How to avoid manual annotation?

Part I: Weakly-supervised learning

Coarse or cheap labels

Part II: Self-supervised learning

No labels

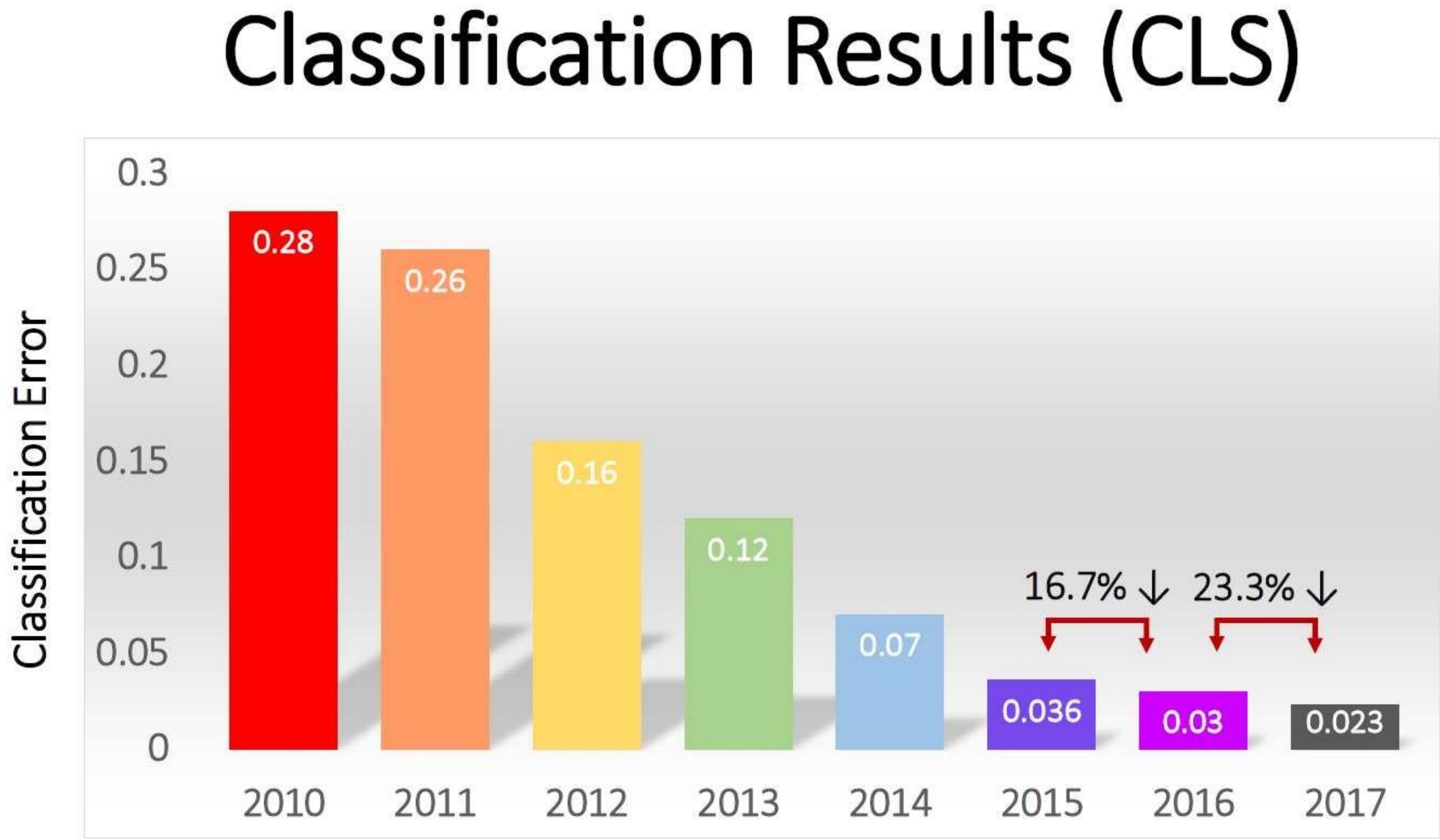
The ImageNet Challenge Story …

IMAGENET

1000 categories

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

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

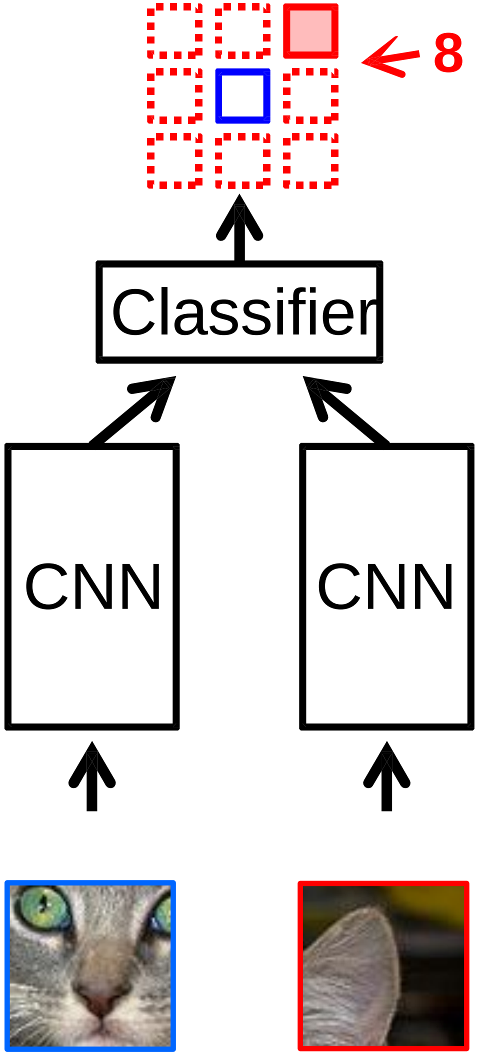
The Development of Embodied Cognition: Six Lessons from Babies
by Linda Smith and Michael Gasser

What is Self-Supervision?

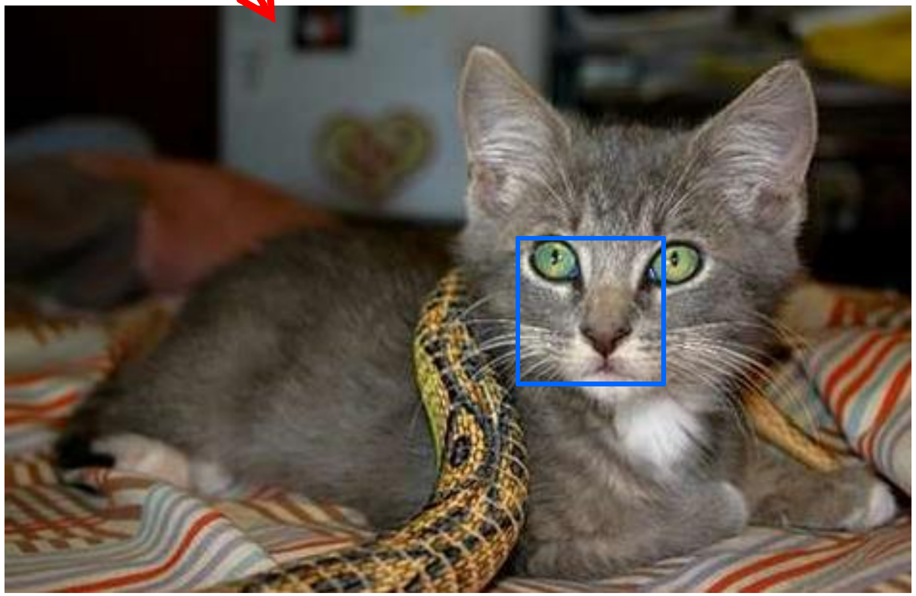
- A form of unsupervised learning where the data provides the **supervision**
- In general, withhold some part of the data, and task the network with predicting it
- The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to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

Example: relative positioning

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



8 possible locations

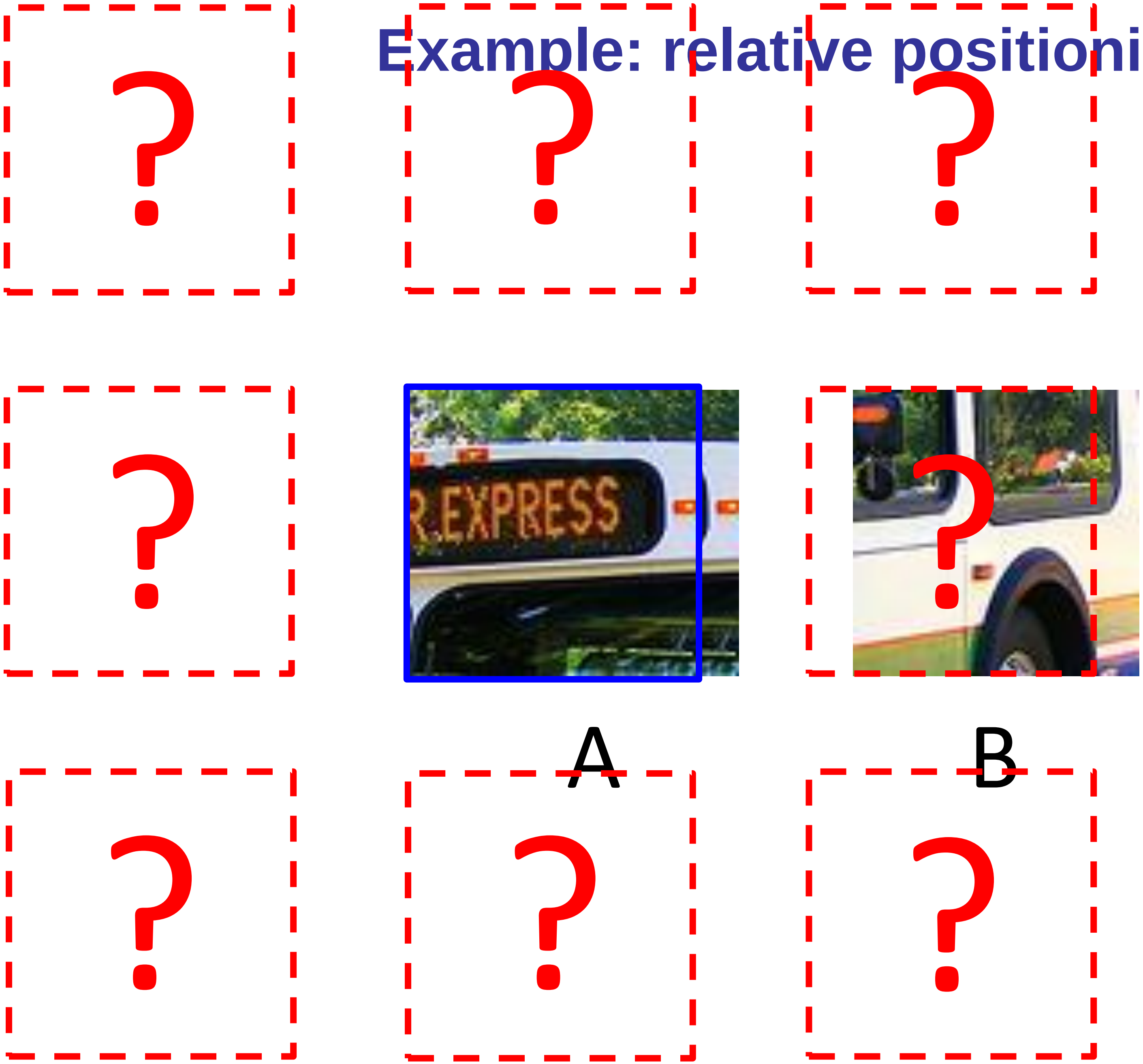


Randomly Sample Patch

Sample Second Patch

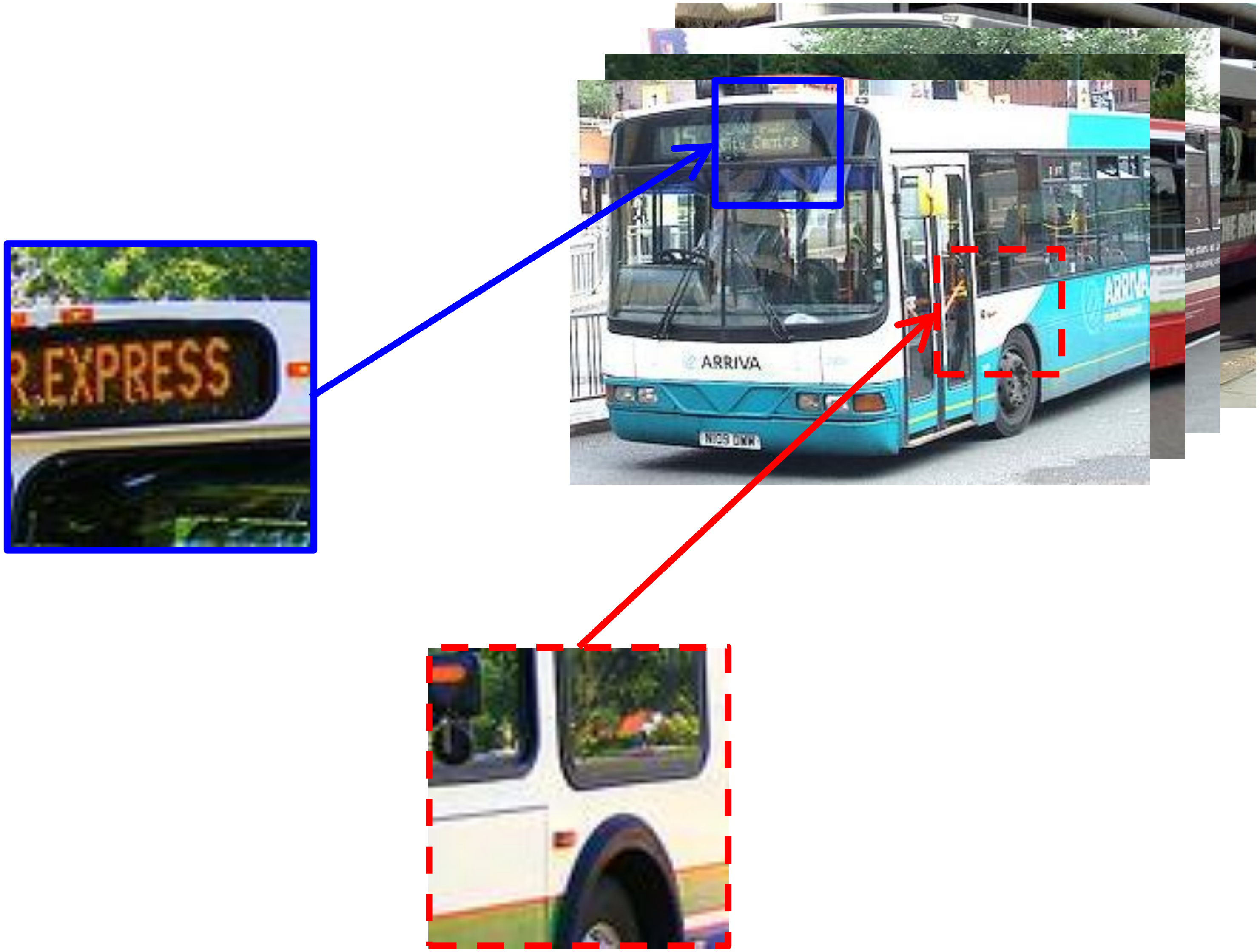
Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

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

Outline

Self-supervised learning in three parts:

A. from images

B. from videos

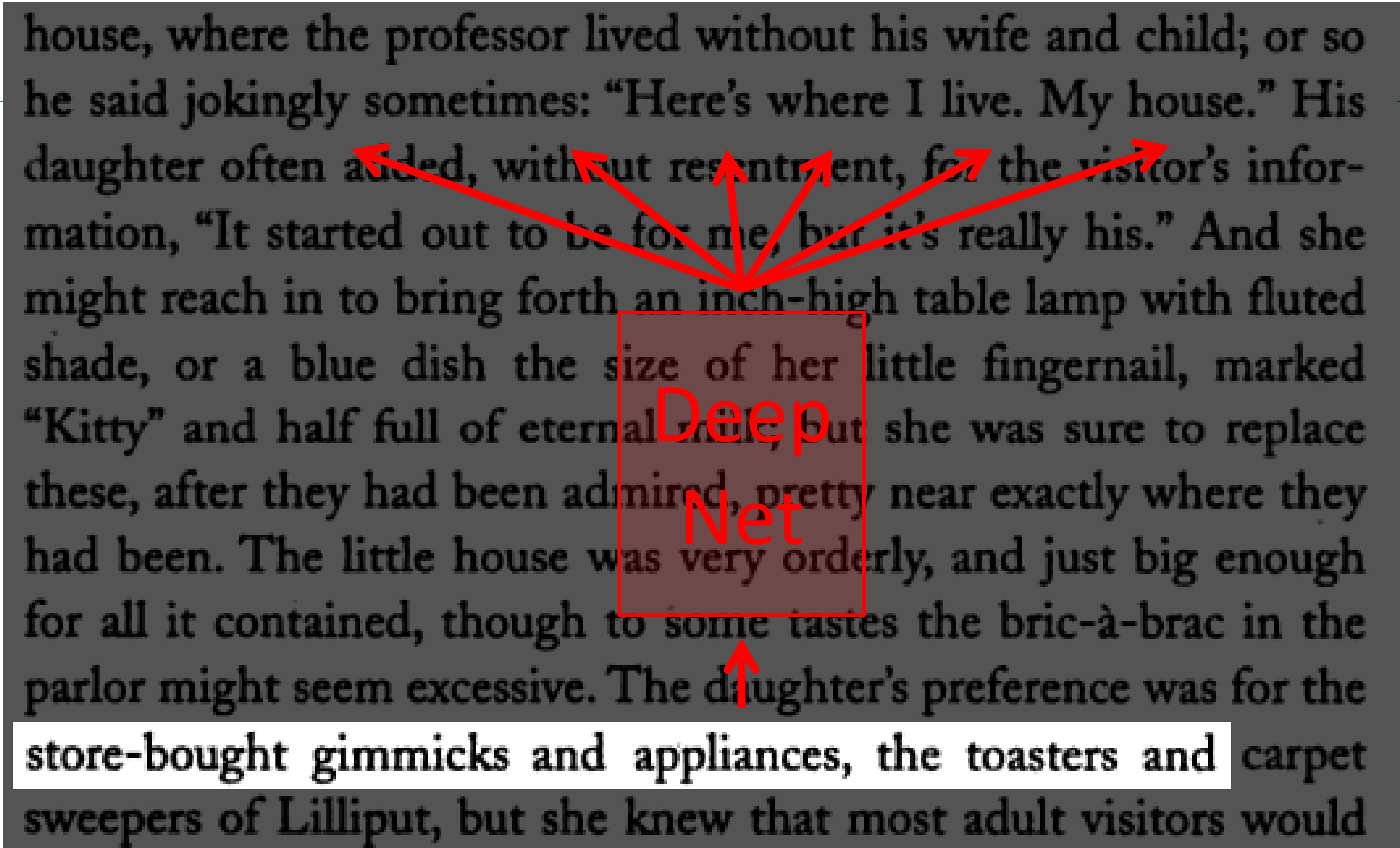
C. from videos and sound

Part A

Self-Supervised Learning from Images

Context as Supervision

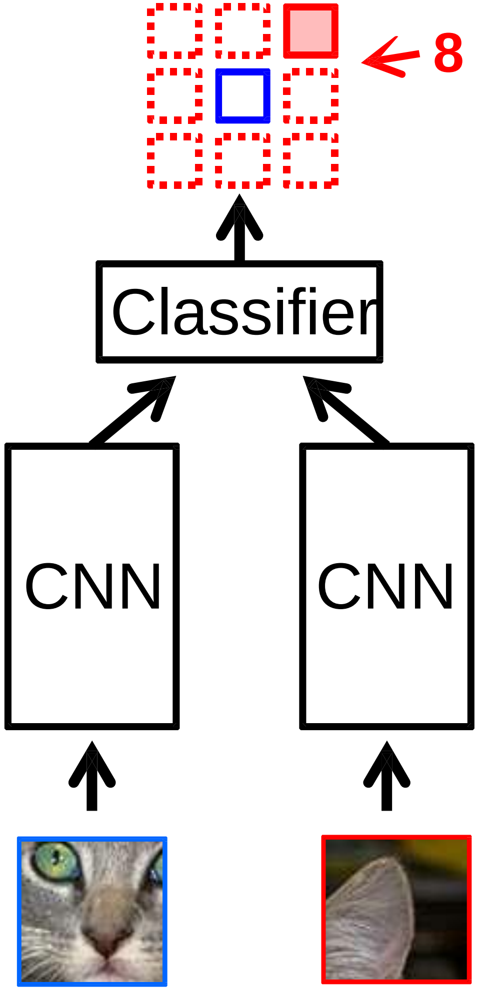
[Collobert & Weston 2008; Mikolov et al. 2013]



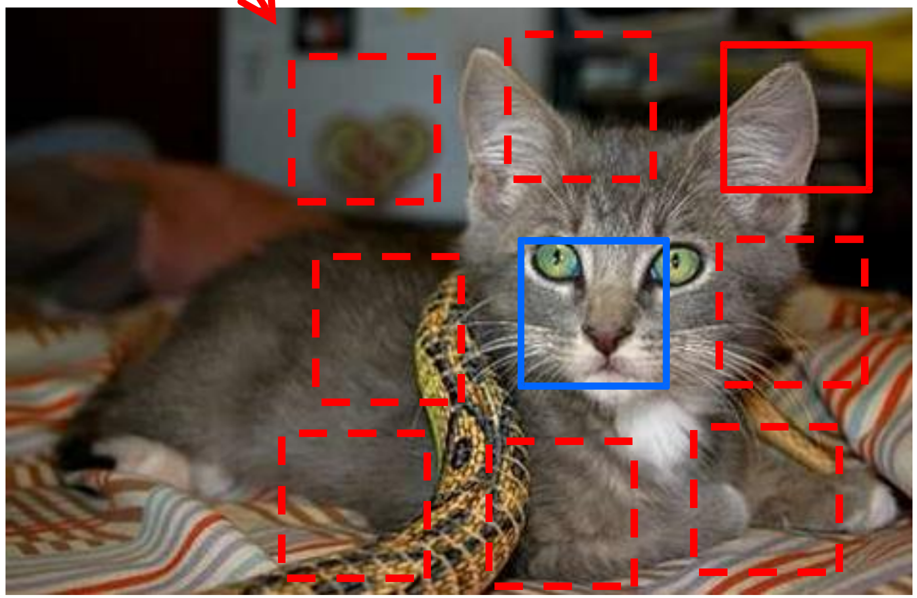
house, where the professor lived without his wife and child; or so he said jokingly sometimes: "Here's where I live. My house." His daughter often added, without resentment, for the visitor's information, "It started out to be for me, but it's really his." And she might reach in to bring forth an inch-high table lamp with fluted shade, or a blue dish the size of her little fingernail, marked "Kitty" and half full of eternal milk, but she was sure to replace these, after they had been admired, pretty near exactly where they had been. The little house was very orderly, and just big enough for all it contained, though to some tastes the bric-à-brac in the parlor might seem excessive. The daughter's preference was for the store-bought gimmicks and appliances, the toasters and carpet sweepers of Lilliput, but she knew that most adult visitors would

Recap: relative positioning

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



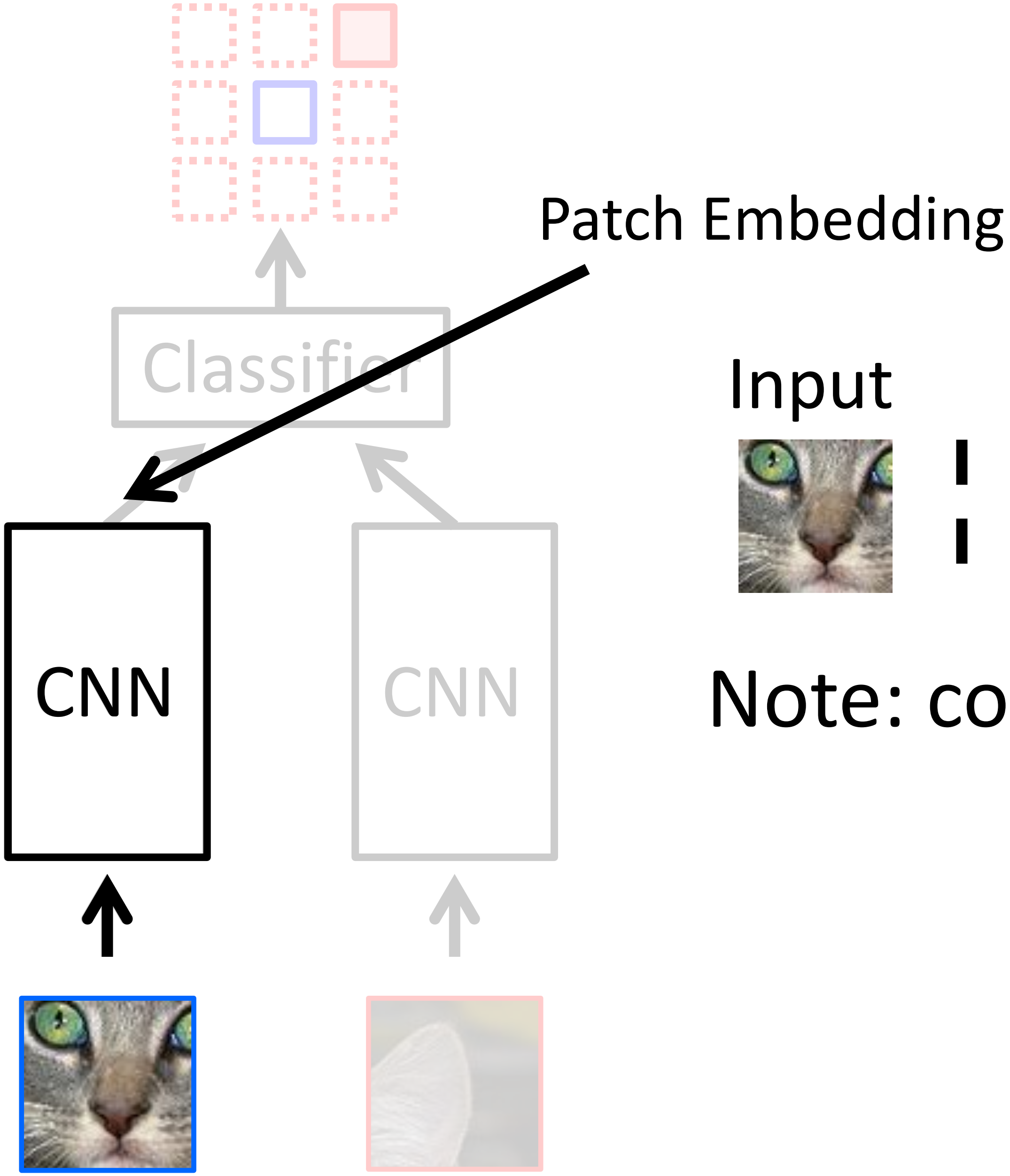
8 possible locations



Randomly Sample Patch

Sample Second Patch

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



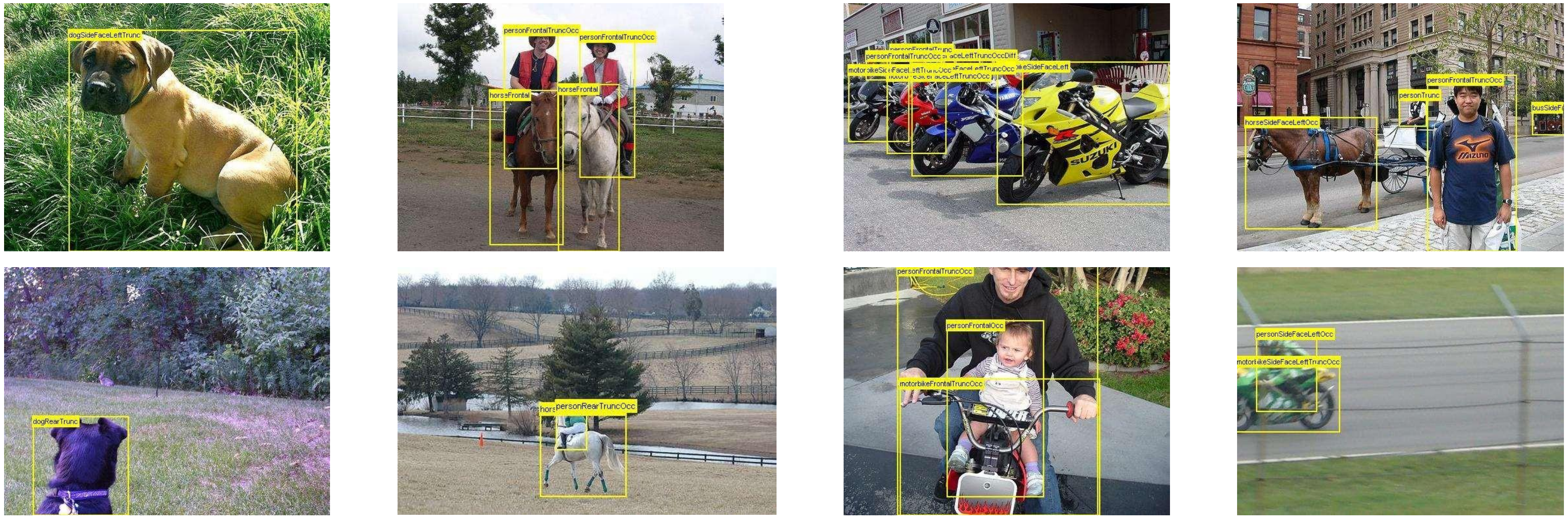
Note: connects ***across*** instances!

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)

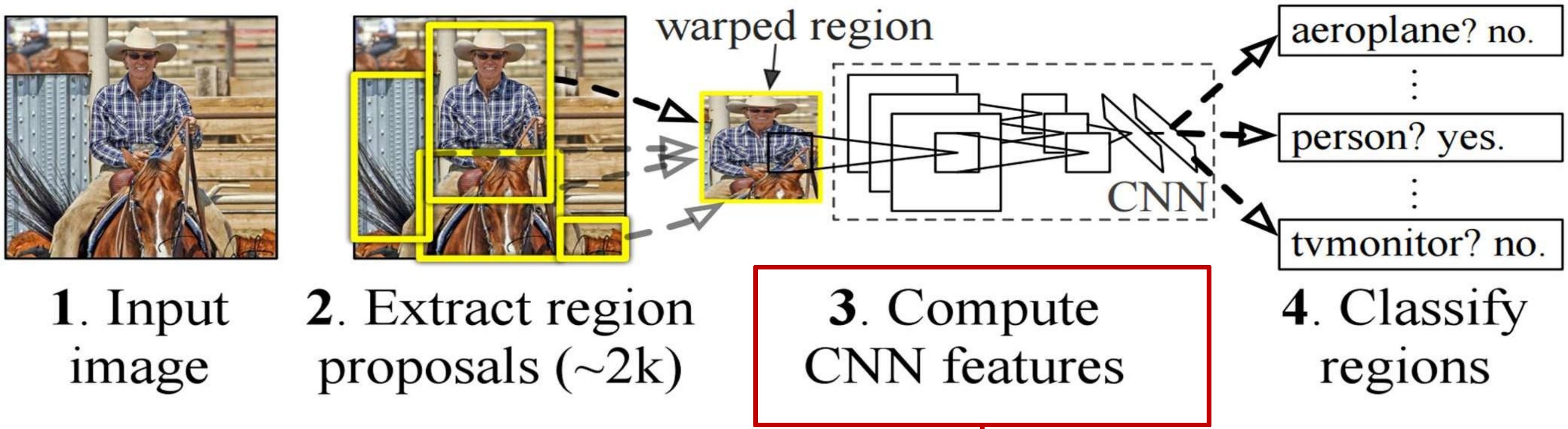
Dog Horse Motorbike Person



Evaluation: PASCAL VOC Detection

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

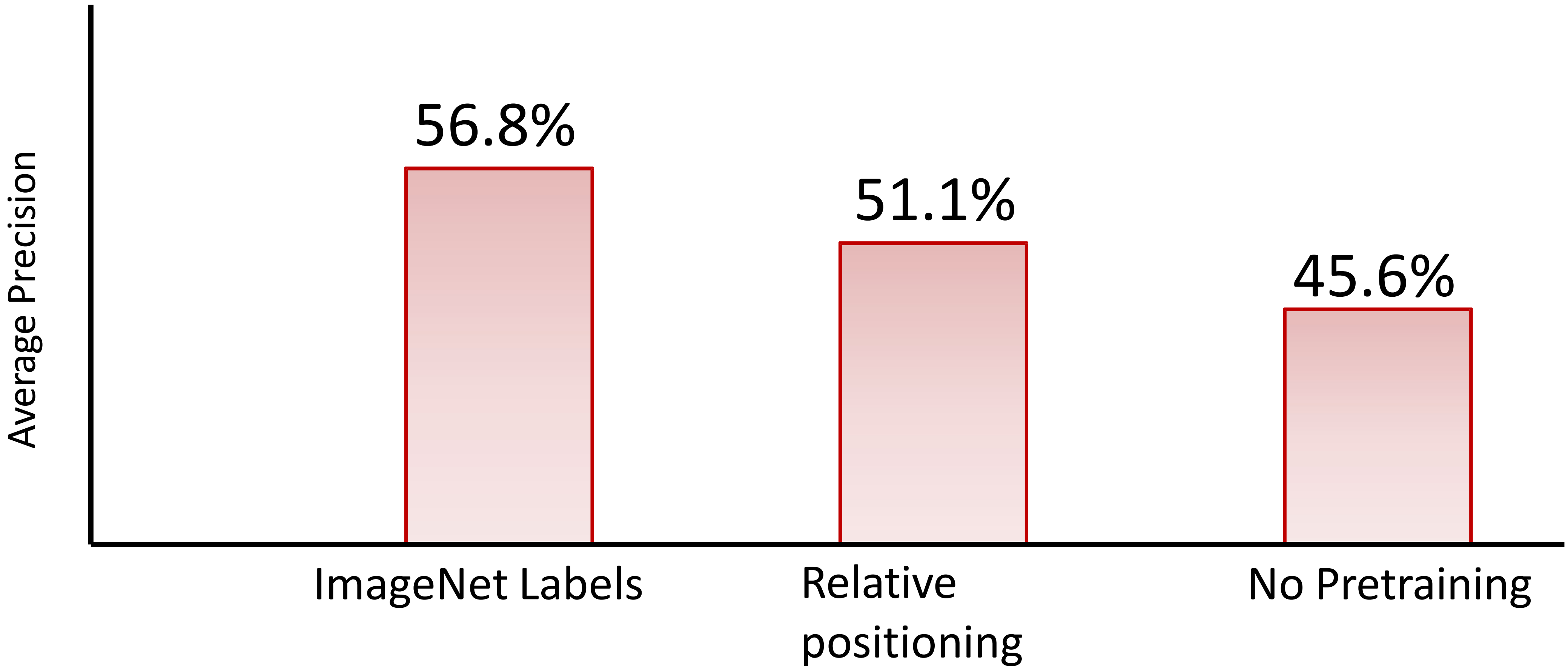
R-CNN



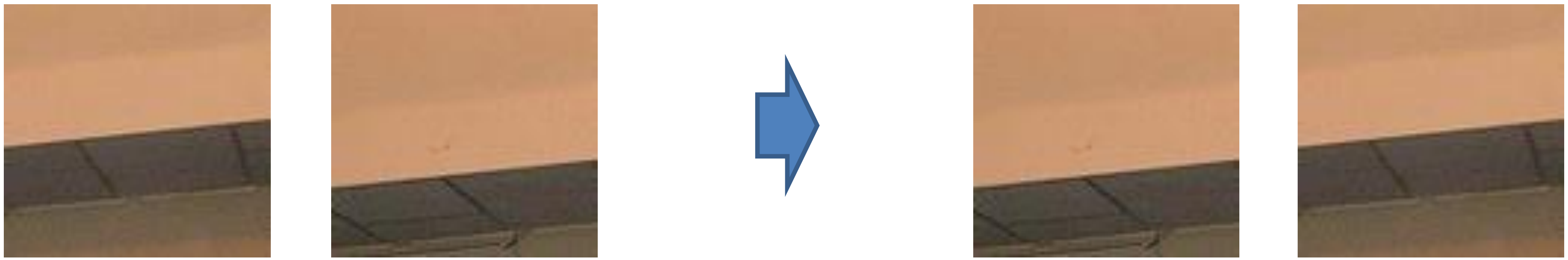
Pre-train on relative-position task, w/o labels

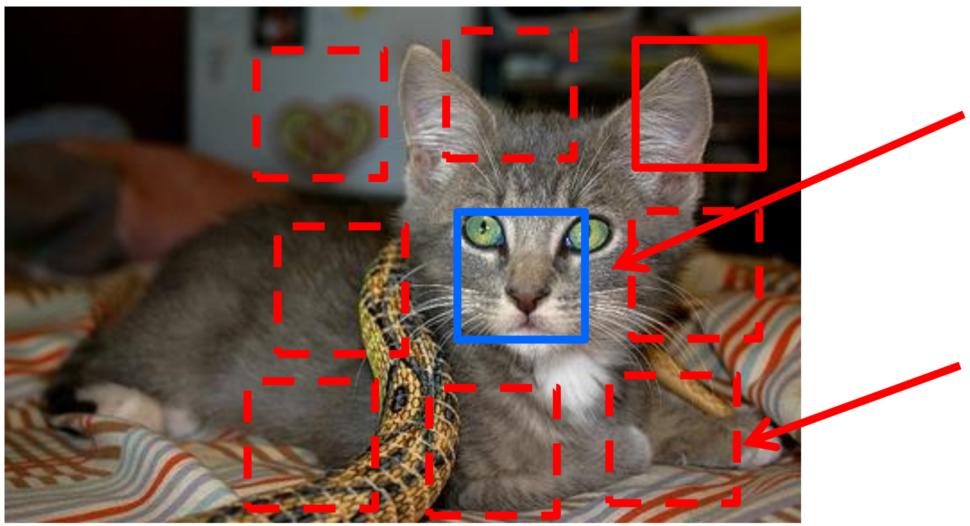
[Girshick et al. 2014]

Evaluation: PASCAL VOC Detection

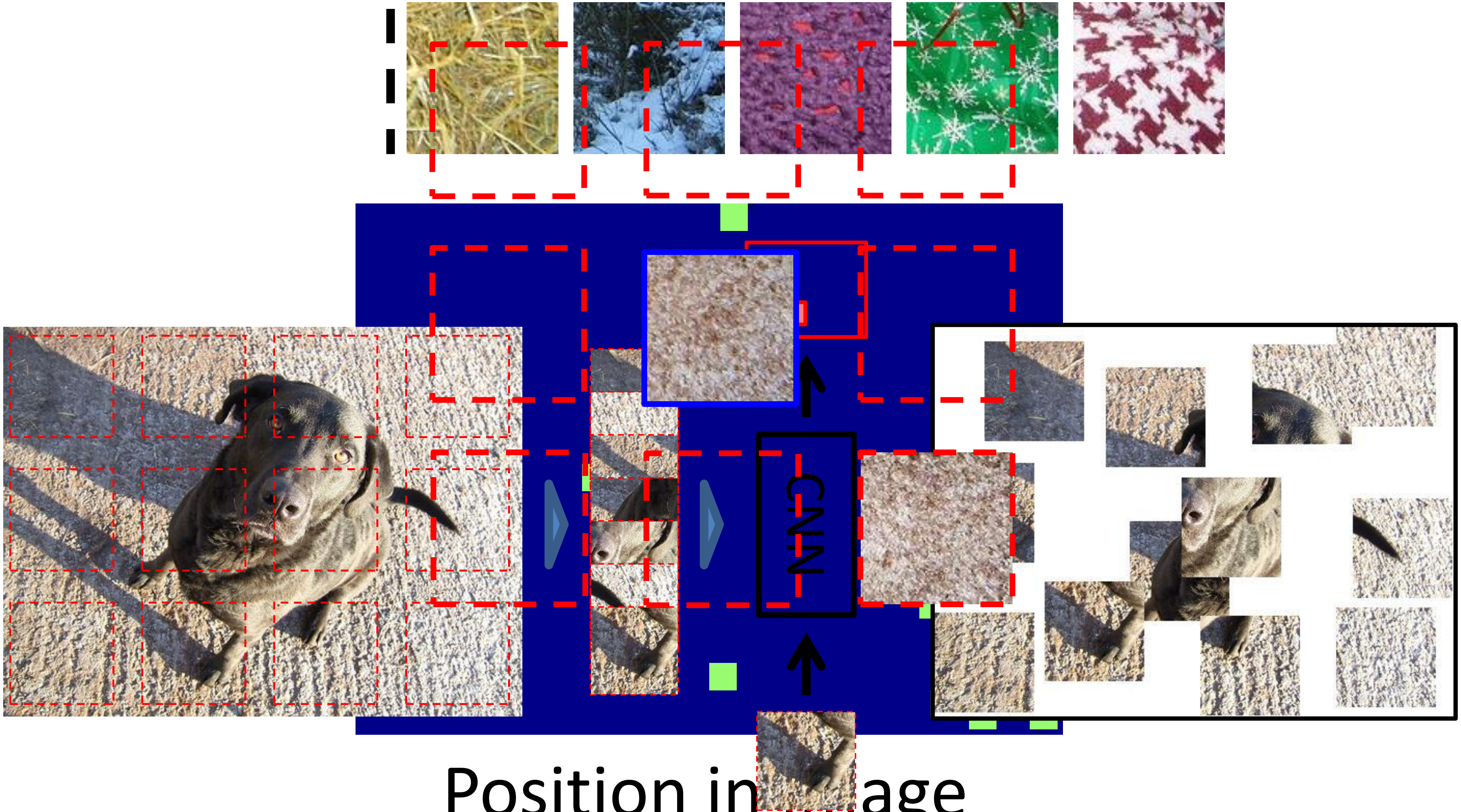


Avoiding Trivial Shortcuts



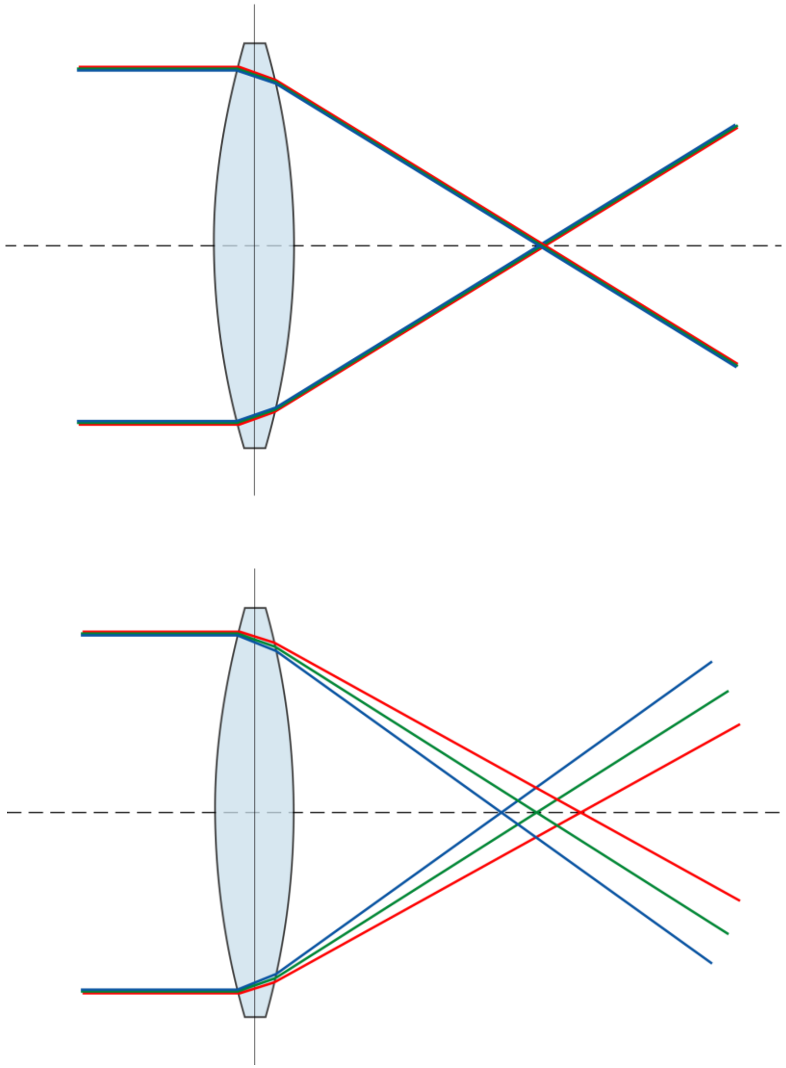


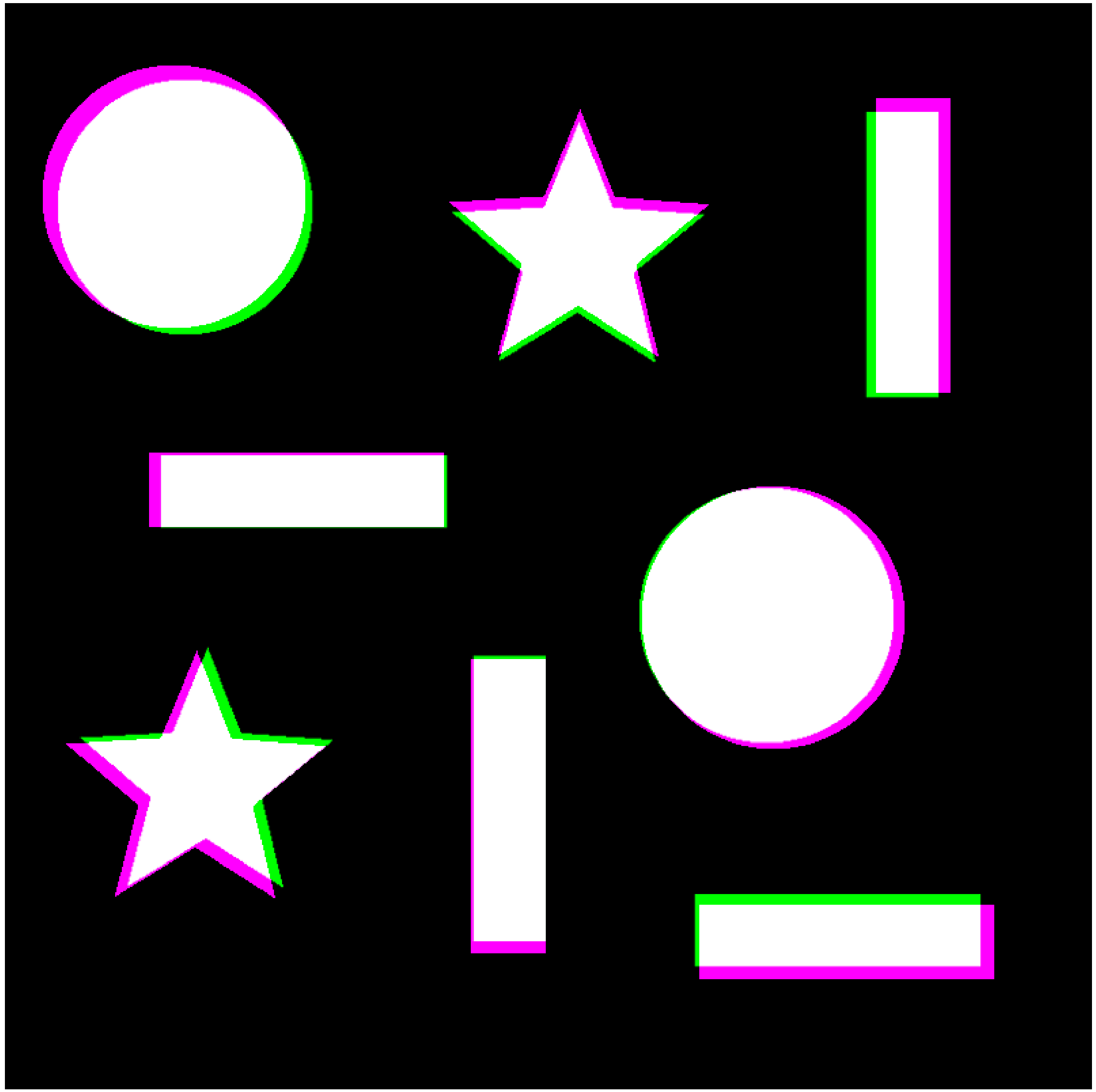
A Not-So “Trivial” Shortcut



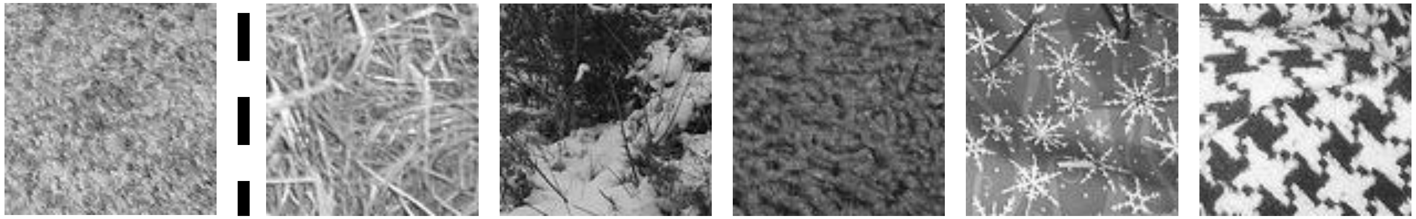
Position in ...age

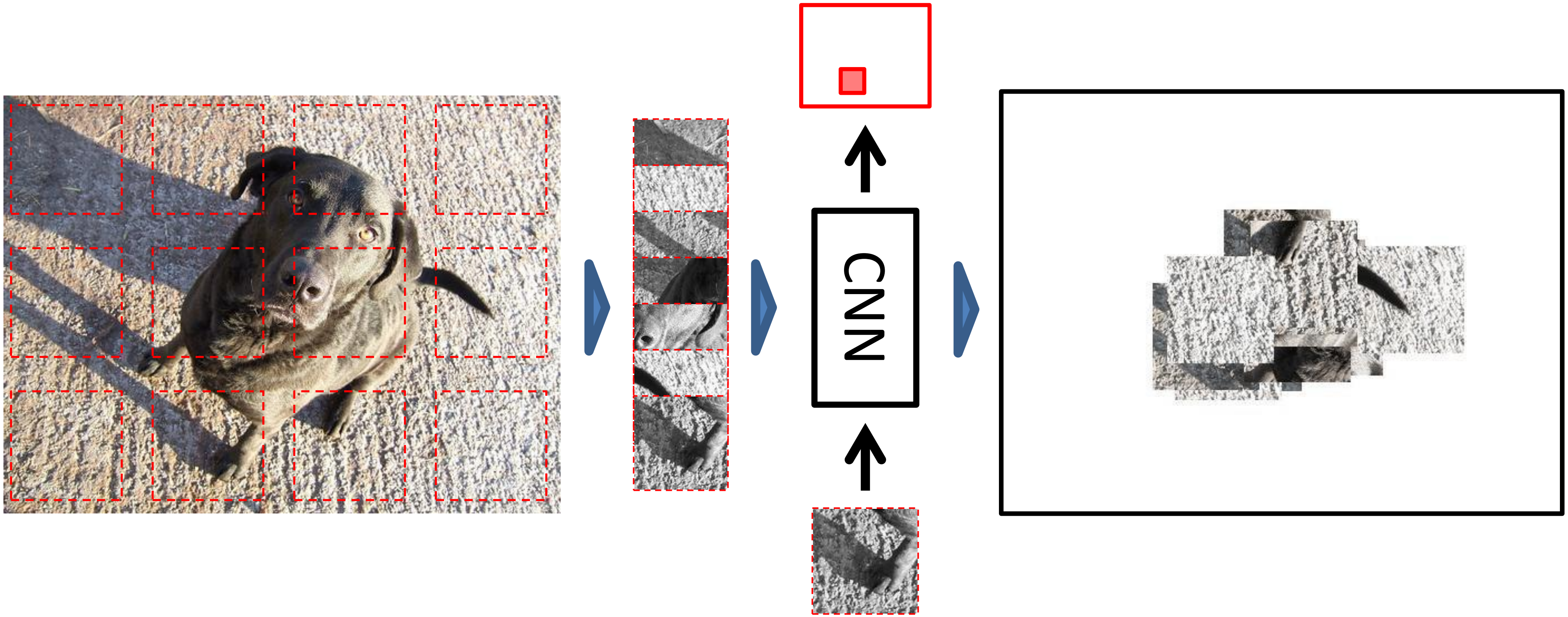
Chromatic Aberration



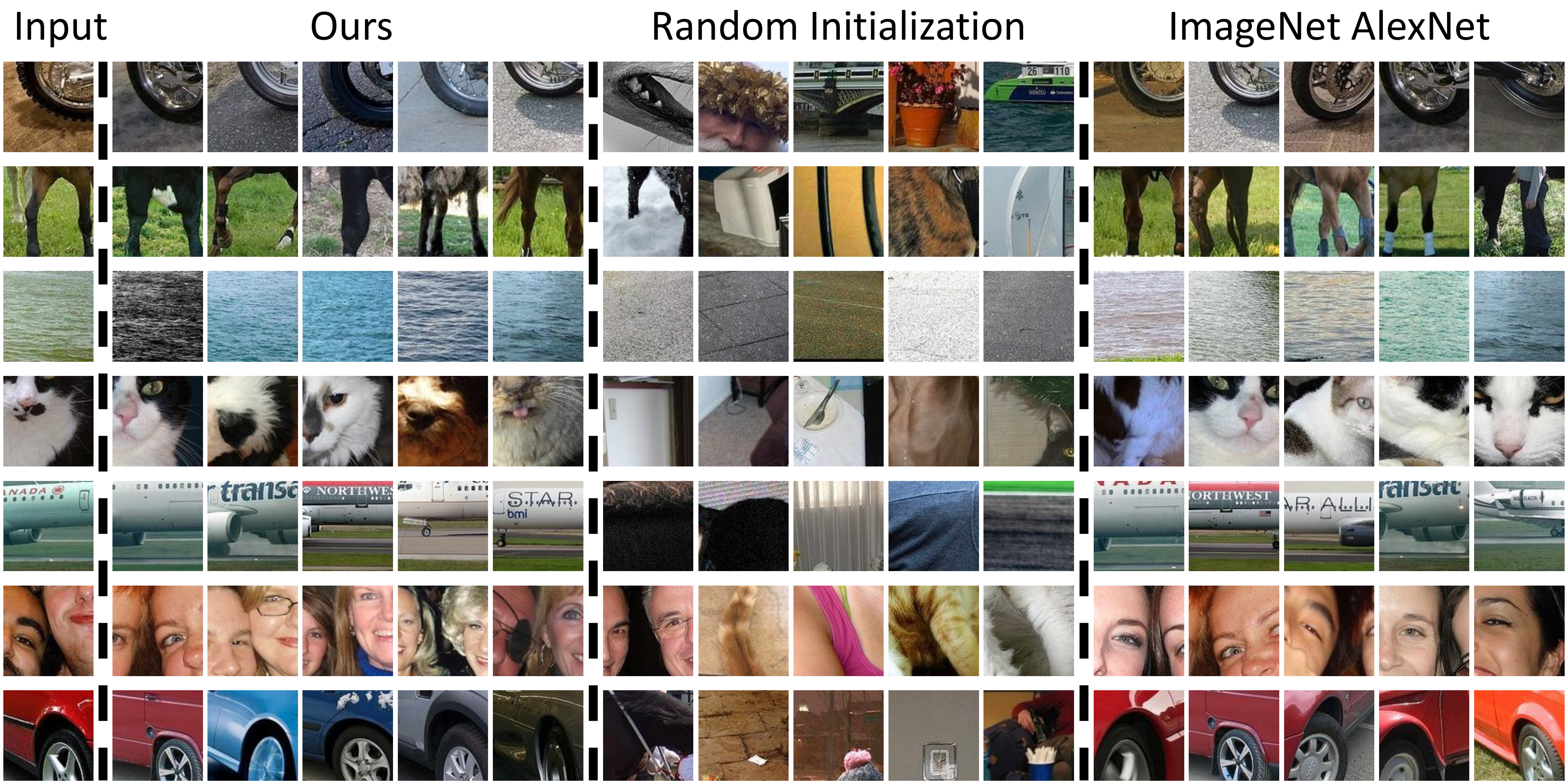


Chromatic Aberration





What is learned?



Visual Data Mining?

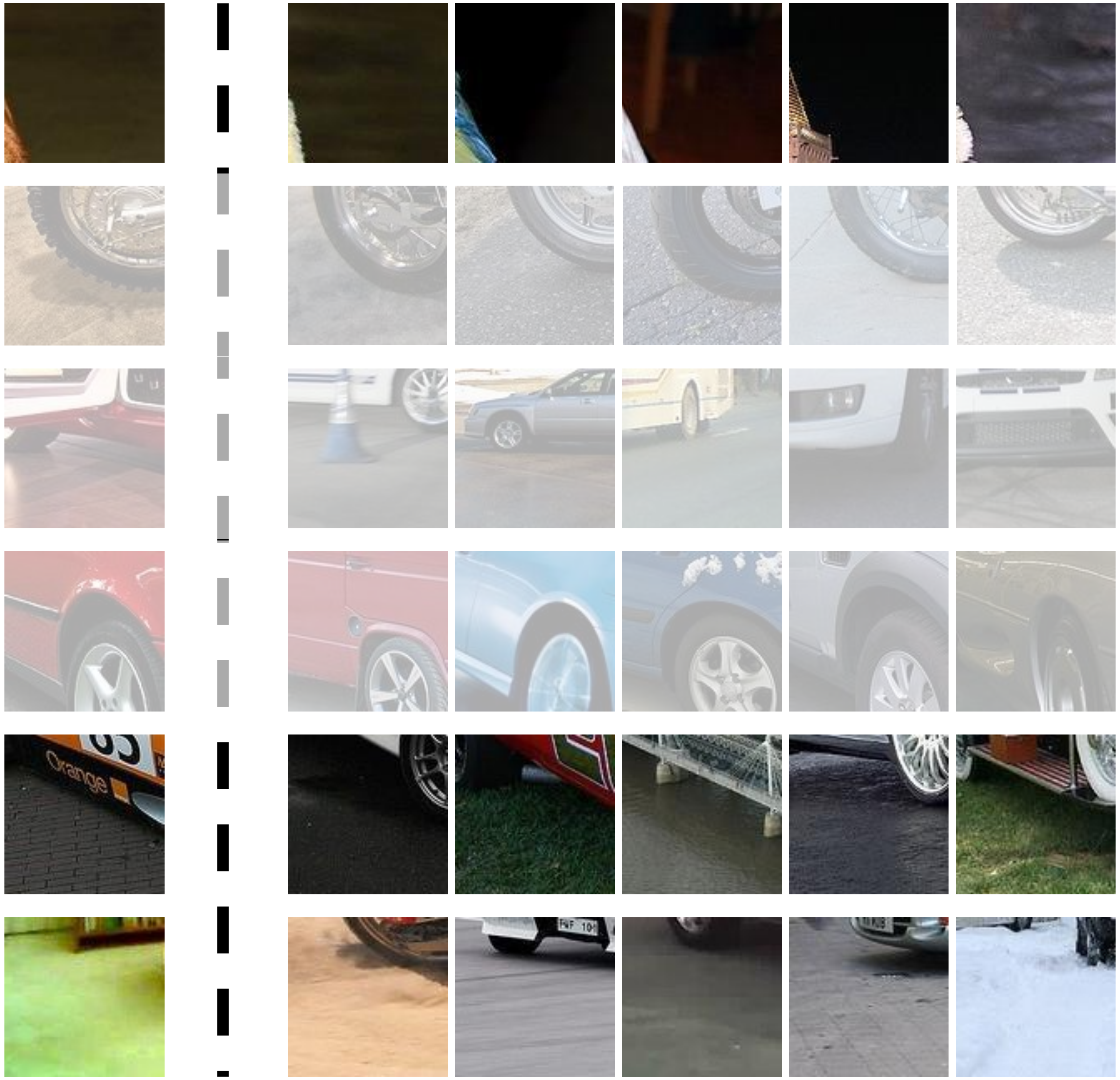
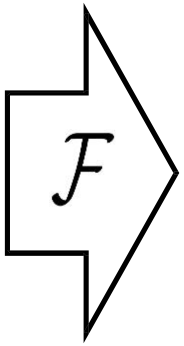


Image example II: colourization

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)

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

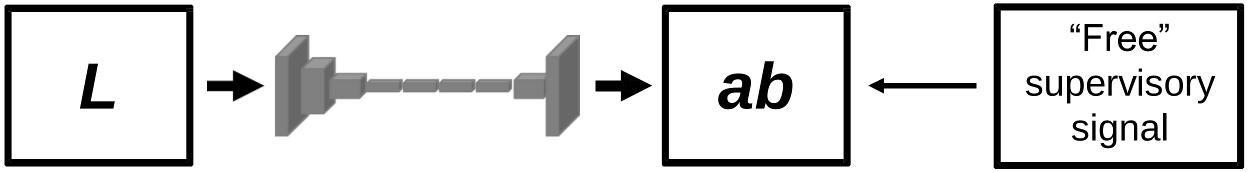
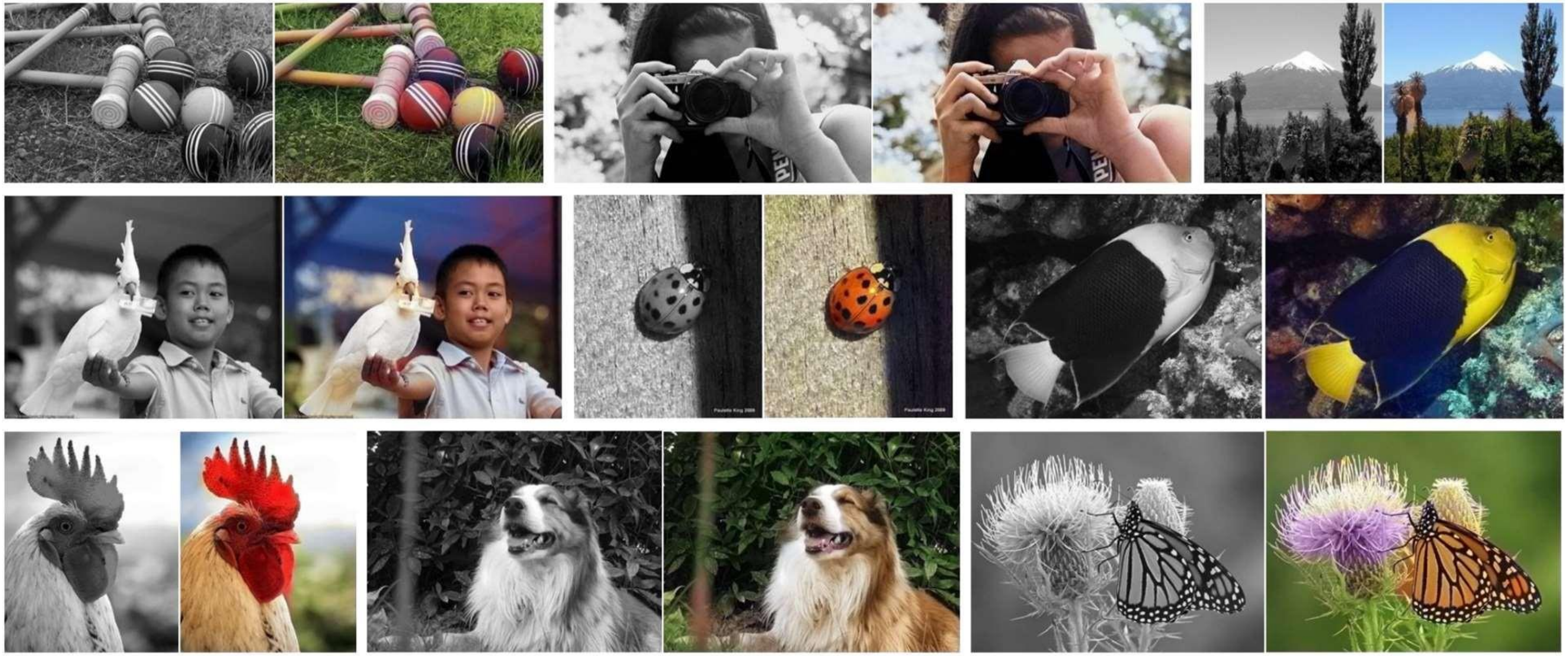


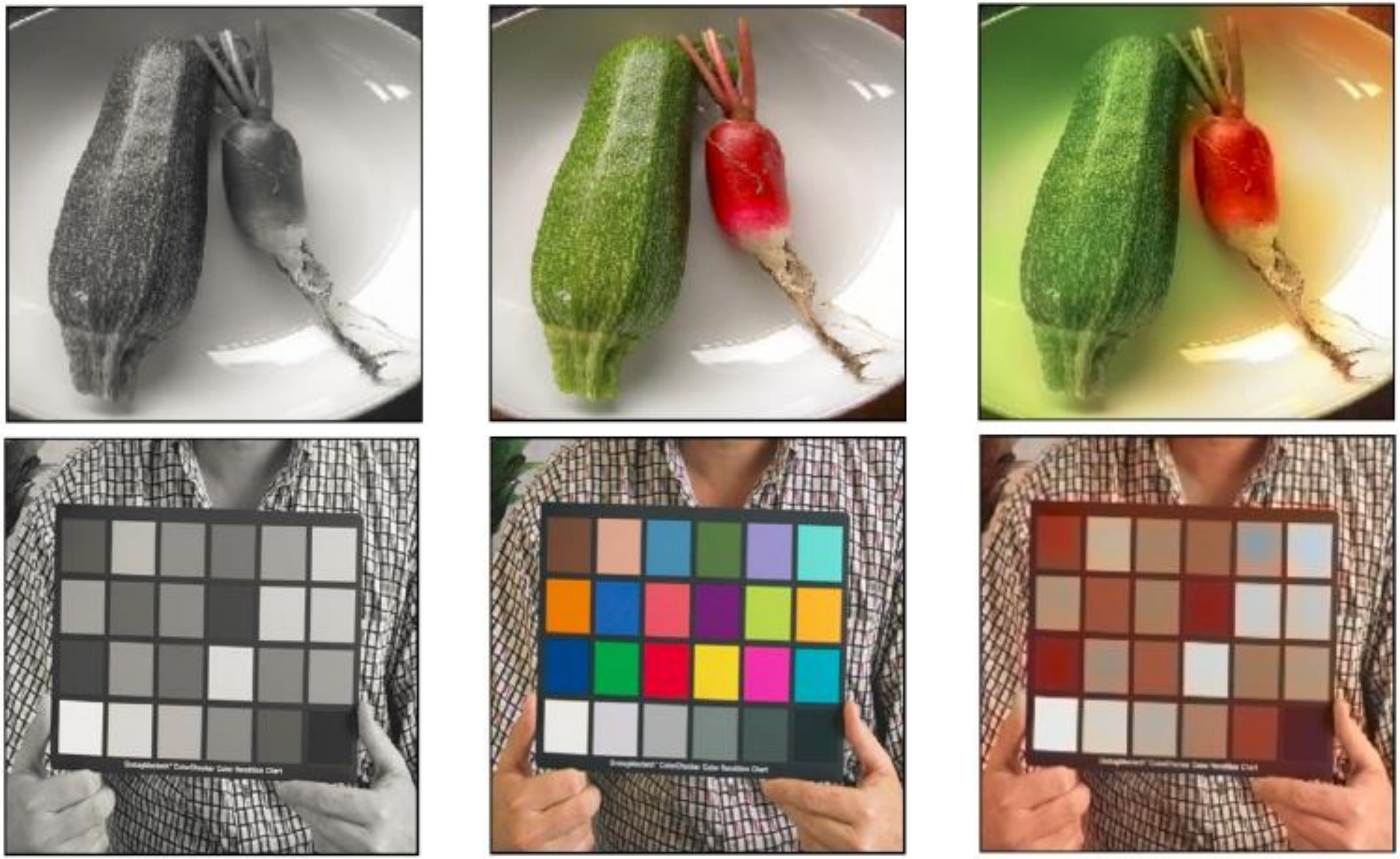
Image example II: colourization

Train network to predict pixel colour from a monochrome input



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

Input **Ground Truth** **Output**



Inherent Ambiguity

Our Output

Ground Truth

Biases

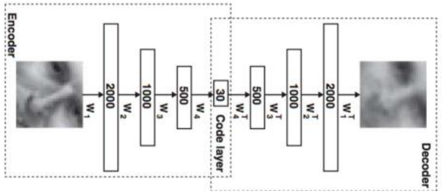
Image example III: exemplar networks

- Exemplar Networks (Dosovitskiy *et al*., 2014)

- Perturb/distort image patches, e.g. by cropping and affine transformations

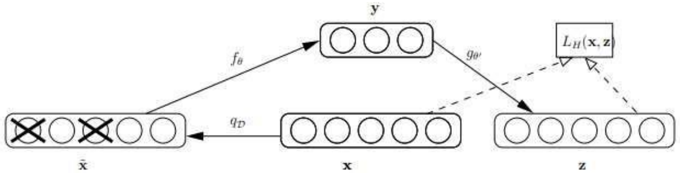
- Train to classify these exemplars as same class

Autoencoders



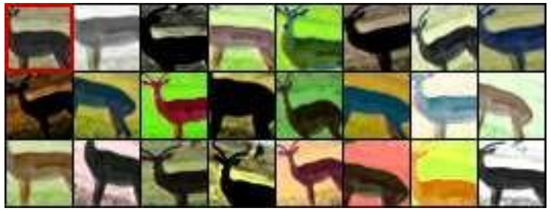
Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders



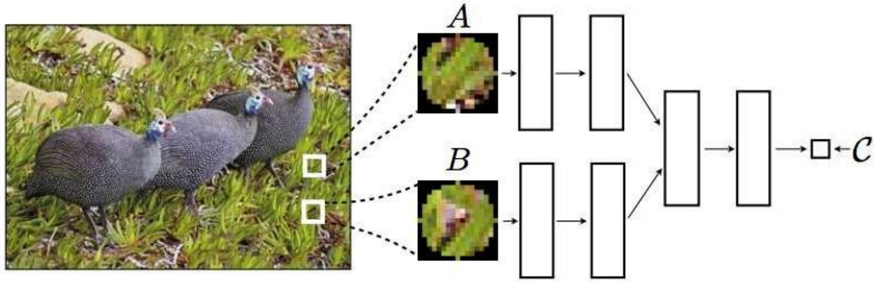
Vincent *et al.* ICML 2008.

Exemplar networks



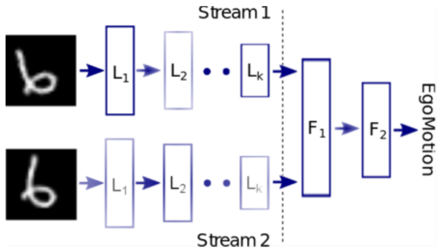
Dosovitskiy *et al*., NIPS 2014

Co-Occurrence



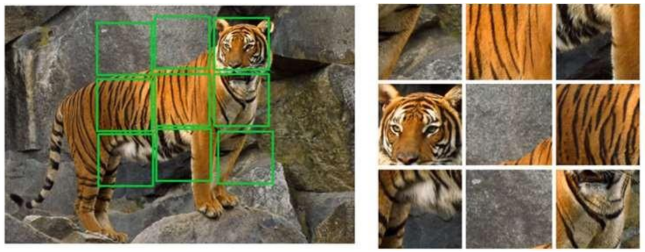
Isola *et al.* ICLR Workshop 2016.

Egomotion



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

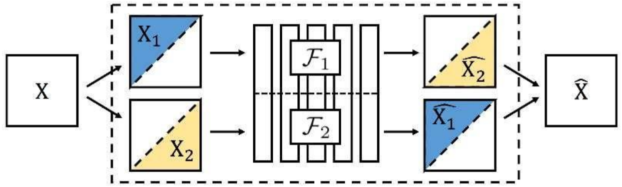
Context



Noroozi et al 2016

Pathak *et al.* CVPR 2016

Split-brain auto-encoders



Zhang *et al.* CVPR 2017

Multi-Task Self-Supervised Learning

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

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

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

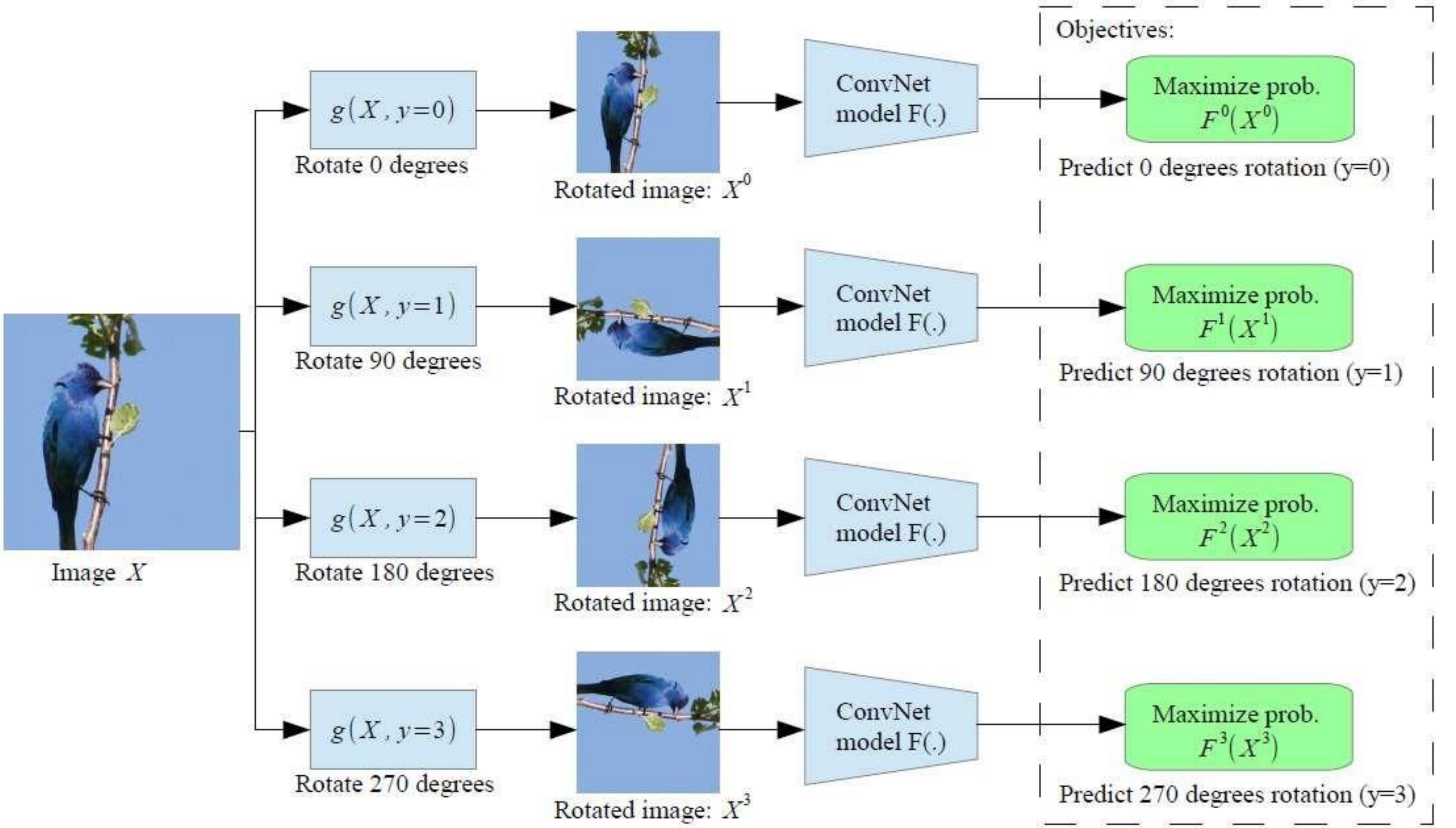
Image Transformations – 2018

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

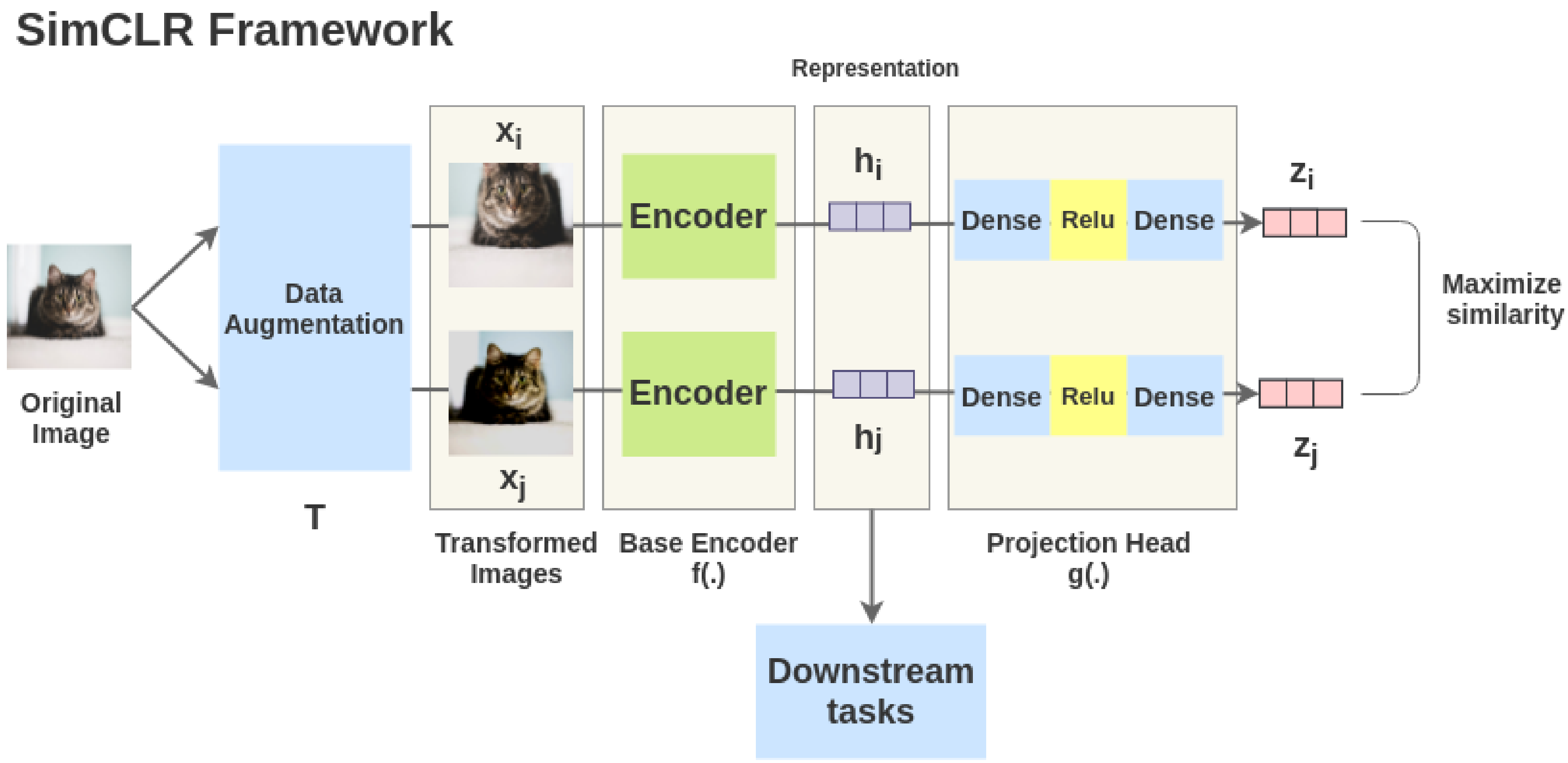
- Uses AlexNet
- Closes gap between ImageNet and self-supervision

	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

SimCLR: Contrastive Learning of Visual Representations

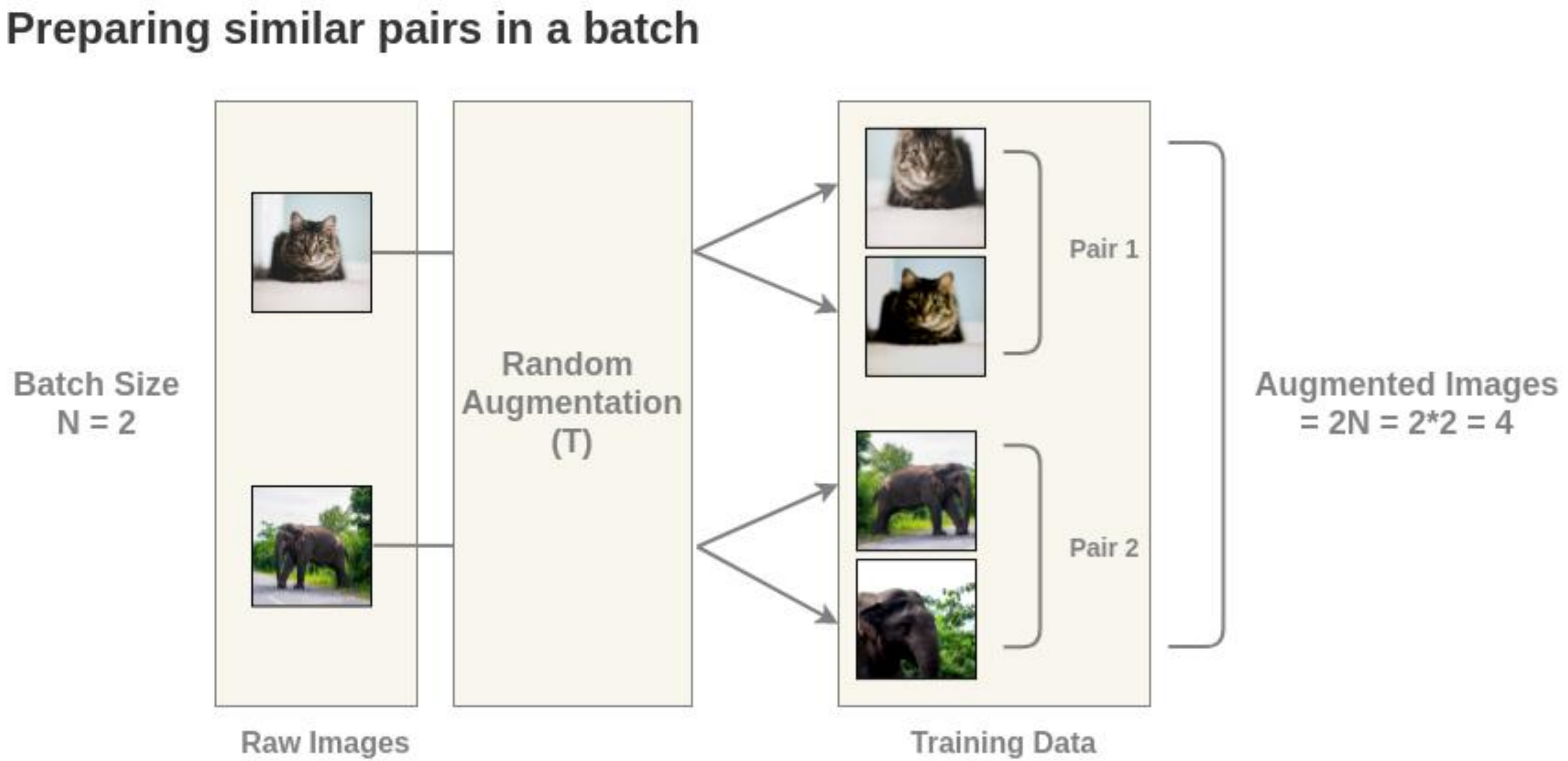
Overview



SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

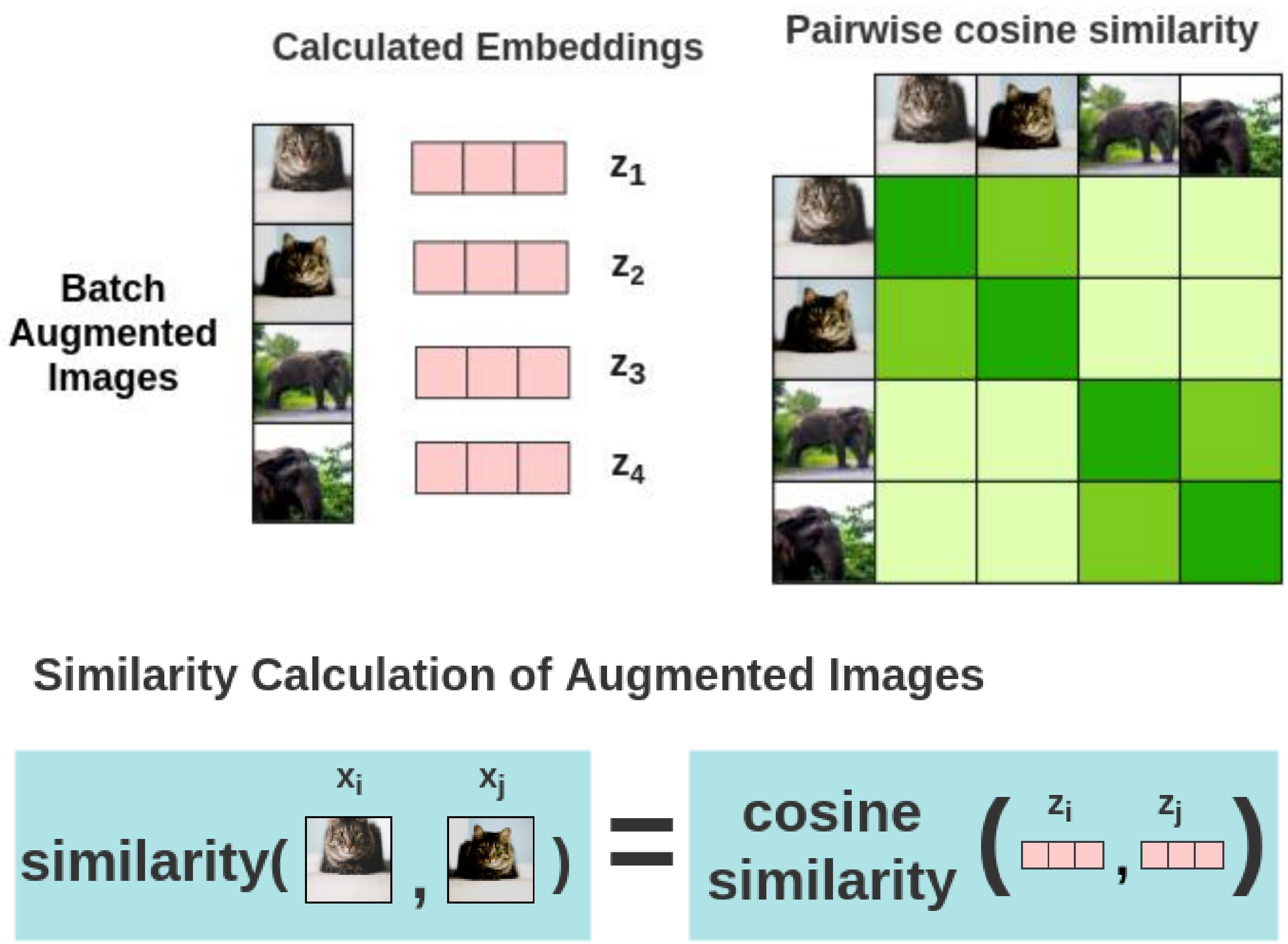
Augmentation



SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

Learning



$$s_{i,j} = \frac{z_i^T z_j}{(\tau \, ||z_i|| \, ||z_j||)}$$

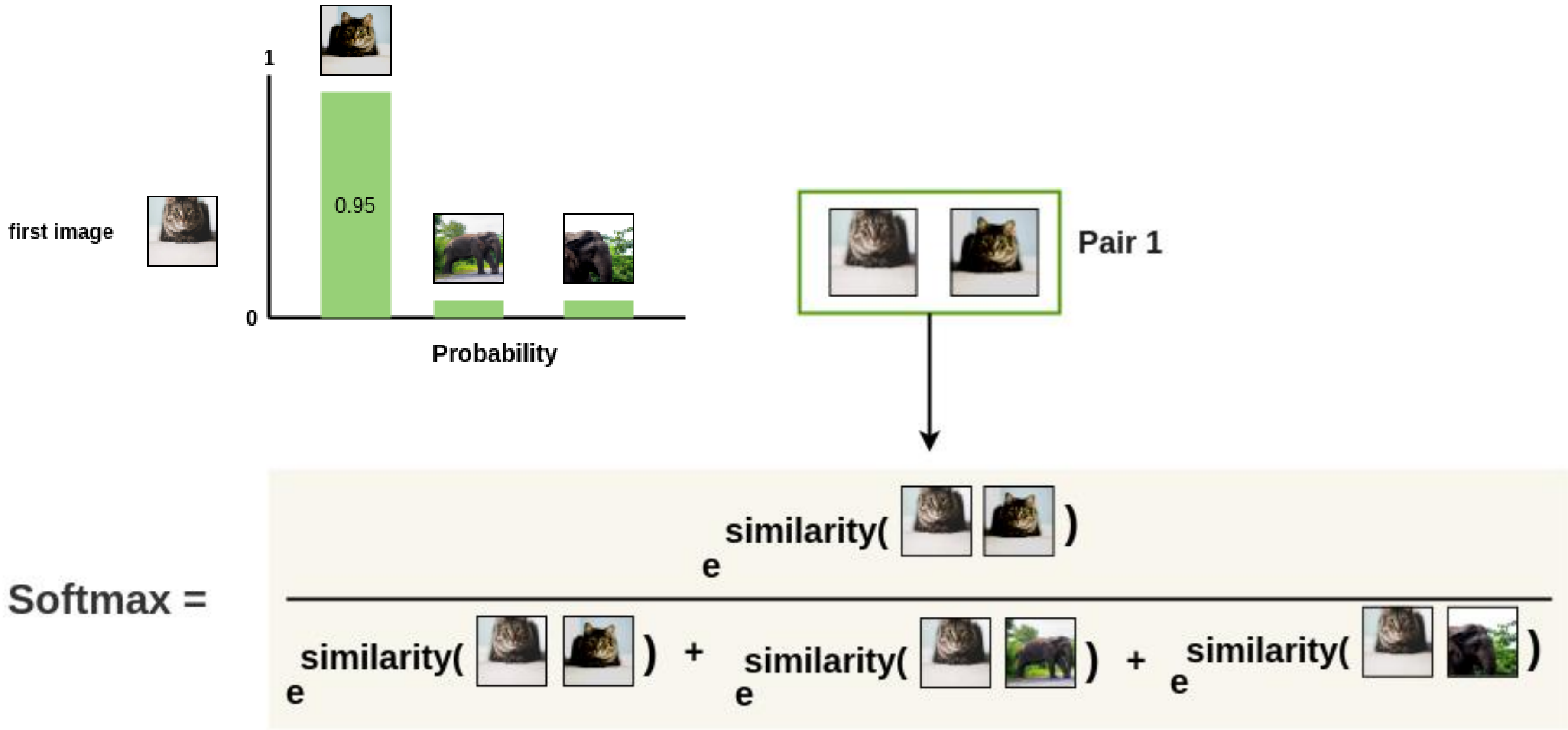
τ = temperature hyperparameter. It can scale the input and widen the range [-1, 1] of cosine similarity

||z|| = vector norm

SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

Learning



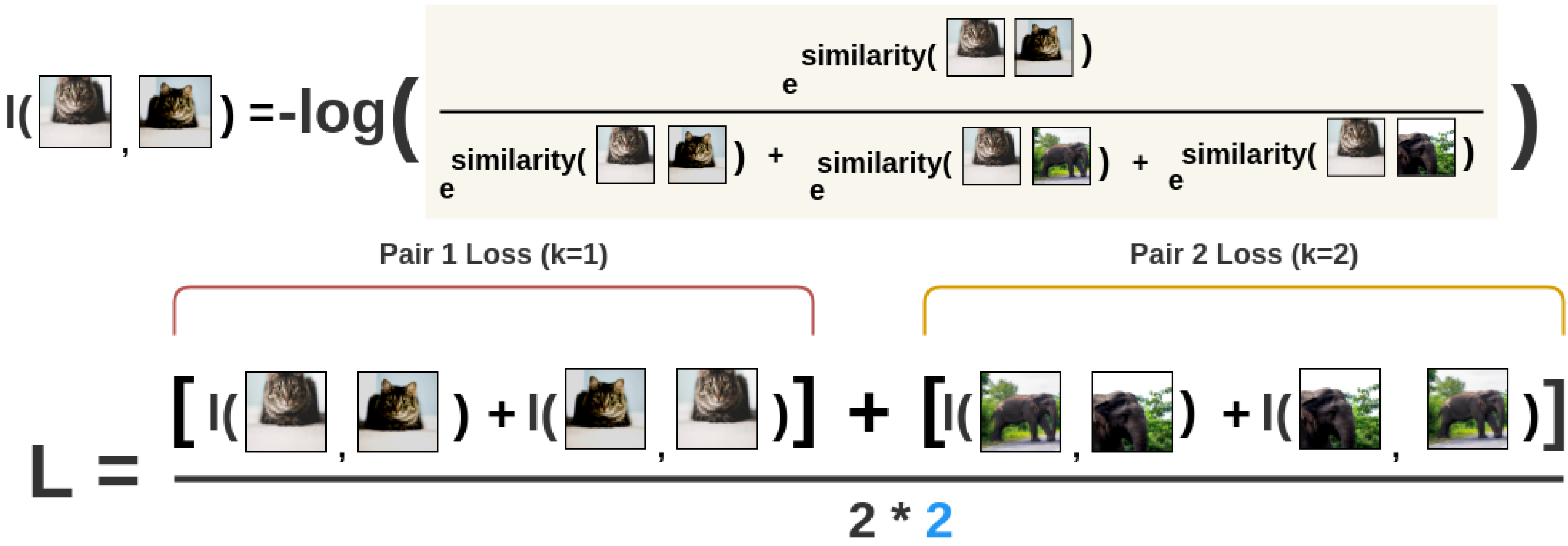
SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

Learning

NCE: Noise Contrastive Estimator

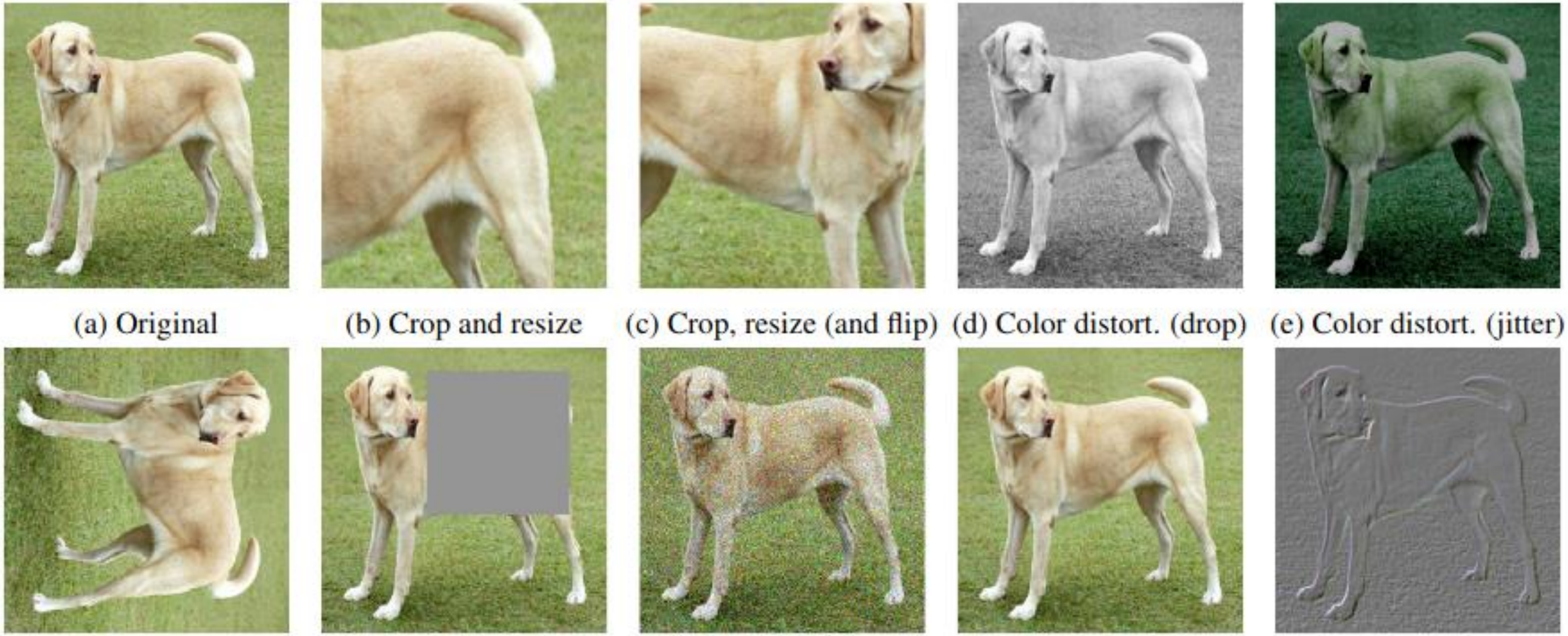
$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$



SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

Augmentations



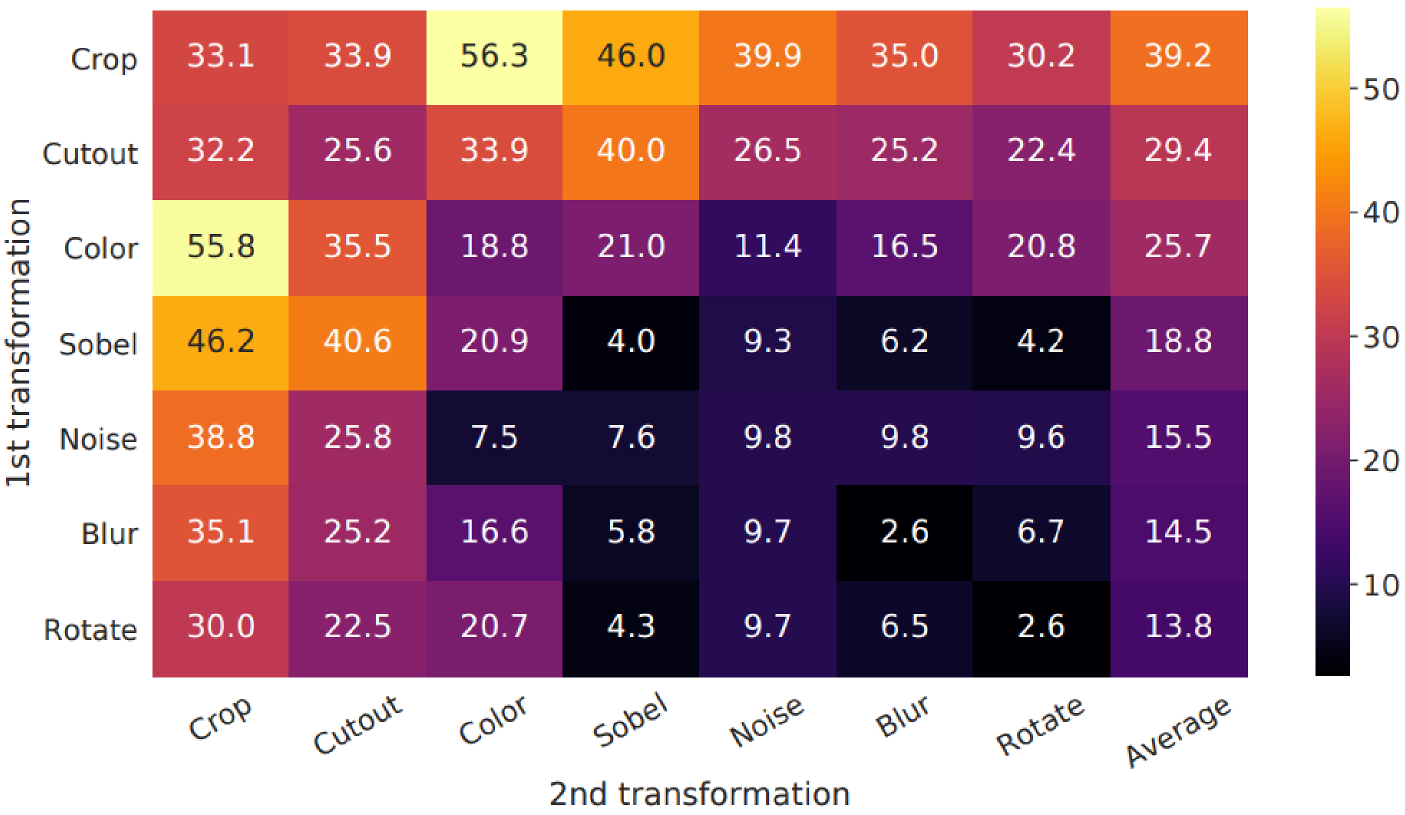
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

Augmentations

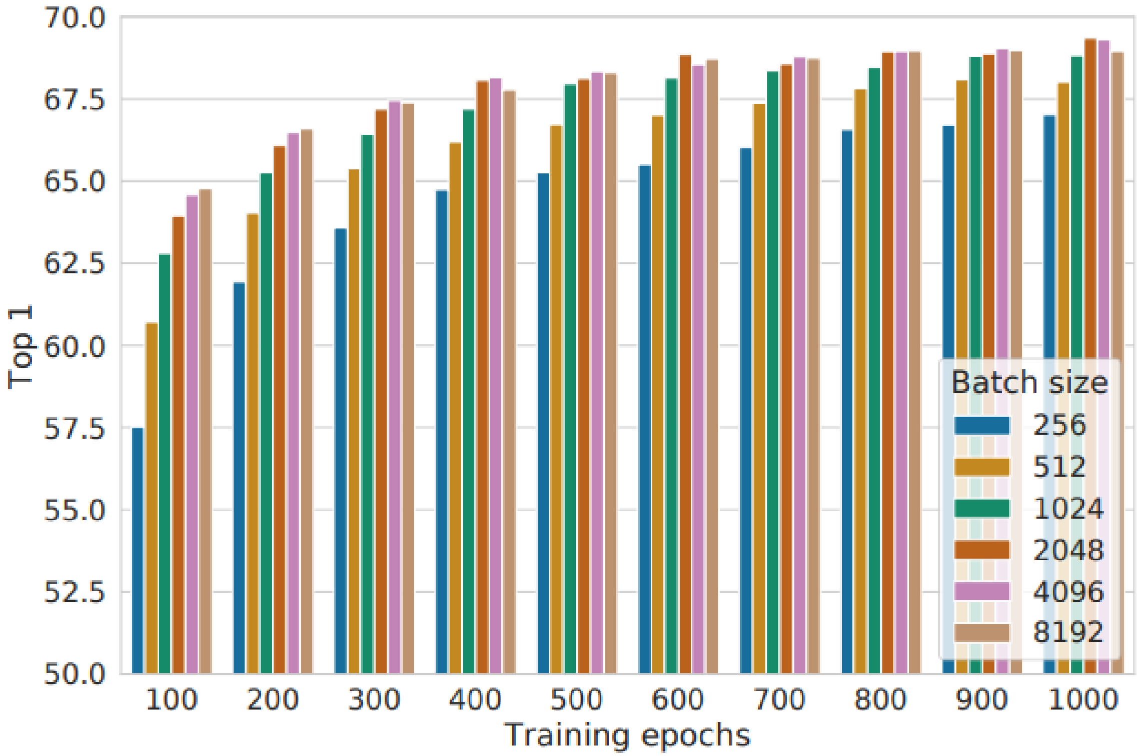


Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations

SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

Evaluation

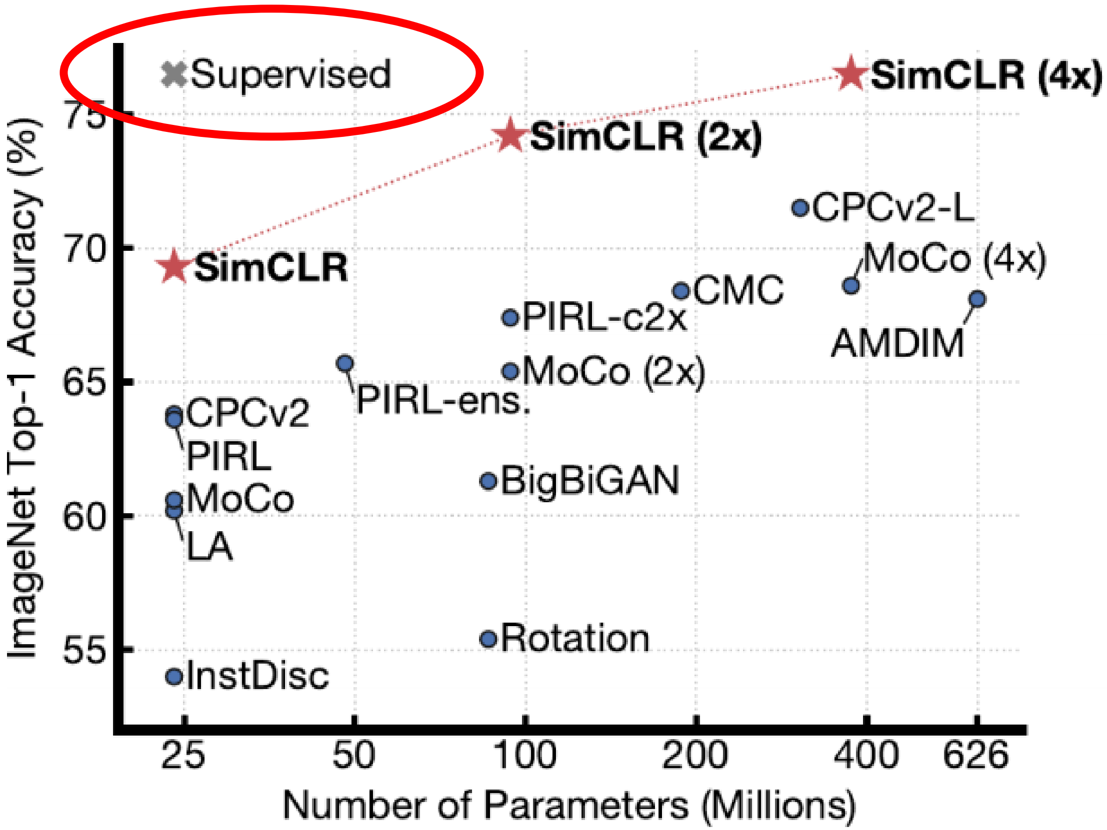


- Larger batch size
- Longer training

SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

SimCLR: Contrastive Learning of Visual Representations

Evaluation



ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet)

SimCLR, A Simple Framework for Contrastive Learning of Visual Representations
Chen T, Kornblith S, Norouzi M, Hinton G., ICML 2020

Masked autoencoders are scalable vision learners

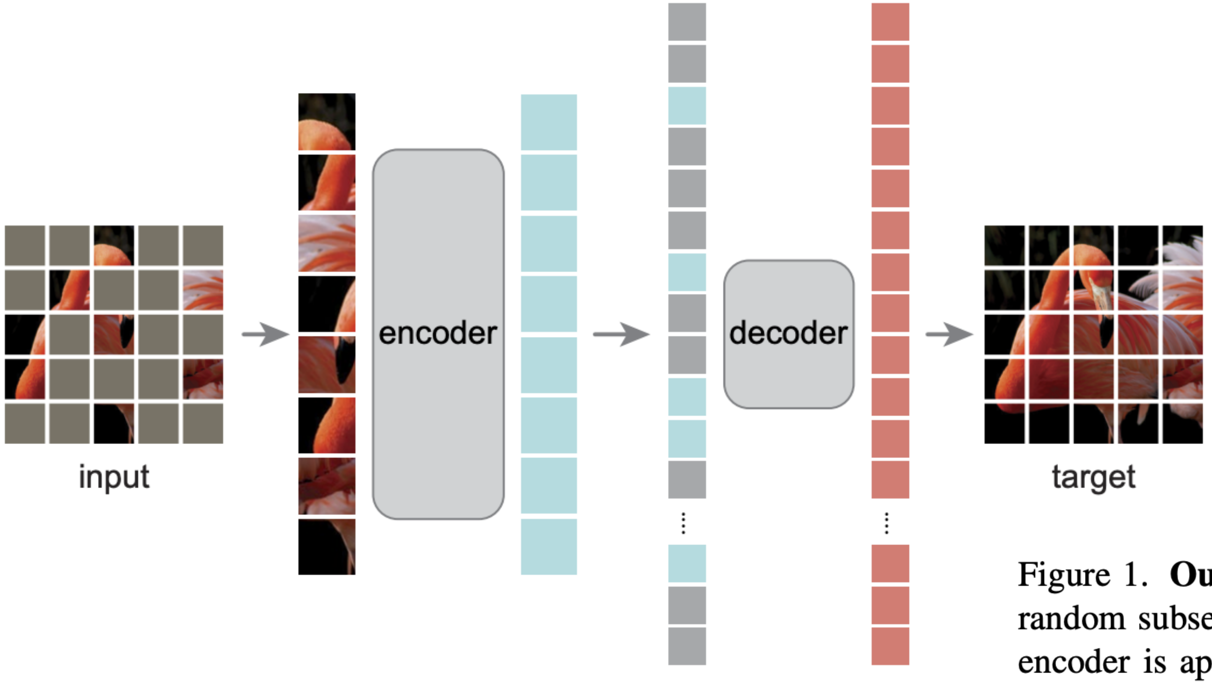


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images to produce representations for recognition tasks.

K. He et al. [Masked autoencoders are scalable vision learners](). CVPR 2022

Masked autoencoders are scalable vision learners

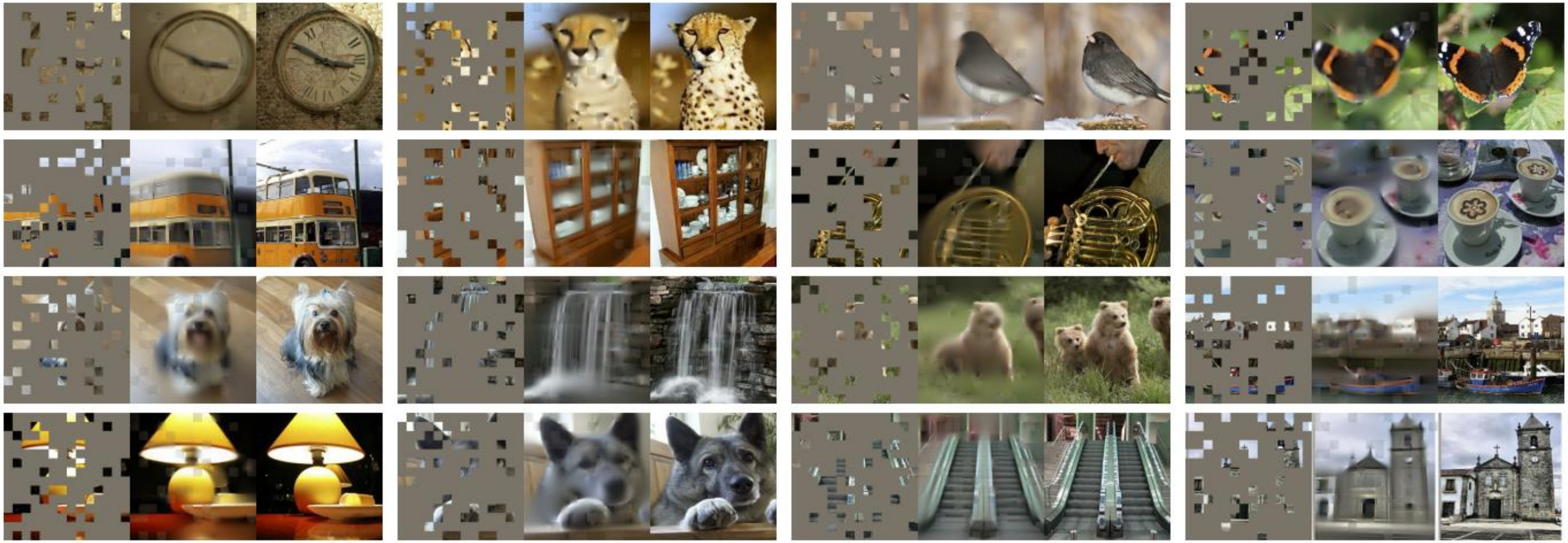


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix. †*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

K. He et al. [Masked autoencoders are scalable vision learners](). CVPR 2022

Masked autoencoders are scalable vision learners

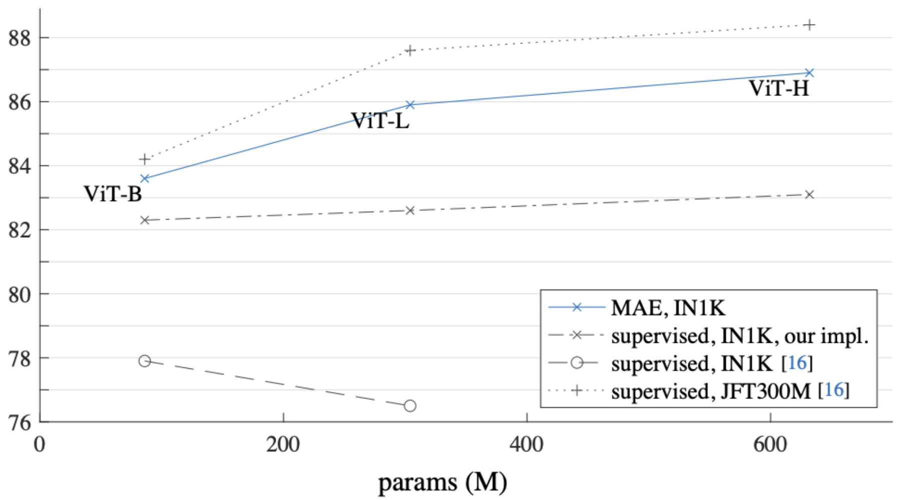


Figure 8. **MAE pre-training *vs*. supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

K. He et al. Masked autoencoders are scalable vision learners. CVPR 2022

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 B

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:

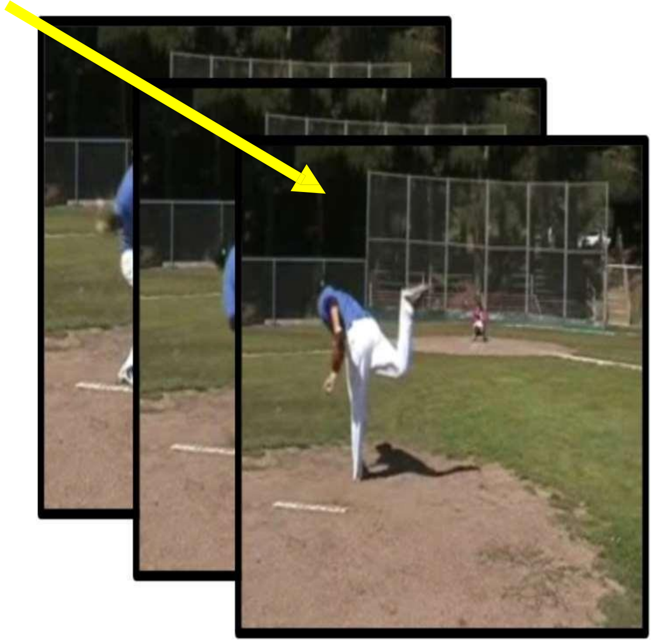
- Video sequence order
- Video direction
- Video tracking

Temporal structure in videos

Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

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

Time



“Sequence” of data

Slide credit: Ishan Misra

Sequential Verification

- Is this a valid sequence?



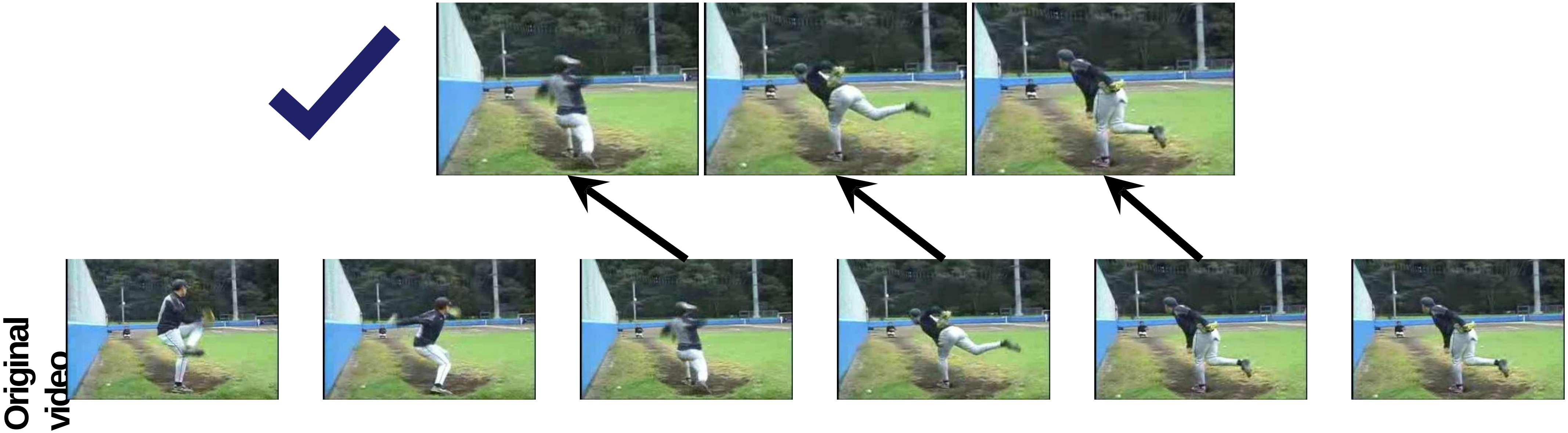
Sun and Giles, 2001; Sun et al., 2001; Cleermans 1993; Reber 1989 Arrow of Time - Pickup et al., 2014

Slide credit: Ishan Misra

Original video

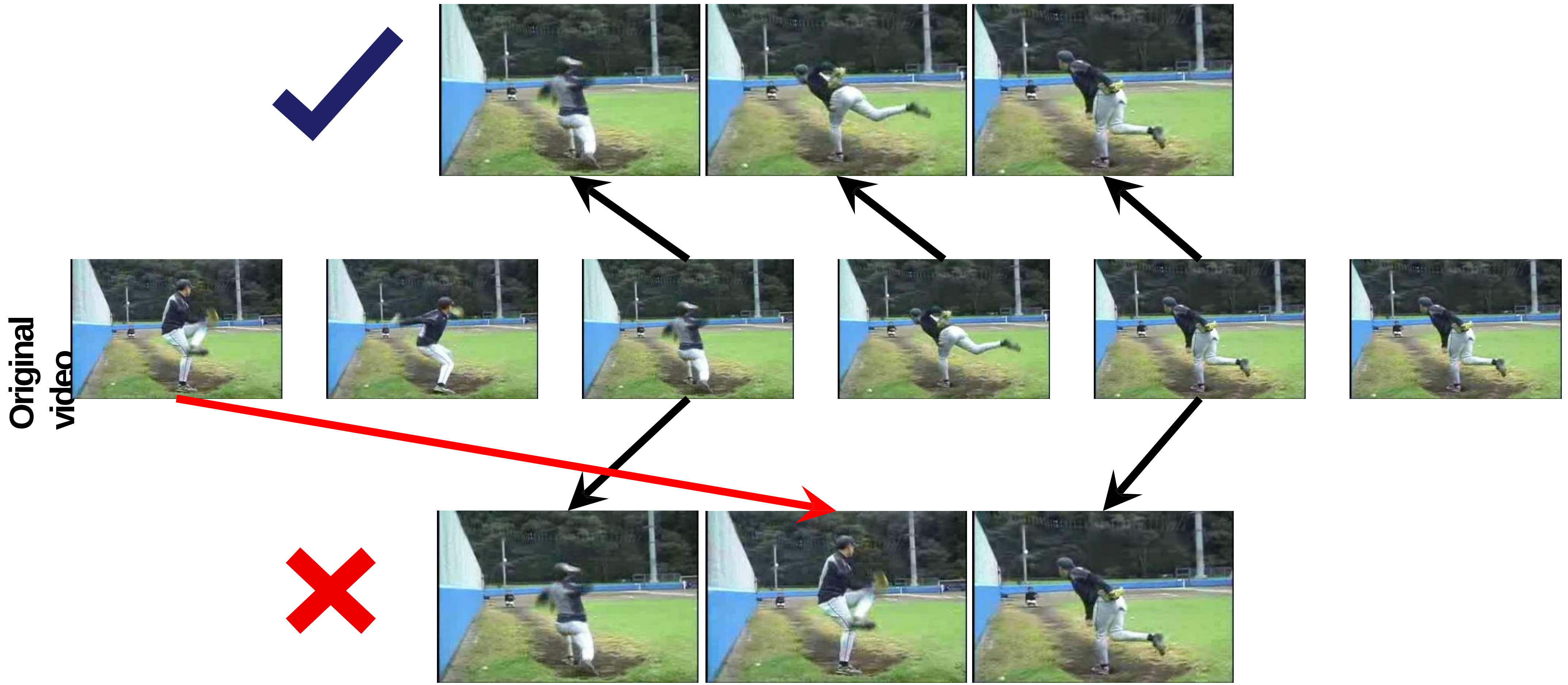
Slide credit: Ishan Misra

Temporally Correct order



Slide credit: Ishan Misra

Temporally Correct order

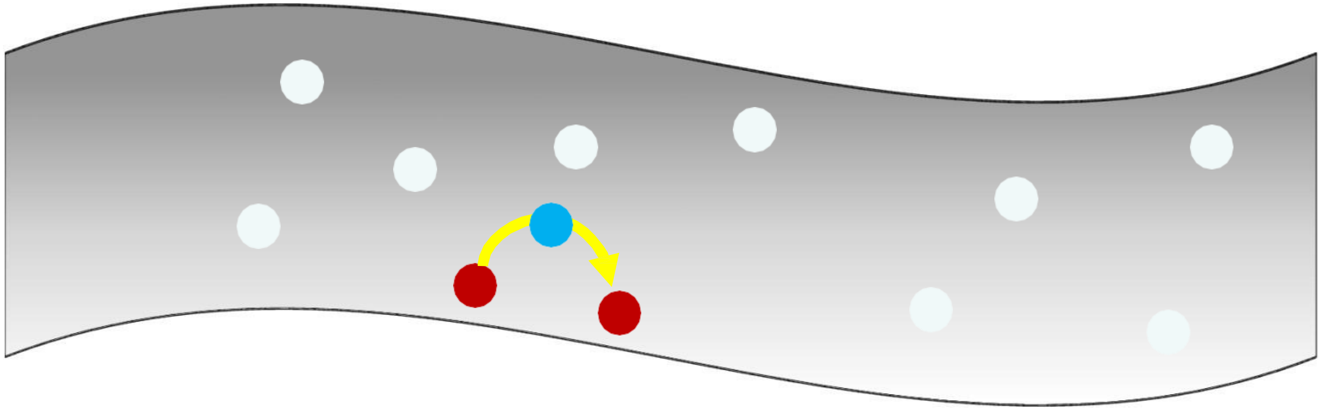


Temporally Incorrect order

Slide credit: Ishan Misra

Geometric View

Images



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

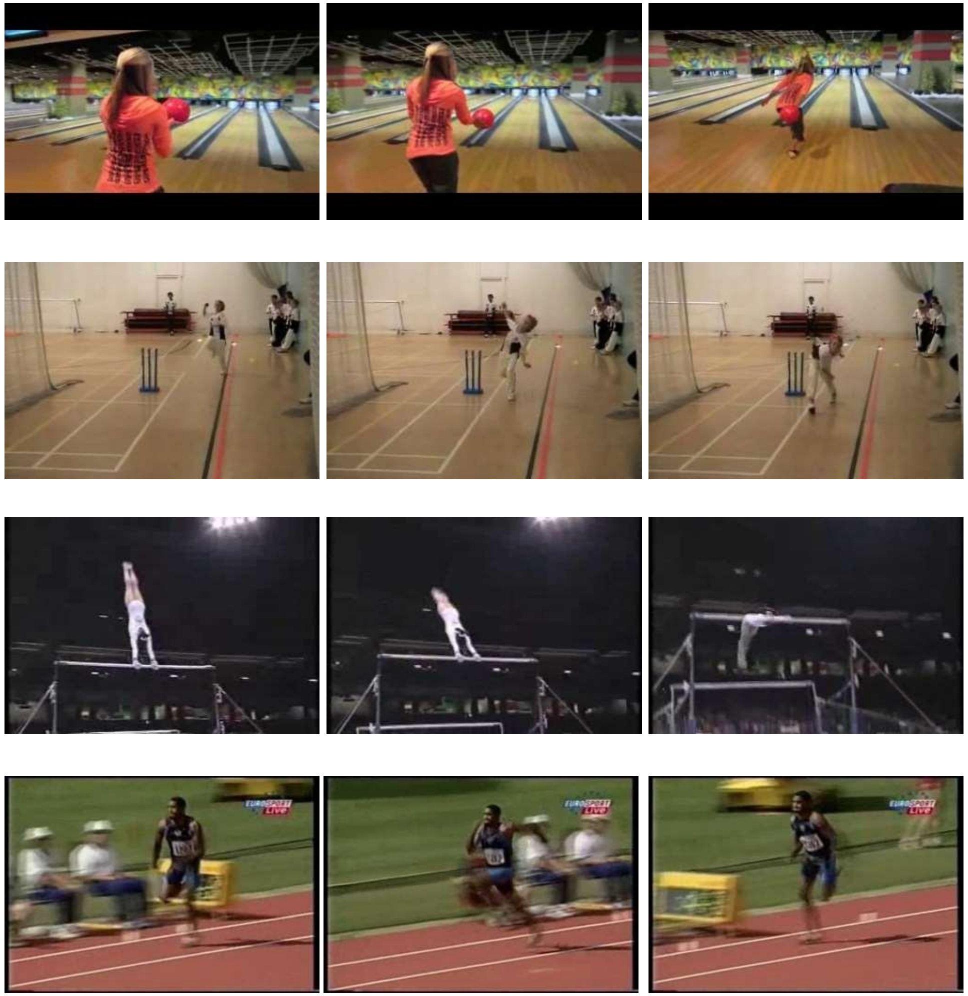
Shuffle and Learn – I. Misra, L. Zitnick, M. Hebert – ECCV 2016

Slide credit: Ishan Misra

Dataset: UCF-101 Action Recognition

UCF101 - Soomro et al., 2012

Positive Tuples

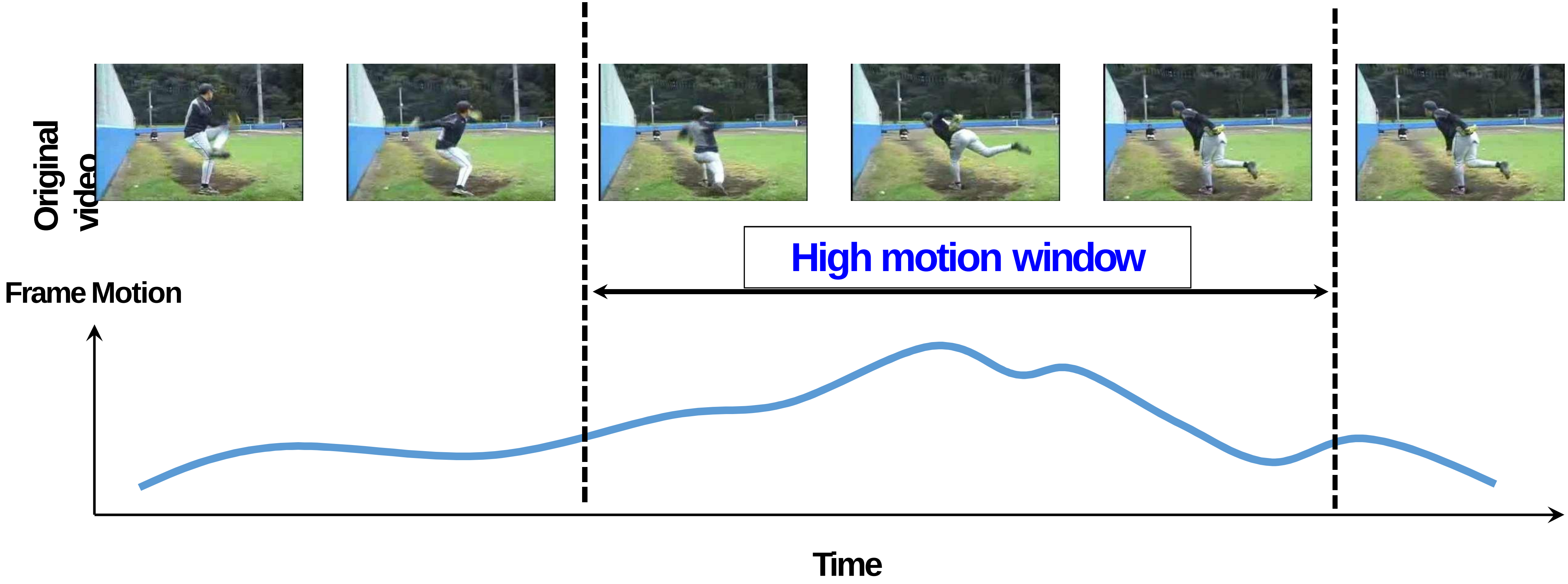


Negative Tuples

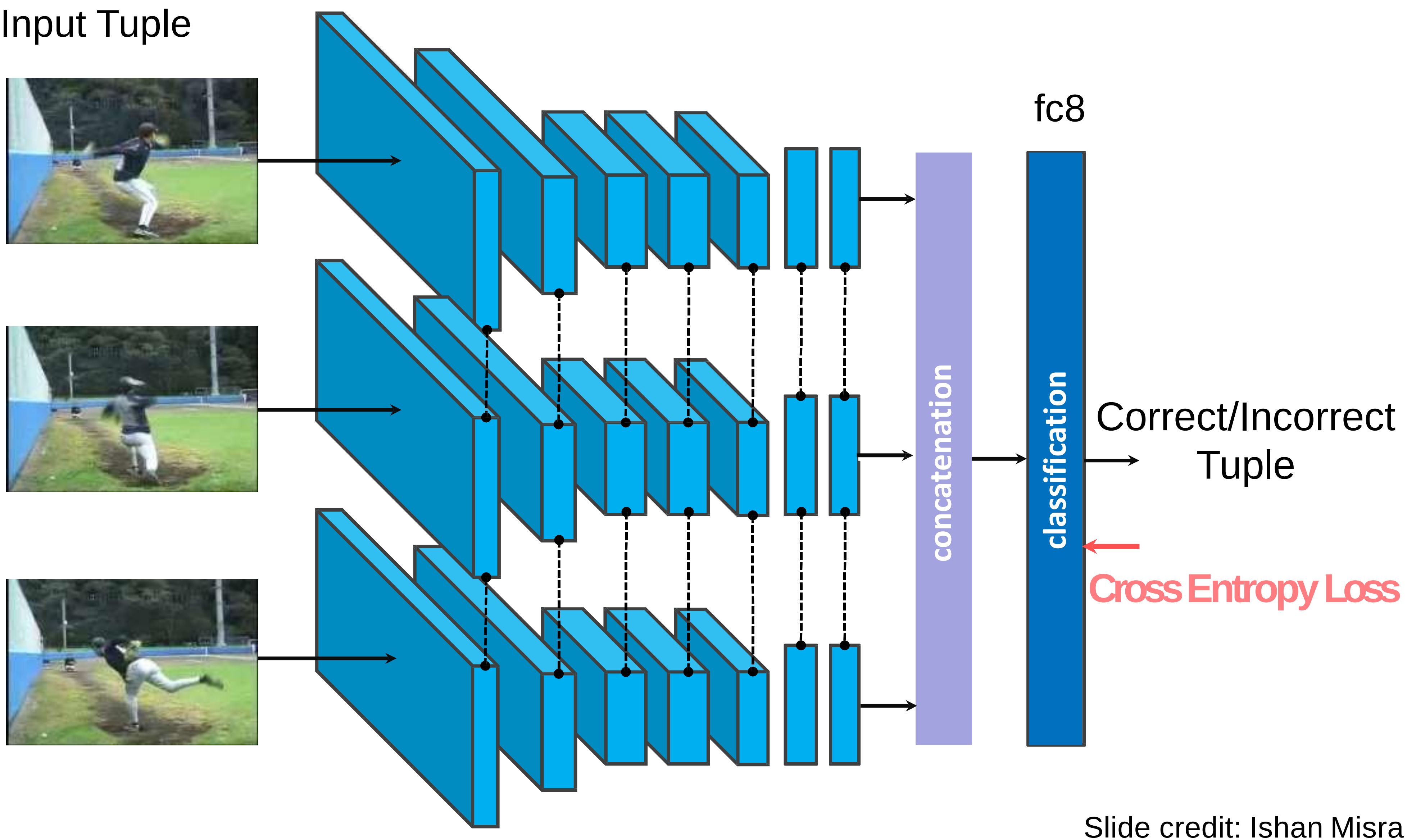
~900k tuples from UCF-101 dataset (Soomro et al., 2012)

Slide credit: Ishan Misra

Informative training tuples



Slide credit: Ishan Misra



Slide credit: Ishan Misra

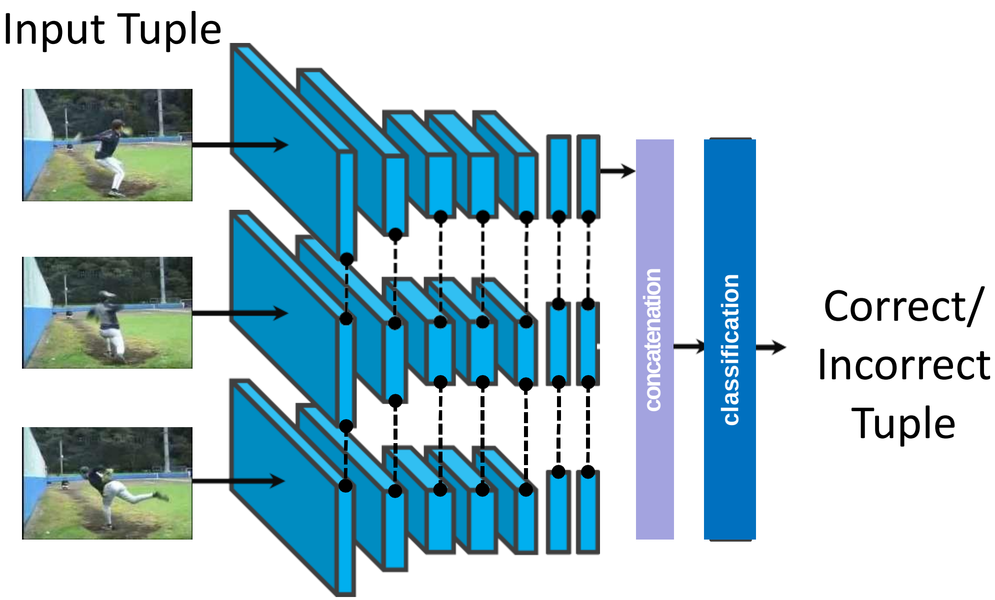
Nearest Neighbors of Query Frame (fc7 features)

Query | **ImageNet** | **Shuffle & Learn** | **Random**

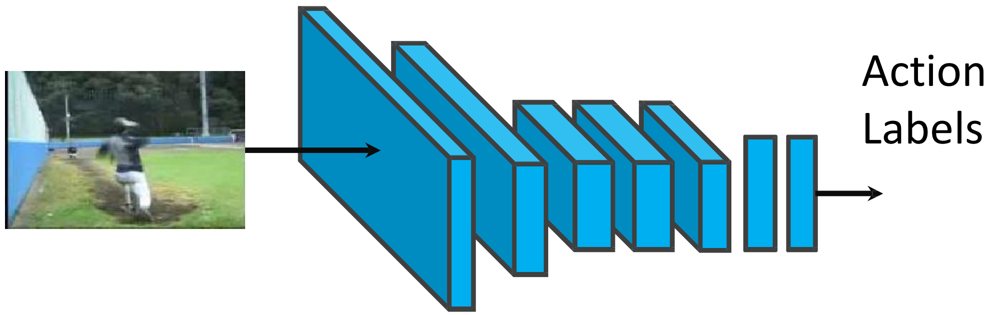
Slide credit: Ishan Misra

Finetuning setup

Self-supervised Pre-train



Test -> Finetune



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

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

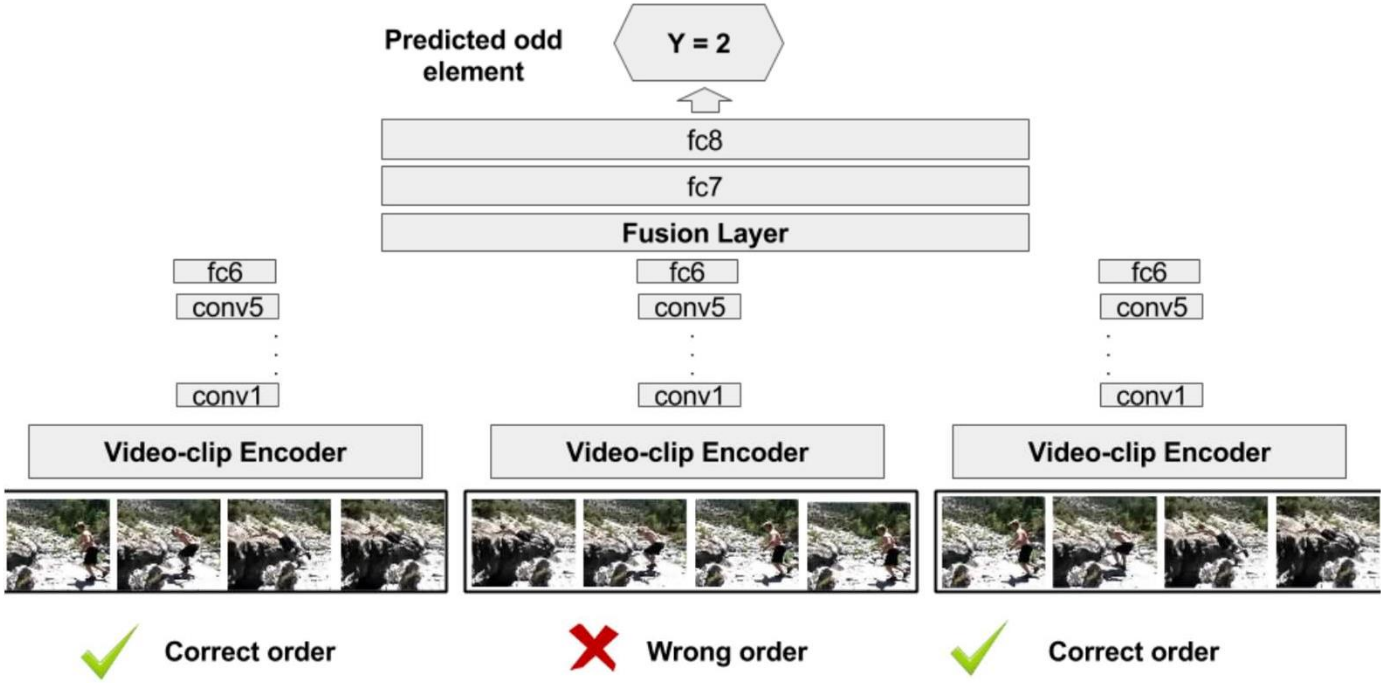
FLIC - Sapp & Taskar, 2013 MPII - Andriluka et al., 2014
Setup fom – Toshev et al., 2013

Slide credit: Ishan Misra

More temporal structure in videos

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

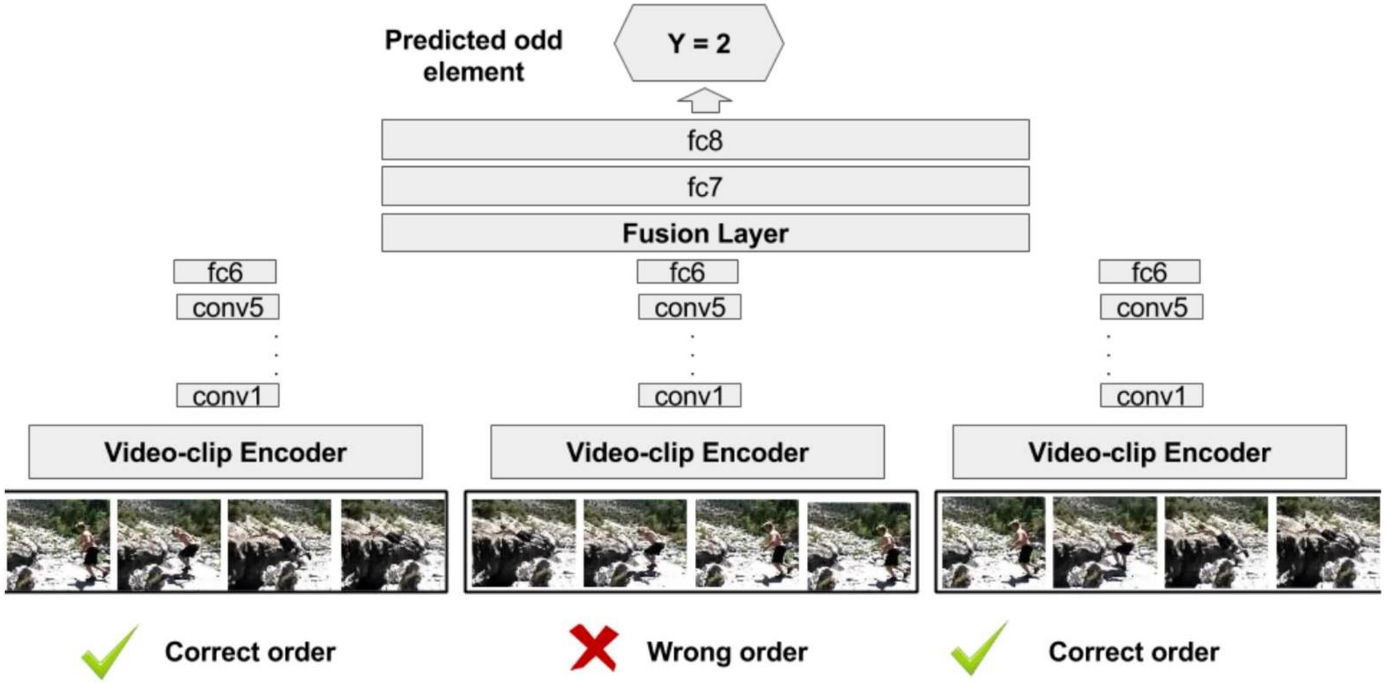
Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould, ICCV 2017



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

- 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

Part C

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**

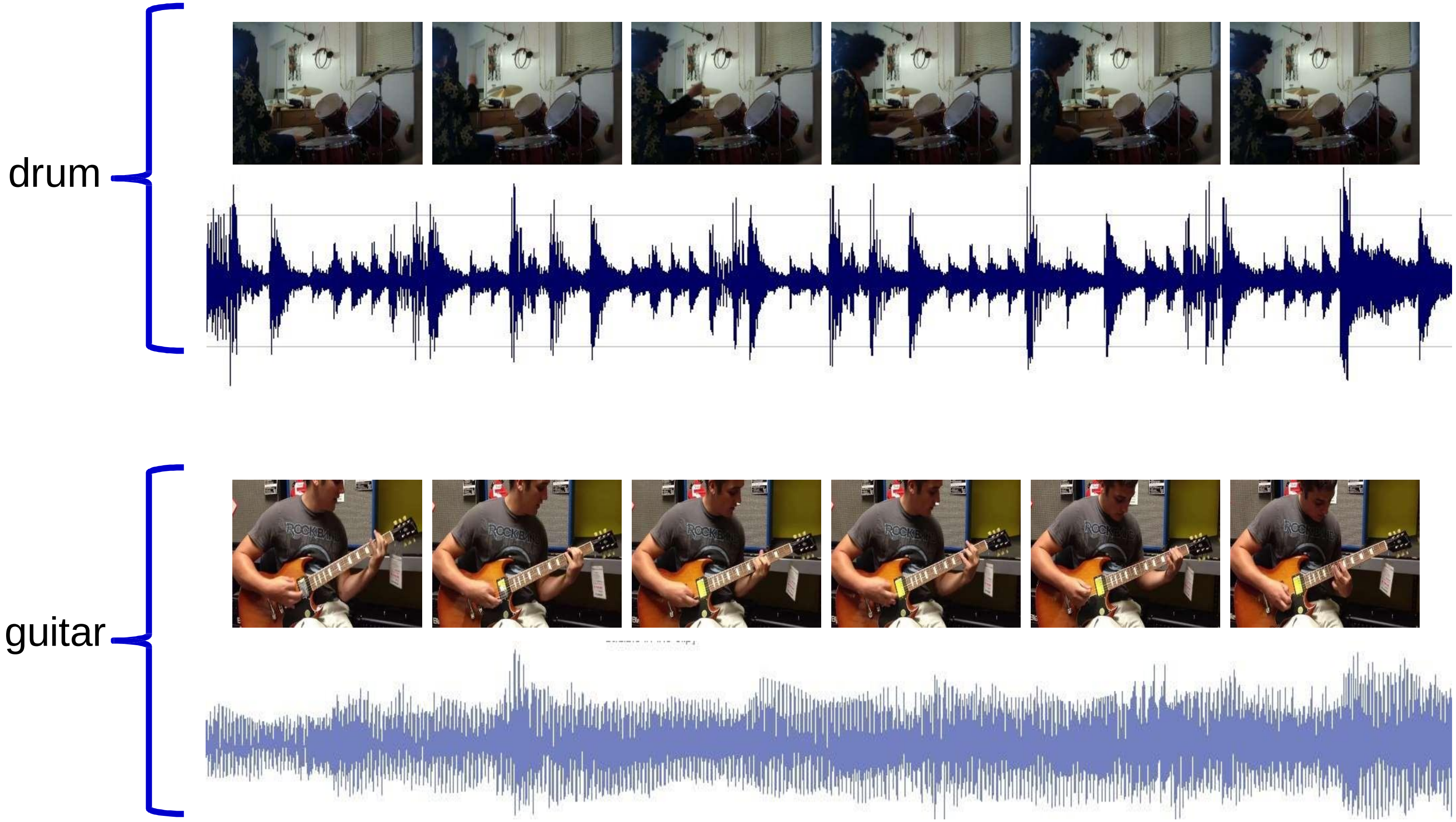
Audio-Visual Co-supervision

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

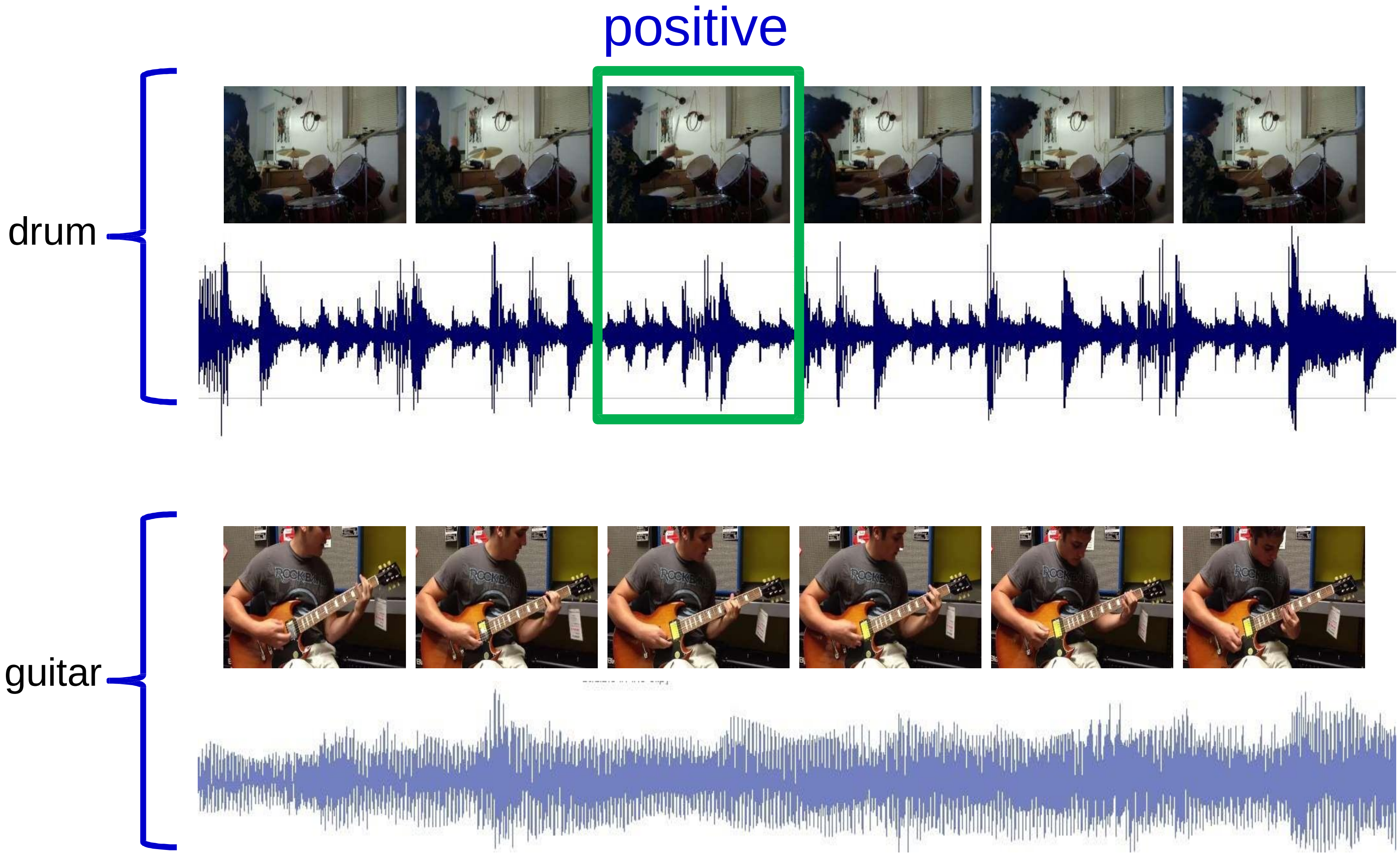
Correspond?

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

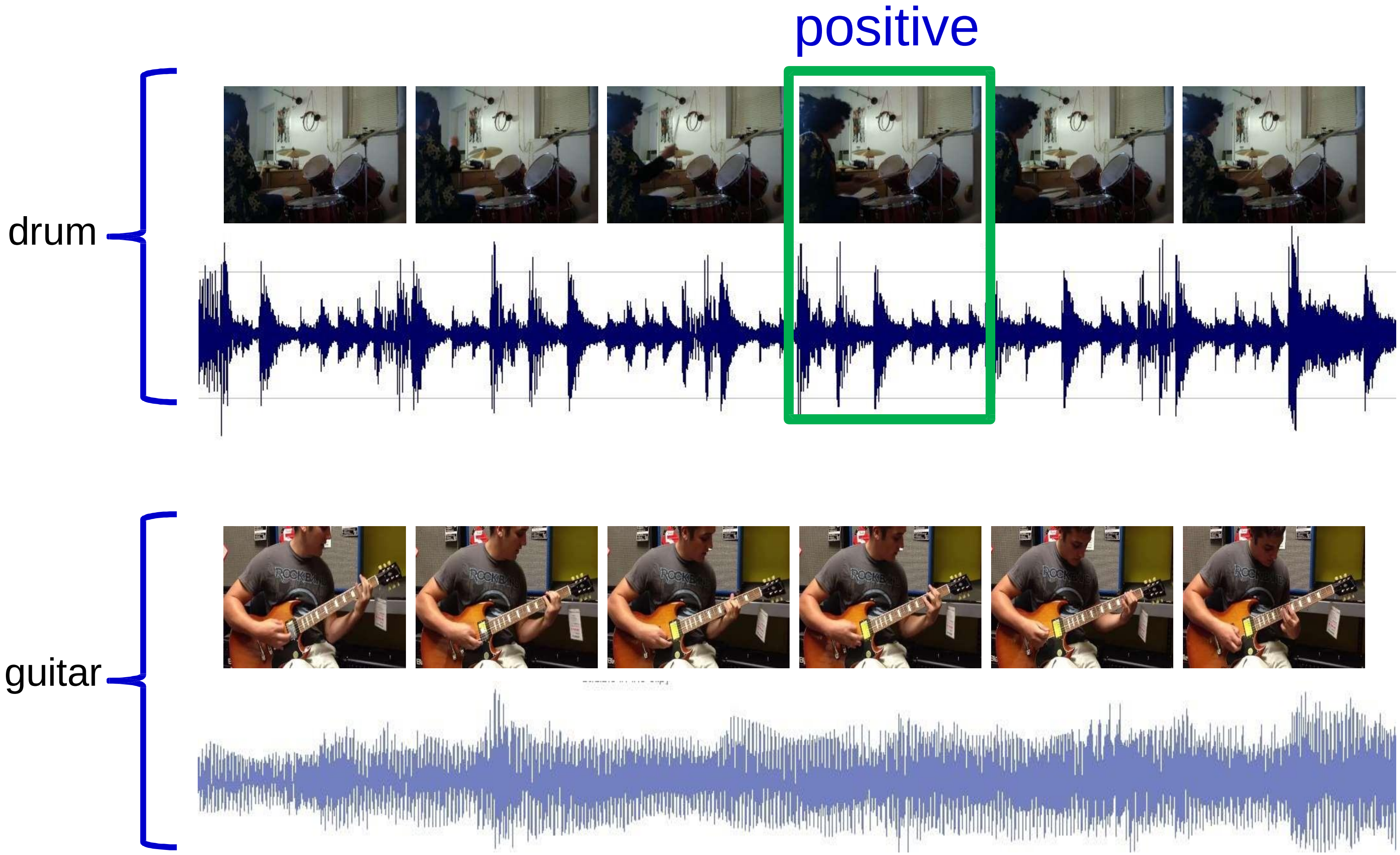
Audio-Visual Correspondence



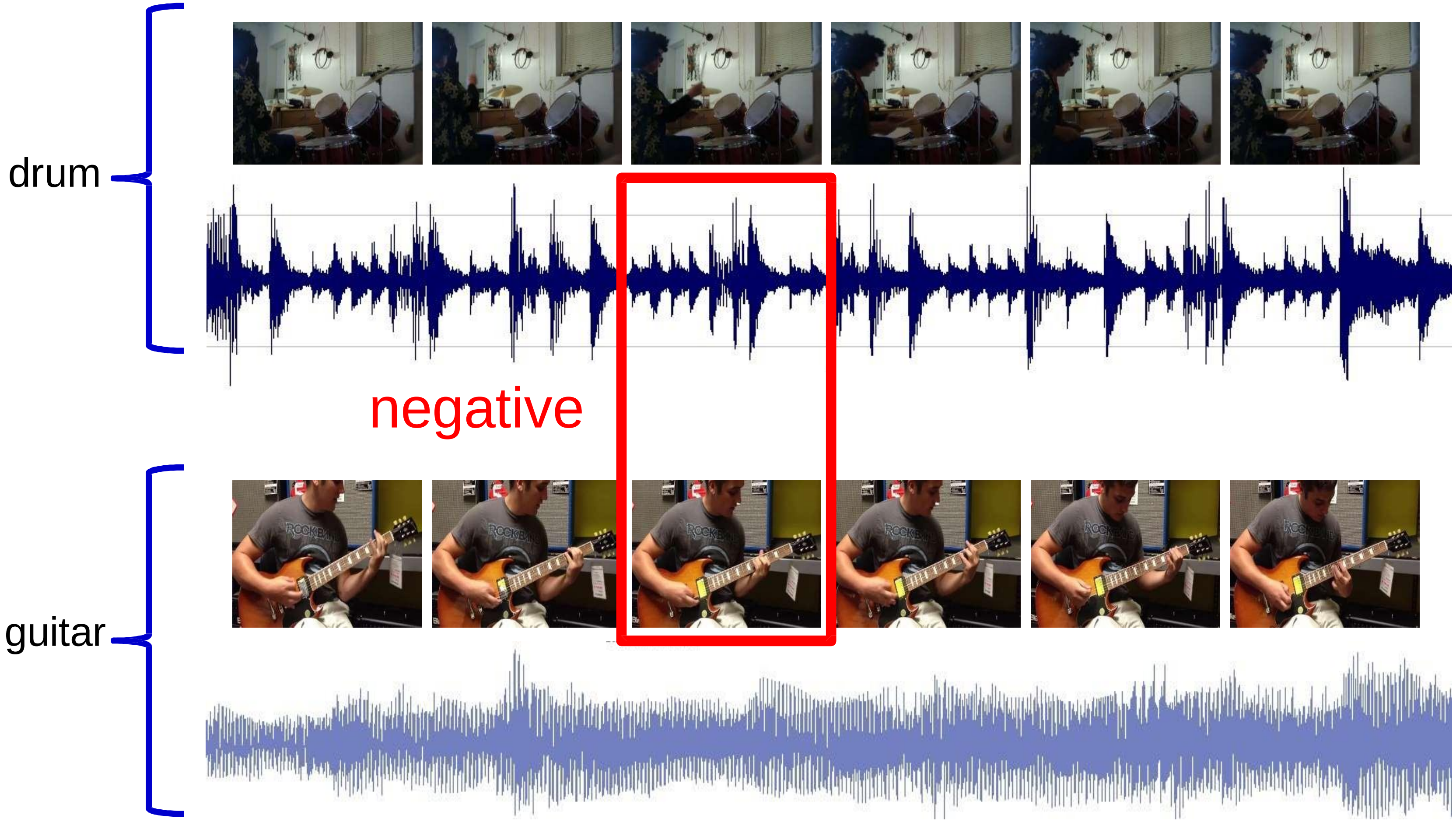
Audio-Visual Correspondence



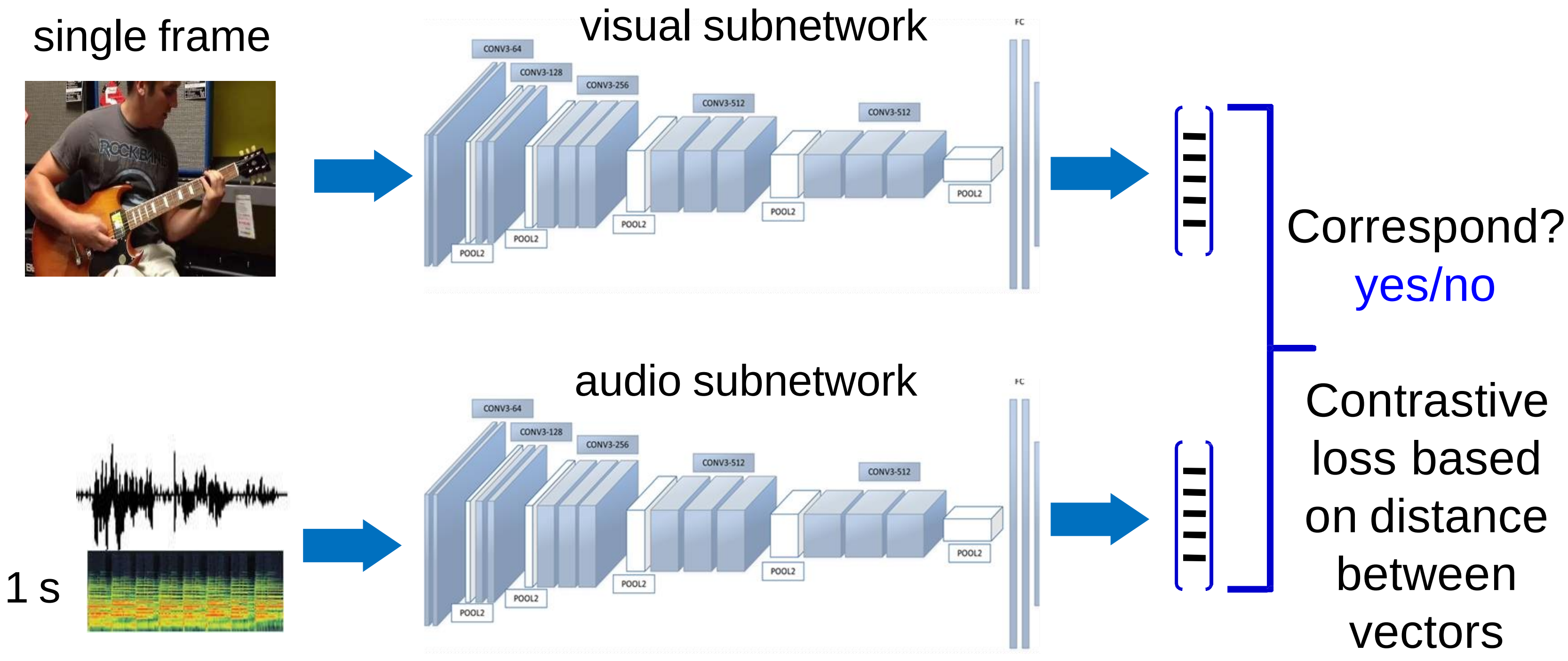
Audio-Visual Correspondence



Audio-Visual Correspondence



Audio-Visual Embedding (AVE-Net)



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

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

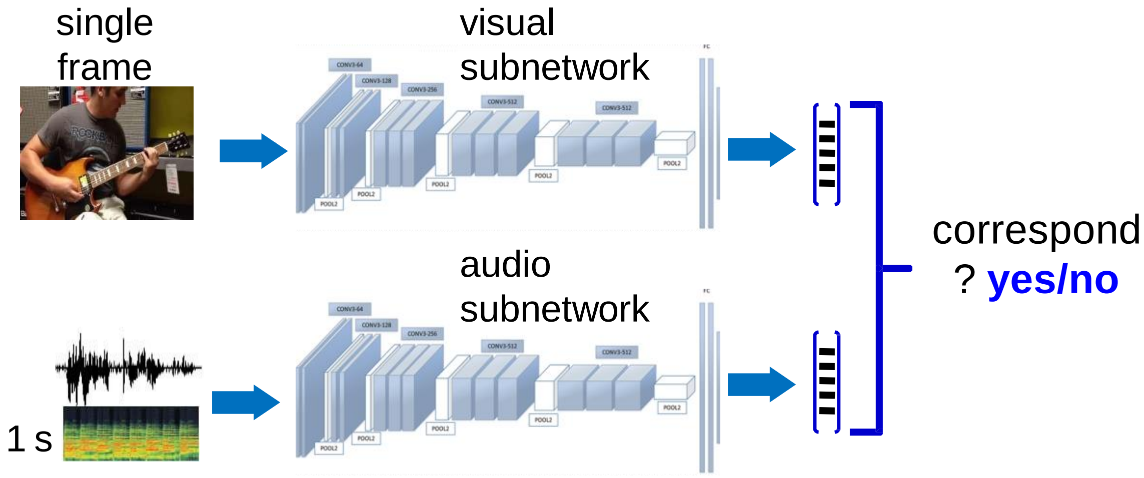
- AudioSet (from YouTube), has labels
  - 200k x 10s clips
  - use musical instruments classes
- Correspondence accuracy on test set: 82% (chance: 50%)

“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



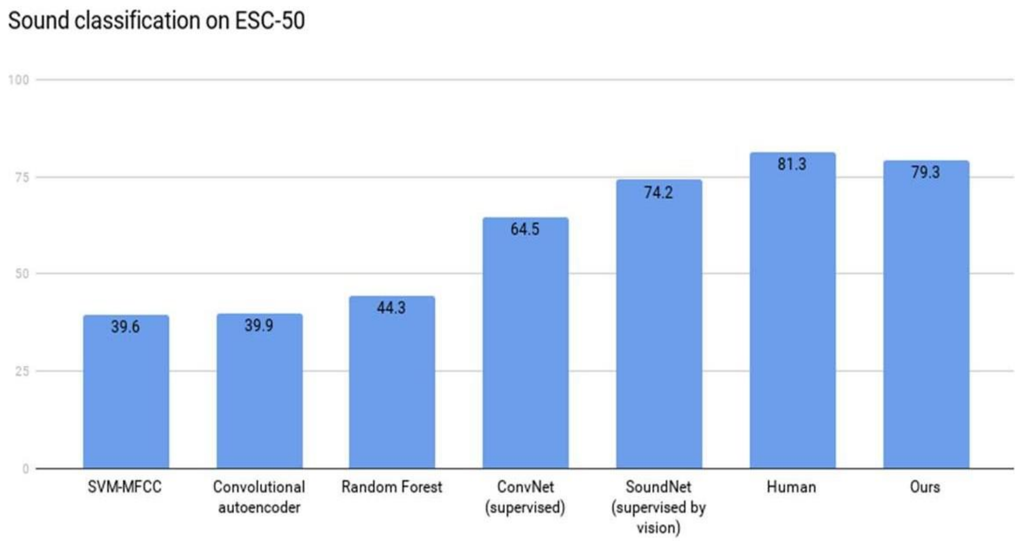
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- "What is making the sound?"
  - Learn to localize objects that sound

"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

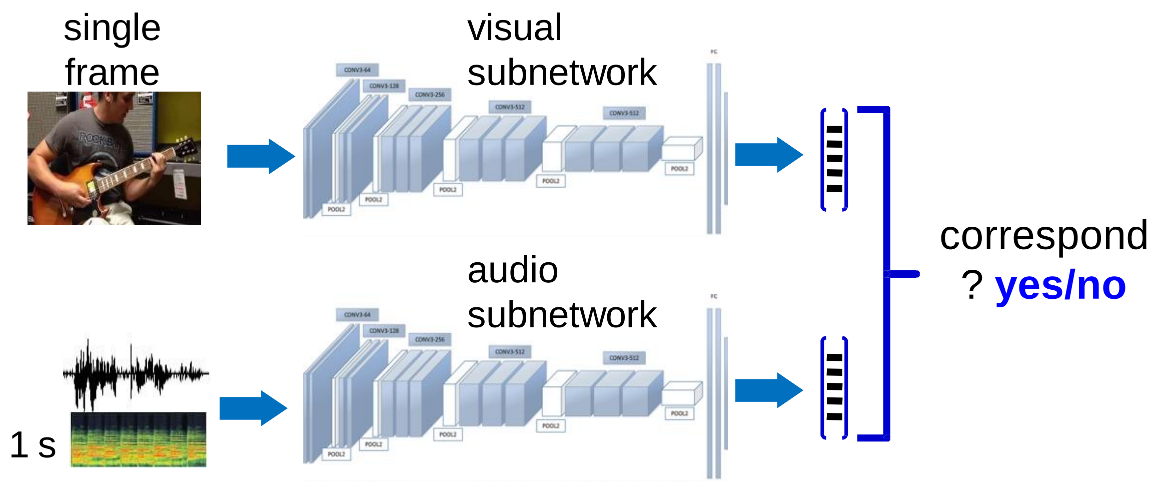
Method	Top 1 accuracy
Random	18.3%
Pathak *et al*. [21]	22.3%
Krähenbühl *et al*. [14]	24.5%
Donahue *et al*. [7]	31.0%
Doersch *et al*. [6]	31.7%
Zhang *et al*. [34] (init: [14])	32.6%
Noroozi and Favaro [18]	34.7%
Ours random	12.9%
Ours	32.3%

- 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

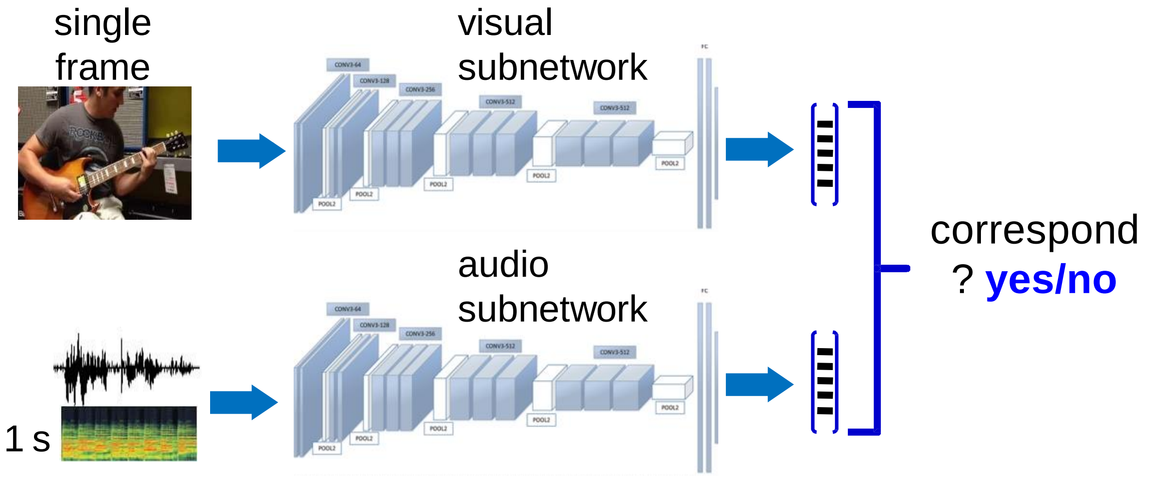
Top 10 ranked audio clips

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

Use audio and visual features

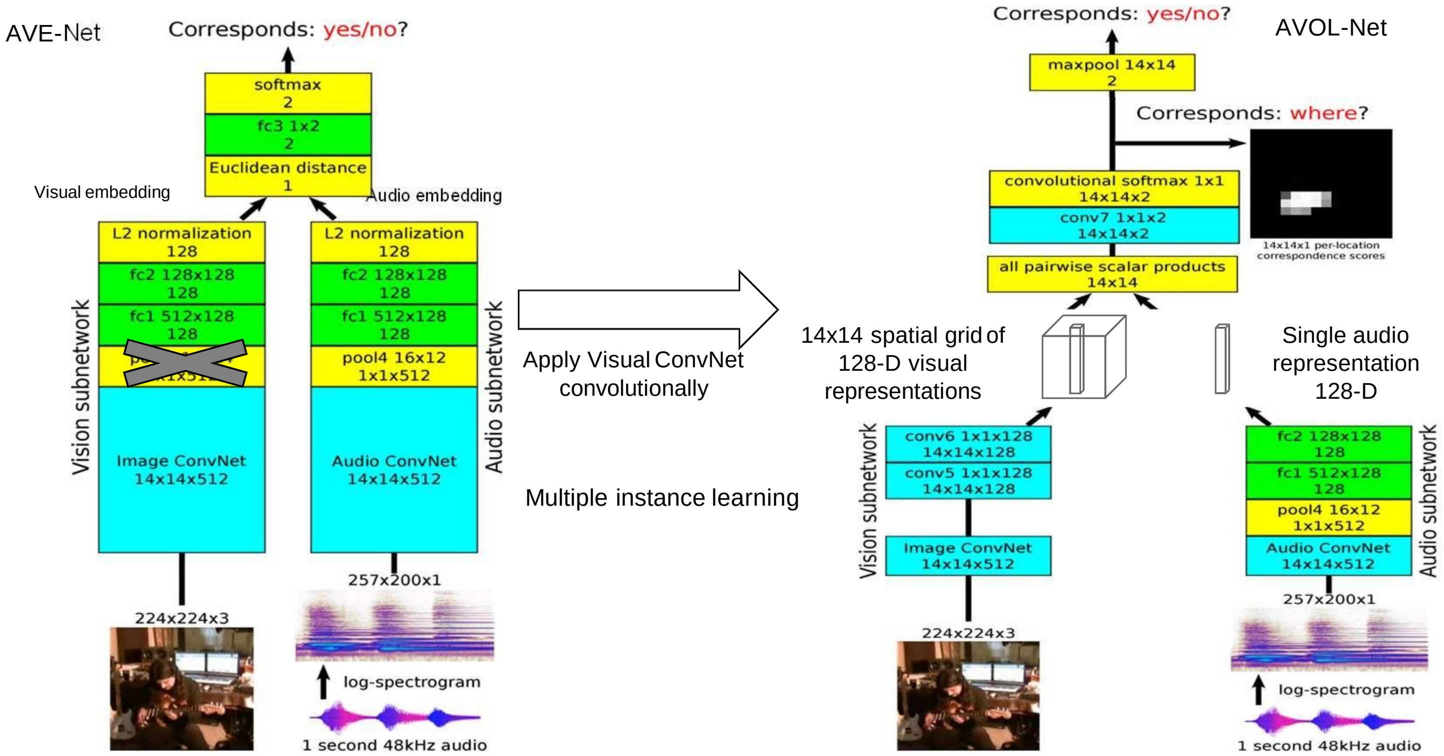
What can be learnt by watching and listening to videos?

- 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



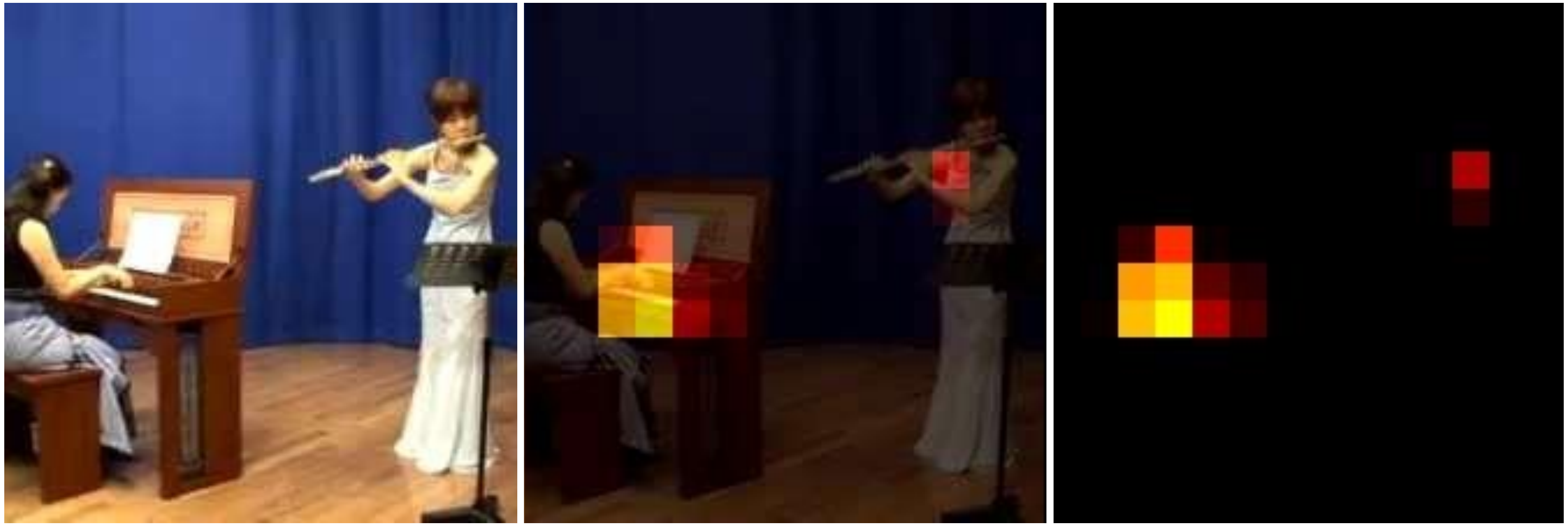
“Objects that Sound”, Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

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

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 …