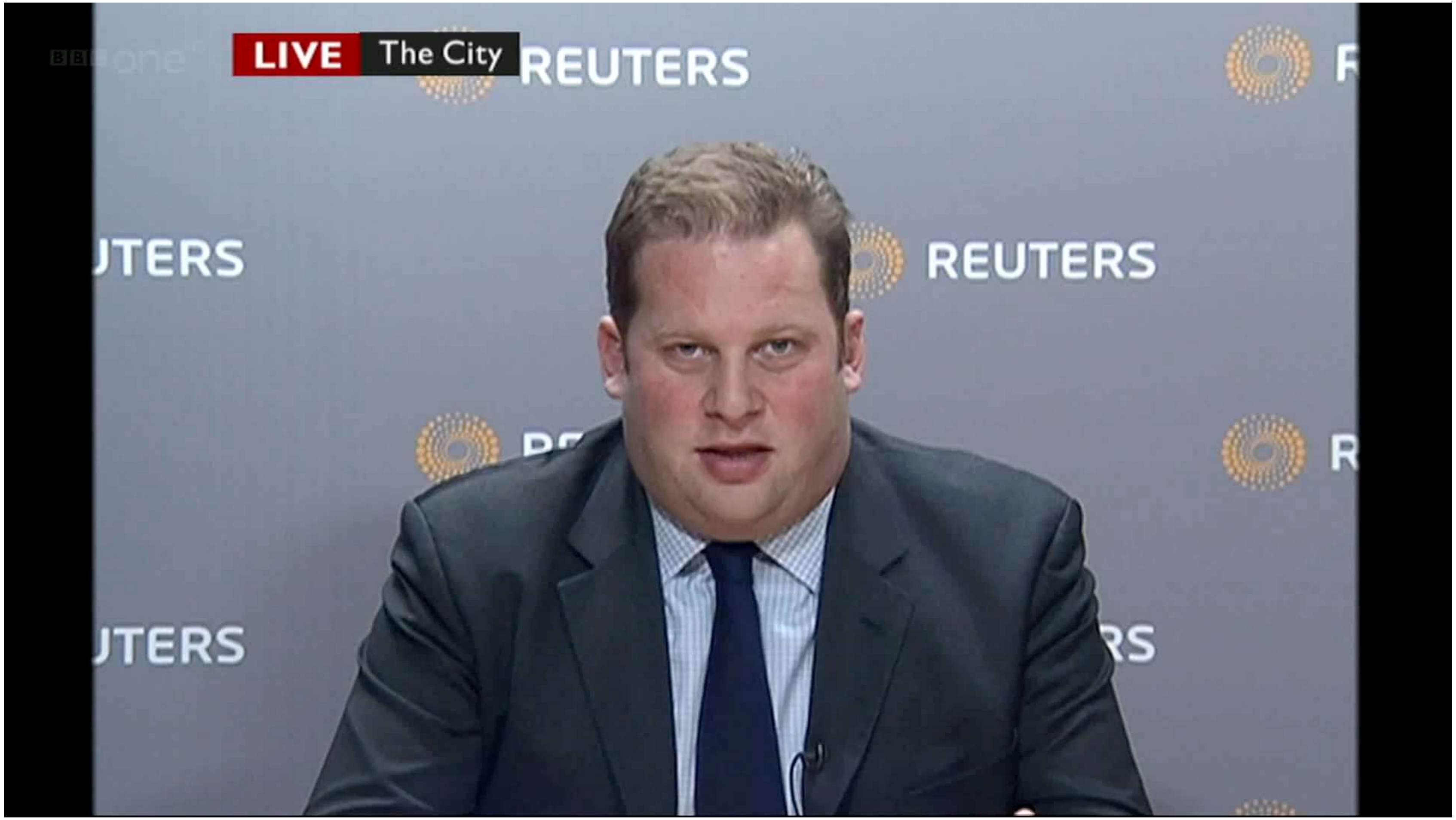
Specialize to talking heads …

Objective: use faces and voice to learn from each other



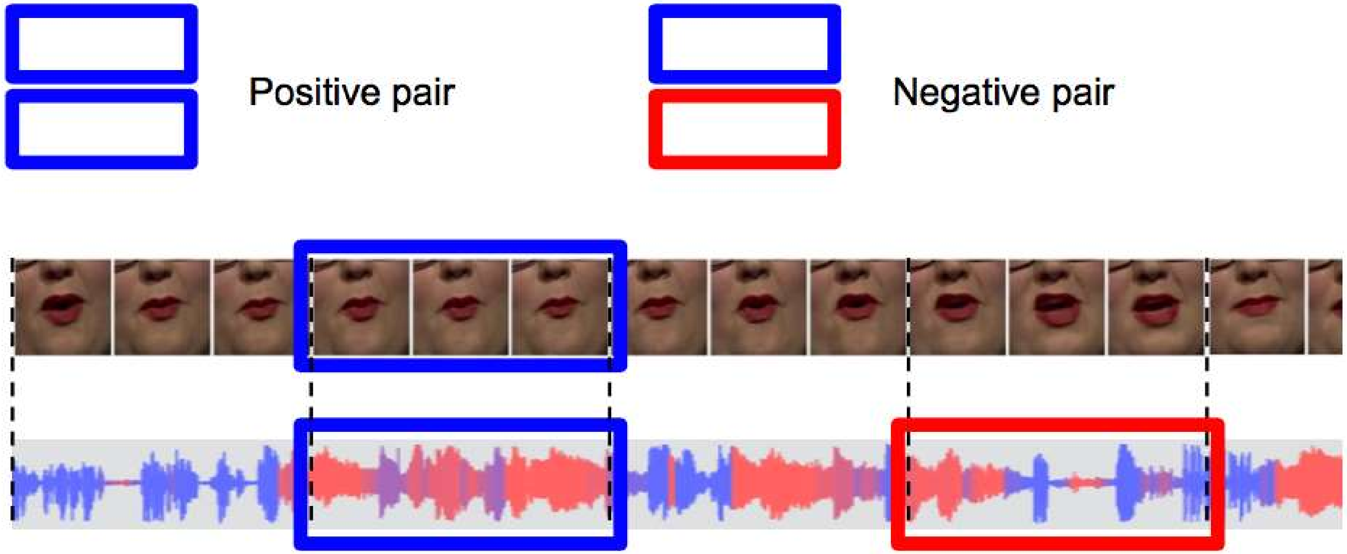
- Two types of proxy task:
    1. Predict audio-visual **correspondence**
    2. Predict audio-visual **synchronization**

Lip-sync problem on TV



Face-Speech Synchronization

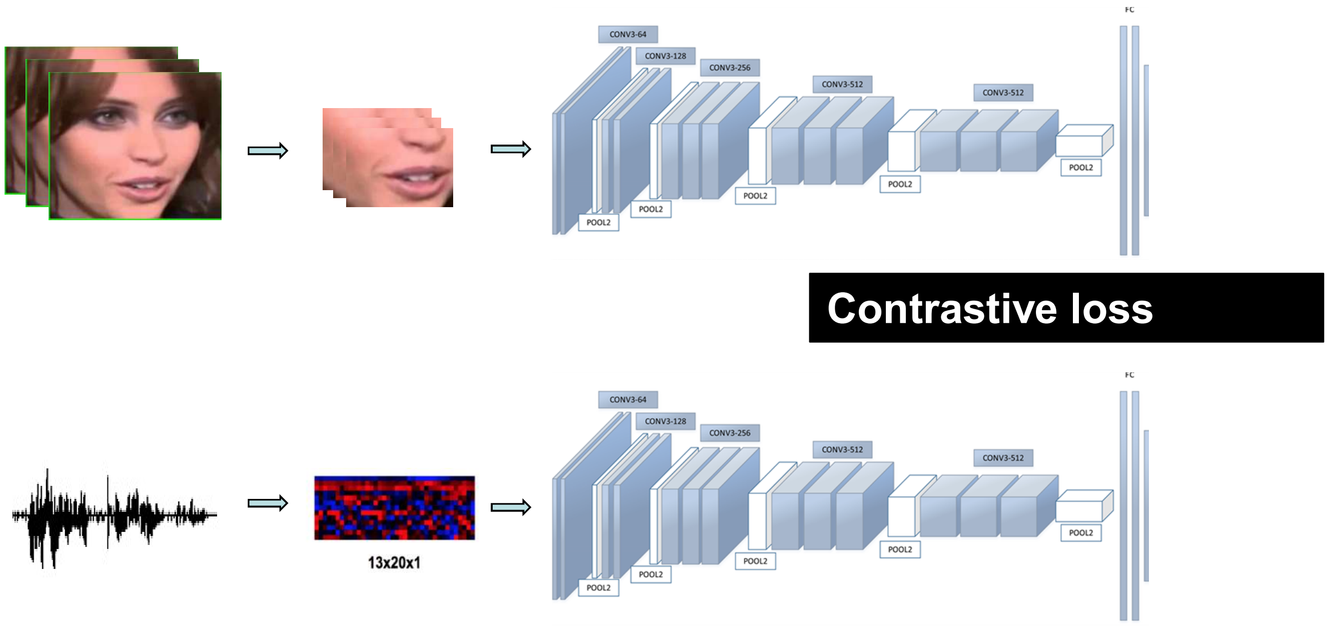
- **Positive samples:** in sync
- **Negative samples:** out of sync (introduce temporal offset)



Chung, Zisserman (2016) “Out of time: Automatic lip sync in the wild”

Sequence-sequence face-speech network

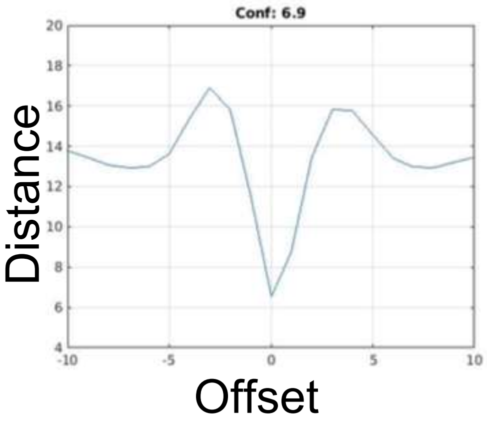
- The network is trained with contrastive loss to:
  - Minimise distance between positive pairs
  - Maximise distance between negative pairs



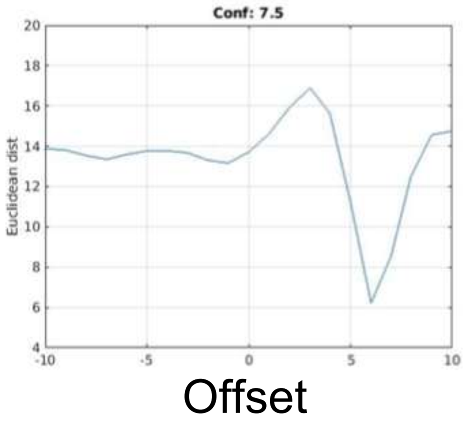
Chung, Zisserman (2016) “Out of time: Automatic lip sync in the wild”

Face-Speech Synchronization

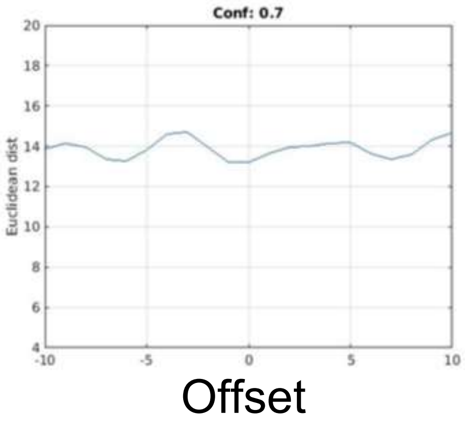
- Averaged sliding windows
- The predicted offset value is >99% accurate, averaged over 100 frames.



In-sync



Off-sync

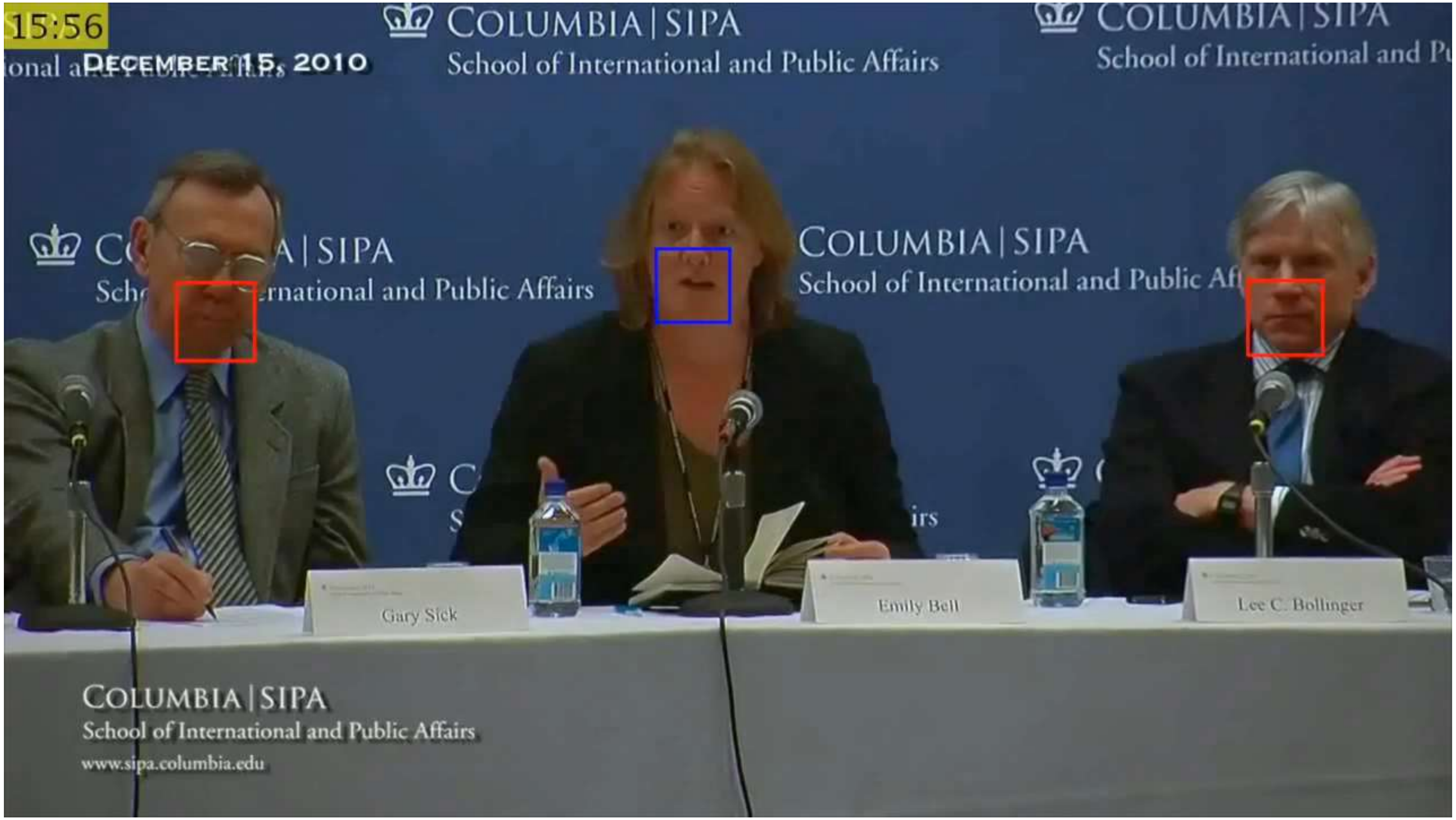


Non-speaker

Chung, Zisserman (2016) “Out of time: Automatic lip sync in the wild”

Application: Lip Synchronization

Application: Active speaker detection



Blue: speaker **Red: non-speaker**

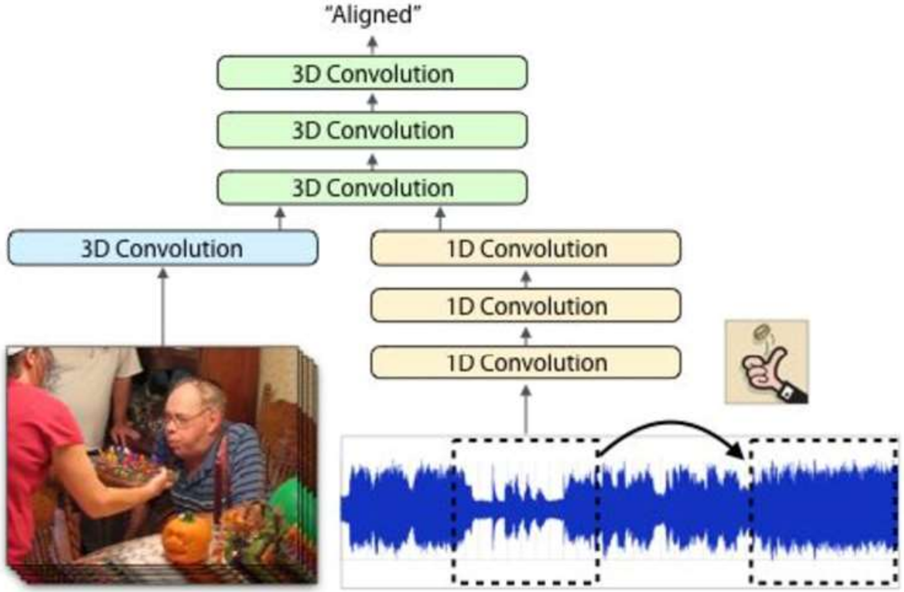
Face-Speech Synchronization - summary

The network can be used for:

- Audio-to-video synchronisation
- Active speaker detection
- Voice-over rejection
- Visual features for lip reading

Audio-Visual Synchronization

Learning by Misaligned Audio

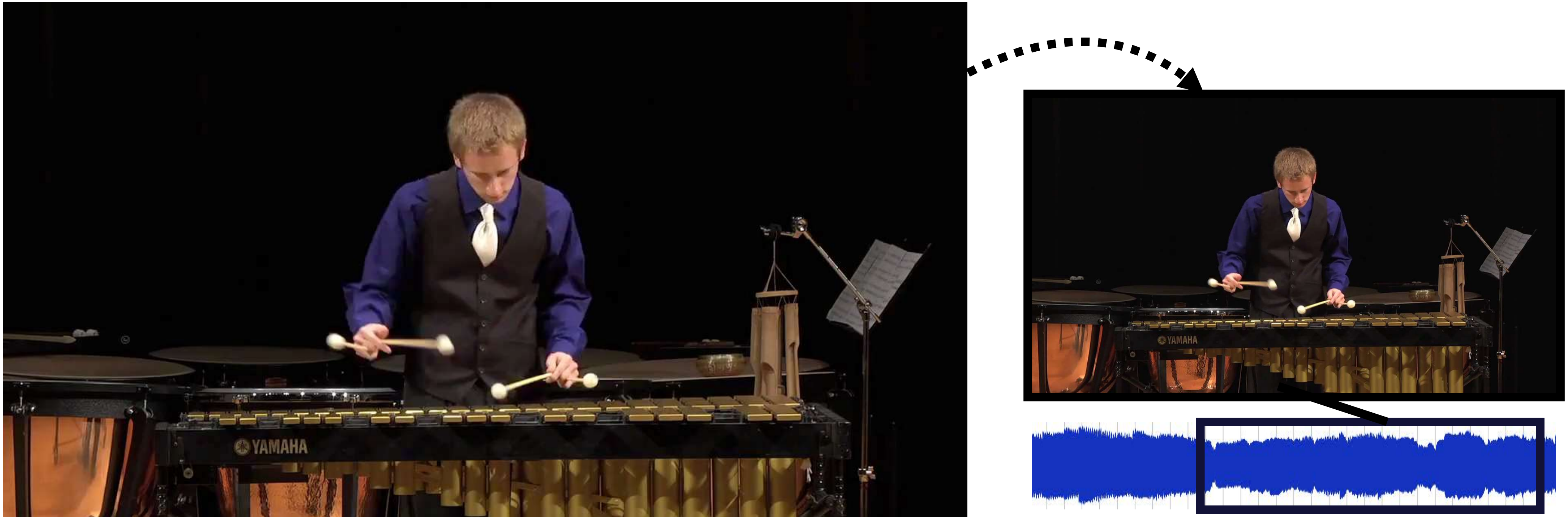


Audio-Visual Scene Analysis with Self-Supervised Multisensory Features
Andrew Owens, Alyosha Efros

Self-supervised Training

Audio-Visual Scene Analysis with Self-Supervised Multisensory Features,
Andrew Owens, Alyosha Efros, 2018

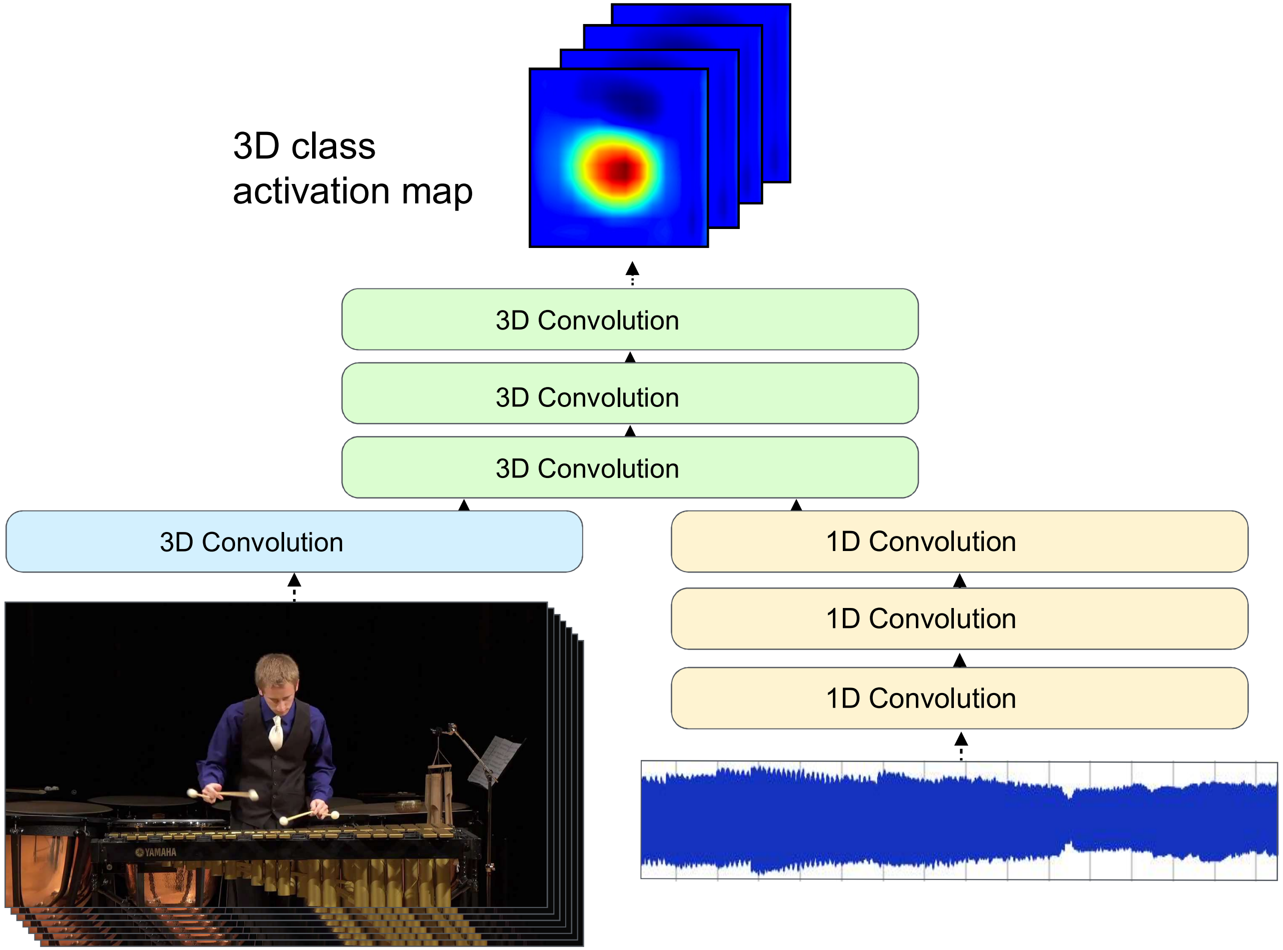
Misaligned Audio



Shifted audio track

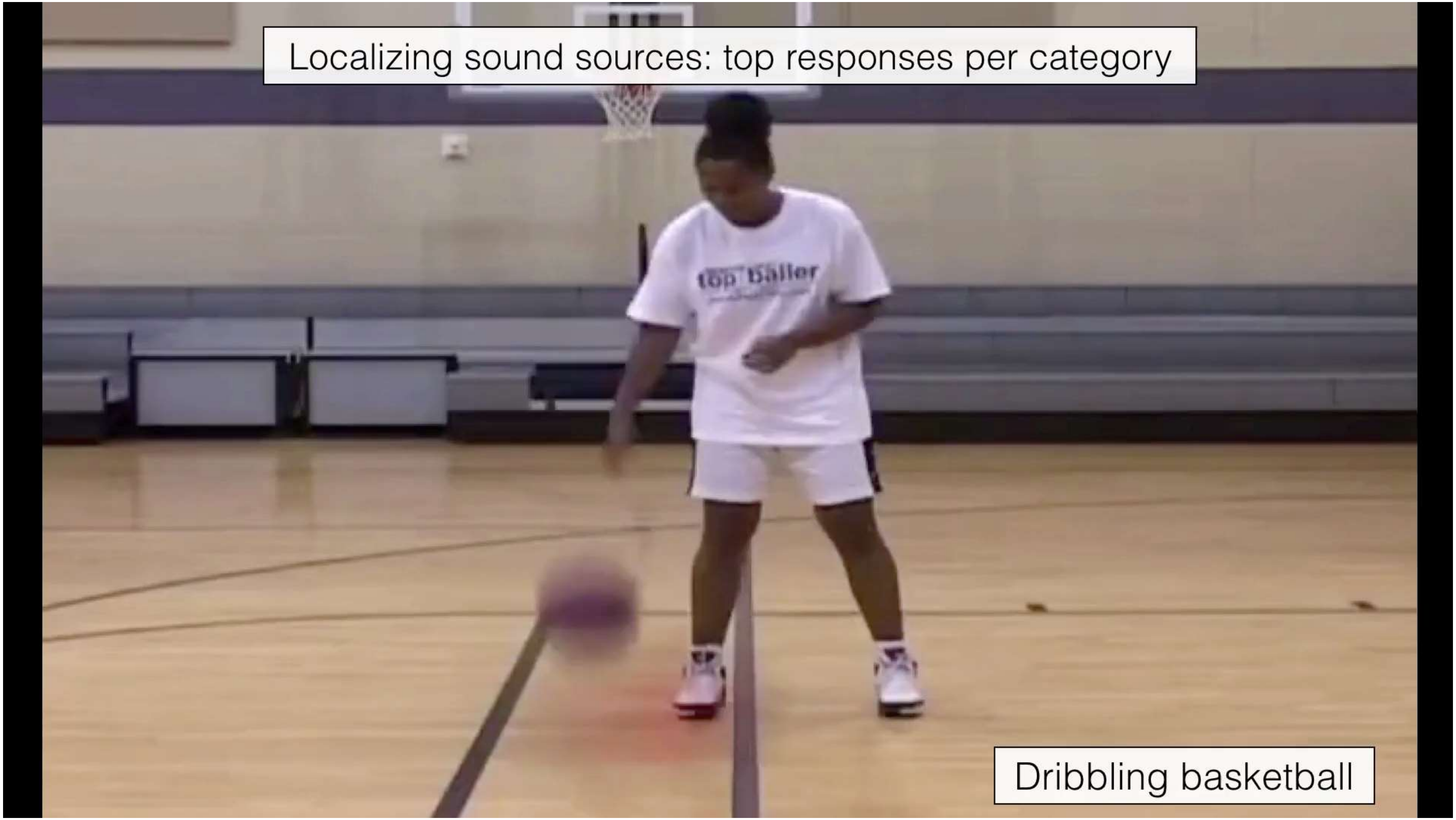
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features,
Andrew Owens, Alyosha Efros, 2018

Visualizing the location of sound sources



Audio-Visual Scene Analysis with Self-Supervised Multisensory Features,
Andrew Owens, Alyosha Efros, 2018

Localizing sound sources: top responses per category



Dribbling basketball

Summary: Audio-Visual Co-supervision

Objective: use vision and sound to learn from each other

- Two types of proxy task:
  1. Predict audio-visual correspondence **-> semantics**
  2. Predict audio-visual synchronization **-> attention**

- Lessons are applicable to any two related sequences, e.g. stereo video, RGB/D video streams, visual/infrared cameras …

Summary

- Self-Supervised Learning from images/video
  - Enables learning without explicit supervision
  - Learns visual representations – on par with ImageNet training

- Self-Supervised Learning from videos with sound
  - Intra- and cross-modal retrieval
  - Learn to localize sounds
  - Tasks not just a proxy, e.g. synchronization, attention, applicable directly

- Applicable to other domains with paired signals, e.g.
  - face and voice
  - Infrared/visible
  - RGB/D
  - Stereo streams …