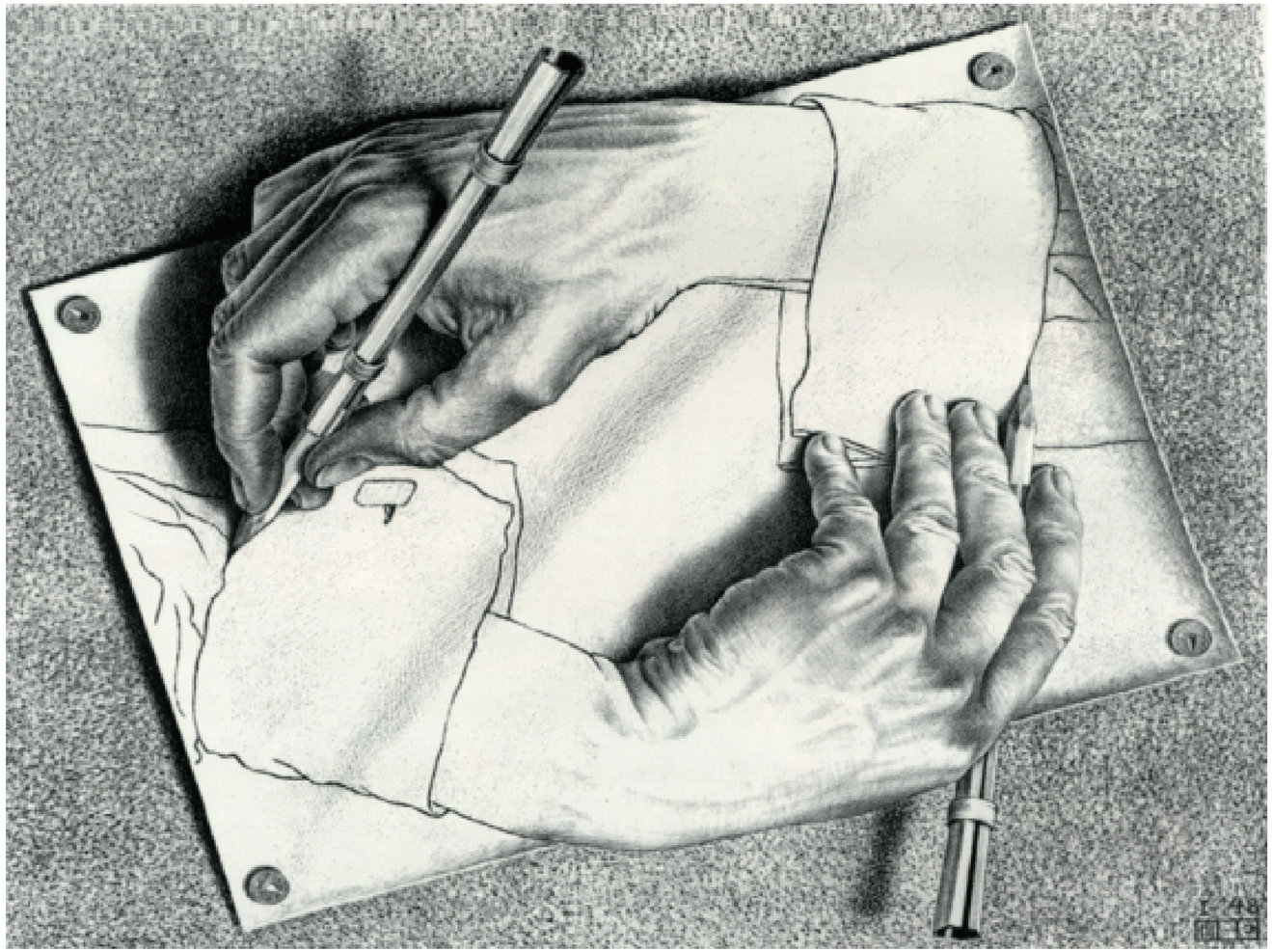
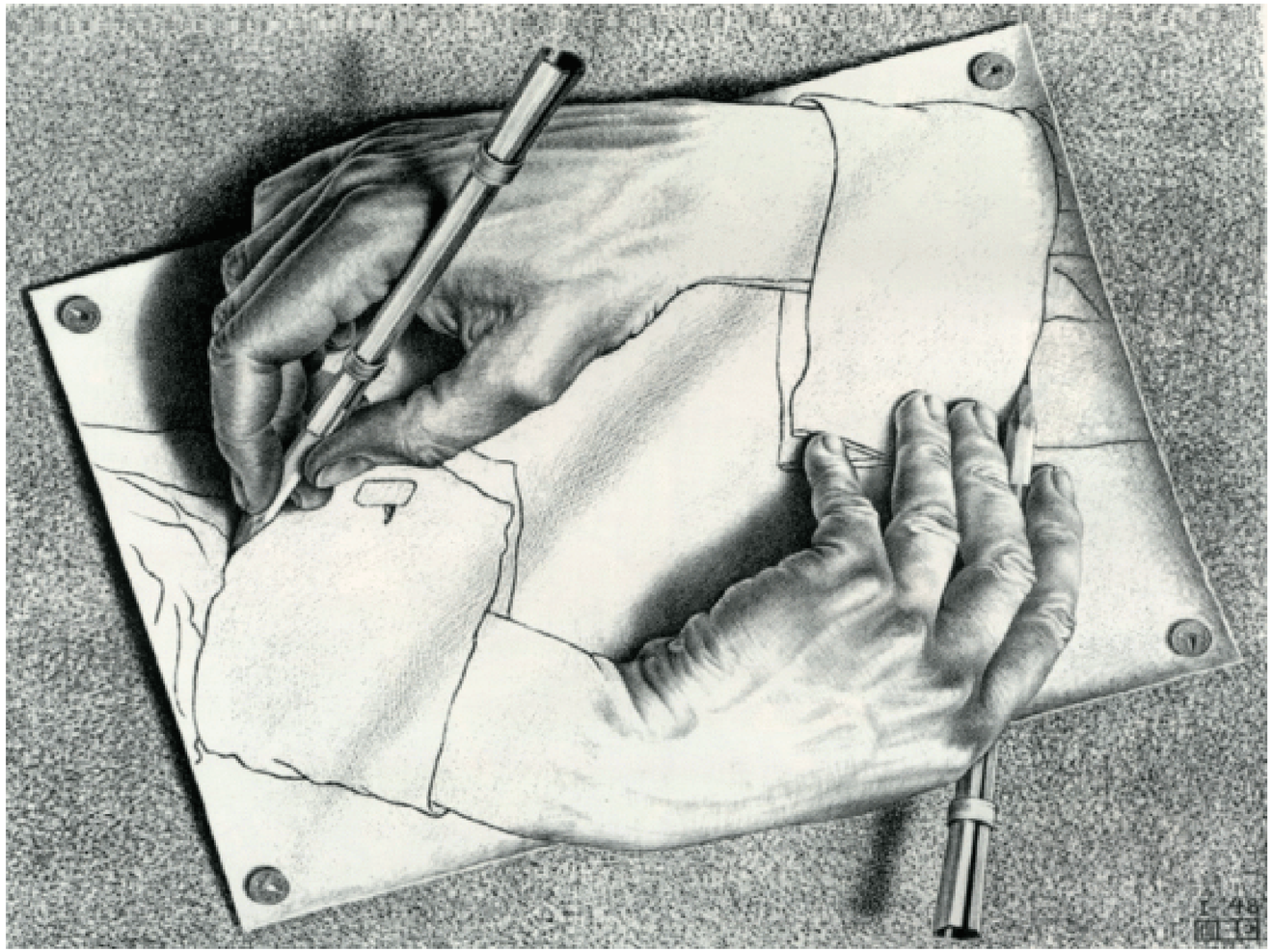
Self-Supervised Visual Learning



CS194: Image Manipulation, Comp. Vision, and Comp. Photo

Alexei Efros, UC Berkeley, Spring 2020

a.k.a. “It’s all about the Data”!



CS194: Image Manipulation, Comp. Vision, and Comp. Photo
Alexei Efros, UC Berkeley, Spring 2020

“AI” is (almost) here!

Deep Learning solves everything!

- “No need to code anymore”
- For any given problem, just:
  1. label some training data
  2. define an objective function
  3. train neural network

  4. sell your startup for millions!

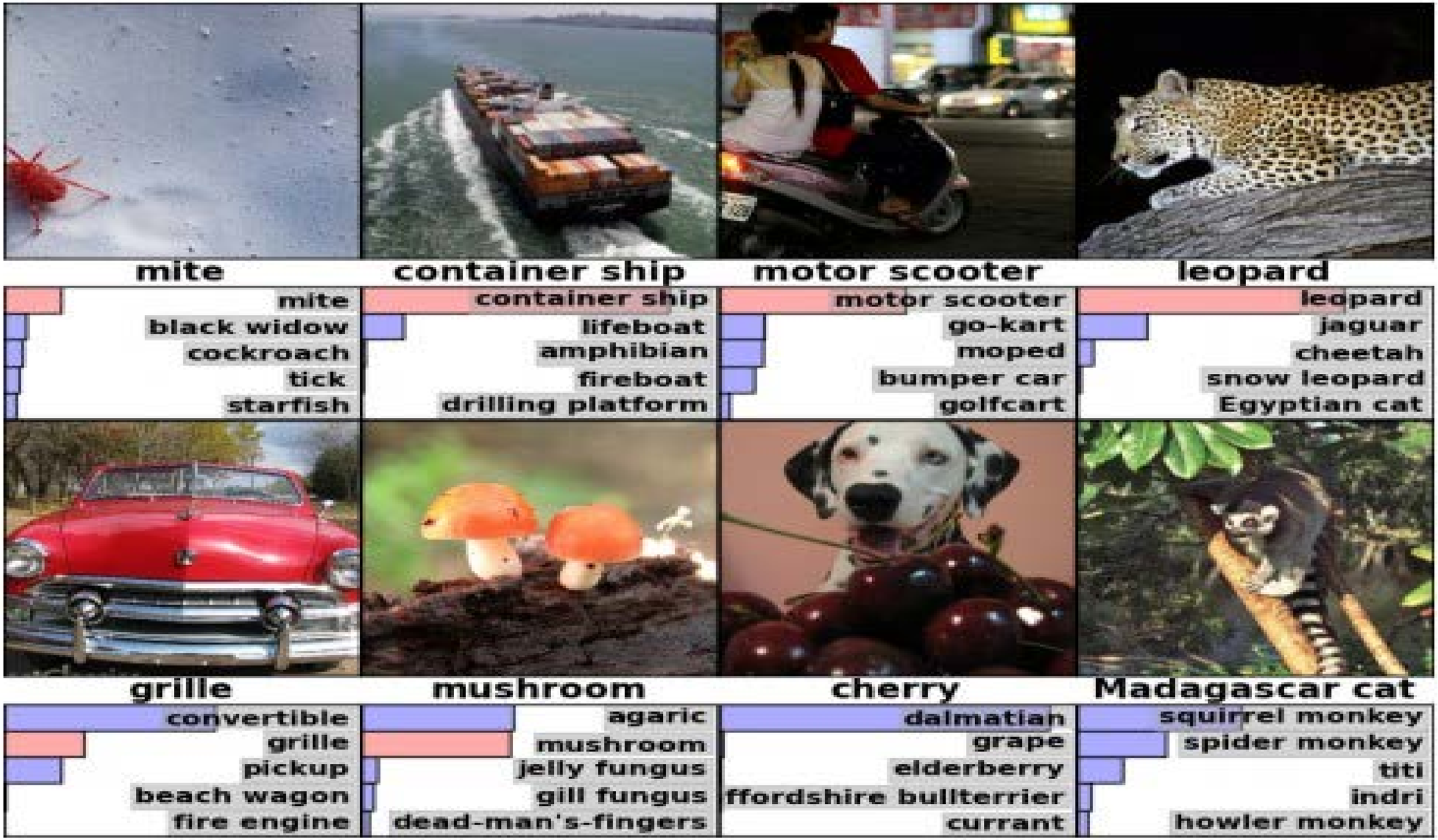


Image Classification: 80%+ correct!

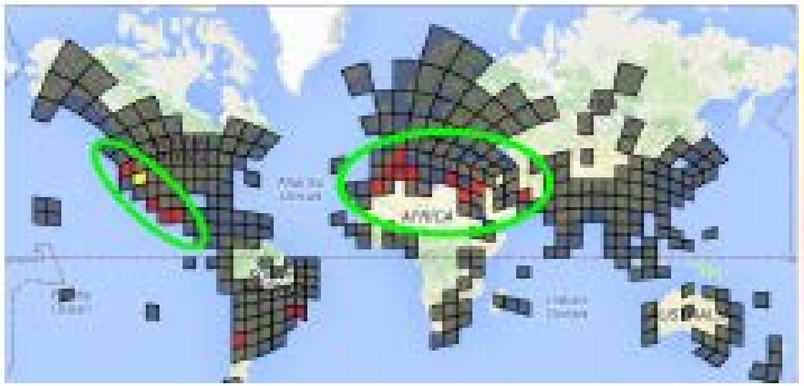
ImageNet Challenge (1000 object classes), Fei-Fei et al.

Planetary-scale geolocation

CC-BY-NC by stevekc

CC-BY-NC by edwin.11

CC-BY-NC by jonathanfh



Weyand et al, “PlaNet – Photo Geolocation with convolutional neural networks”, ECCV 2016

Automatic Image Captioning

a pot of broccoli on a stove

a group of people posing for a picture on a ski lift

Microsoft Research (Fang et al, 2015) + many other groups

BUT…

something seems too good to be true

Image Captioning

“a car parked on the side of the road”

“a car parked on the side of the road”

“a car parked on the side of the road”

Image Classification

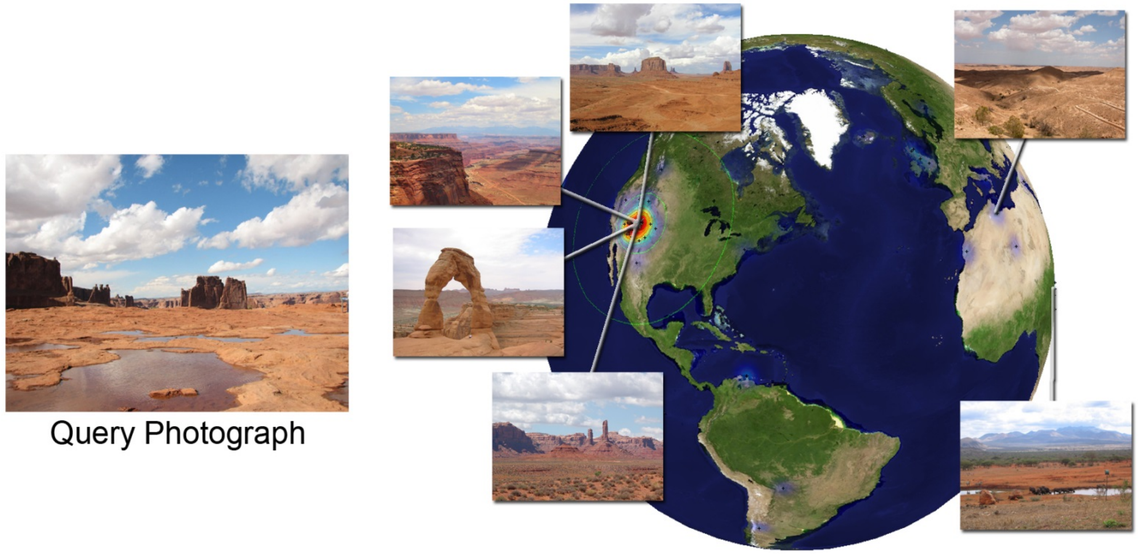
- Performance on ImageNet: ~80%
- Performance in the real world: ~30%

“T-Shirt” class in ImageNet

T-Shirts in the real world

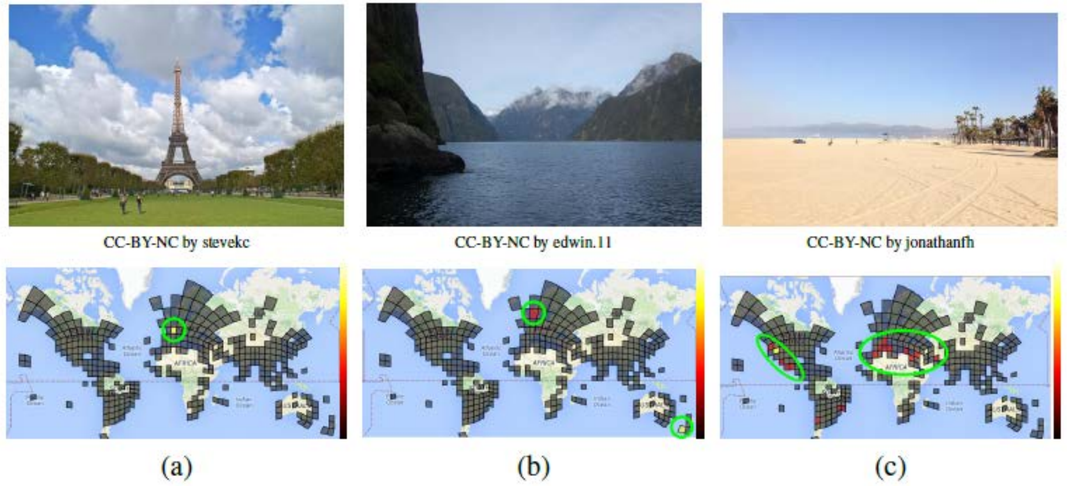
Geolocation

im2gps, 2008



- Nearest Neighbors
- 6 million images

PlaNet, 2016



- Deep Net
- 91 million images

Algorithm vs. Data

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS (orig) [19]		12.0%	15.0%	23.0%	47.0%
Im2GPS (new) [20]	2.5%	21.9%	32.1%	35.4%	51.9%
PlaNet (900k)	0.4%	3.8%	7.6%	21.6%	43.5%
PlaNet (6.2M)	6.3%	18.1%	30.0%	45.6%	65.8%
PlaNet (91M)	**8.4%**	**24.5%**	**37.6%**	**53.6%**	**71.3%**

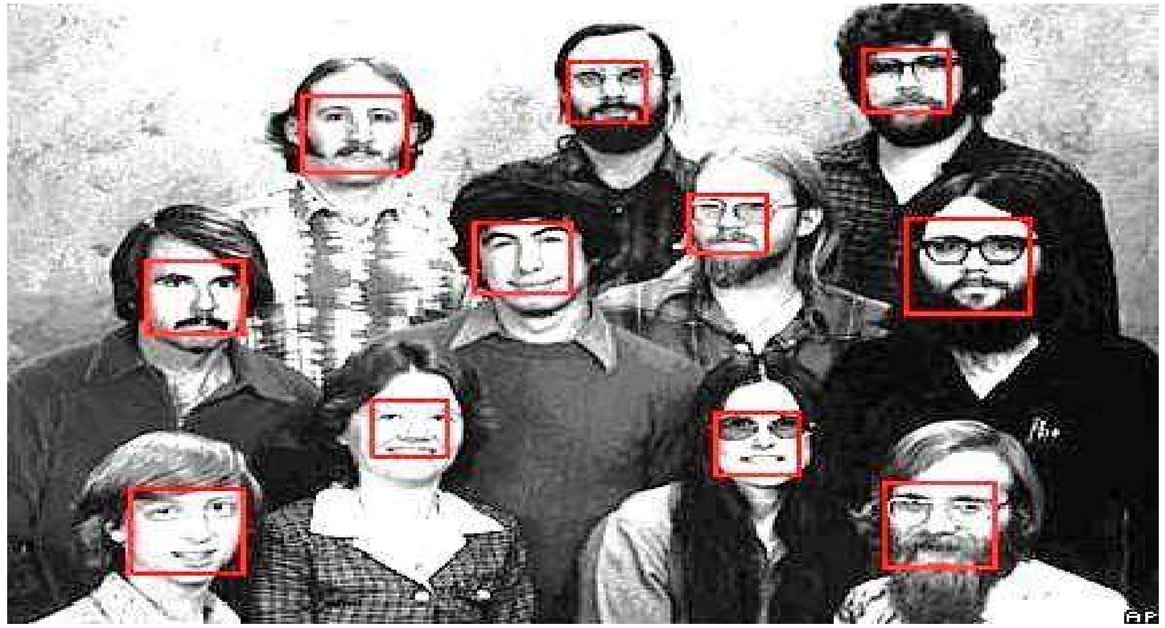
Data gets little respect…

Data

Features

Learning Algorithm

Face Detection: Early Success Story



- Rowley, Baluja, and Kanade, 1998
  - features: **pixels**, classifier**: neural network**
- Schniderman & Kanade, 1999
  - features: **pairs of wavelet coeff**., classifier**: naïve Bayes**
- Viola & Jones, 2001
  - features: **haar**, classifier: **boosted cascade**

Our Scientific Narcissism

All things being equal, we prefer to credit our own cleverness

“Unreasonable Effectiveness of Data”

[Halevy, Norvig, Pereira 2009]

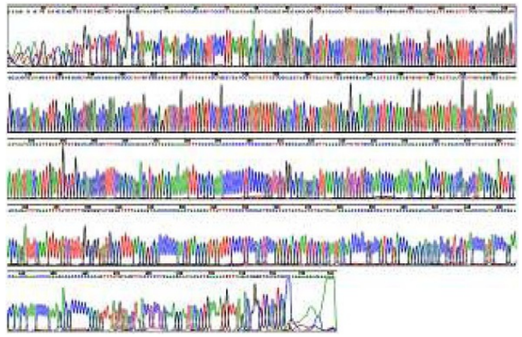
- Parts of our world can be explained by elegant mathematics:
  - physics, chemistry, astronomy, etc.
- But some cannot:
  - psychology, biology, economics, AI, etc.

- That’s exactly where we need **Magic of Data**

Some things are just messy

Navier-Stokes Equation

$$\frac{\partial \mathbf{u}}{\partial t} = -\left(\mathbf{u} \cdot \nabla\right)\mathbf{u} + v\nabla^2\mathbf{u} - \frac{1}{d}\nabla p + \mathbf{f}$$



+ weather
+ location
+ …

Brain-dead lookup (aka Nearest Neighbor) often works surprisingly well

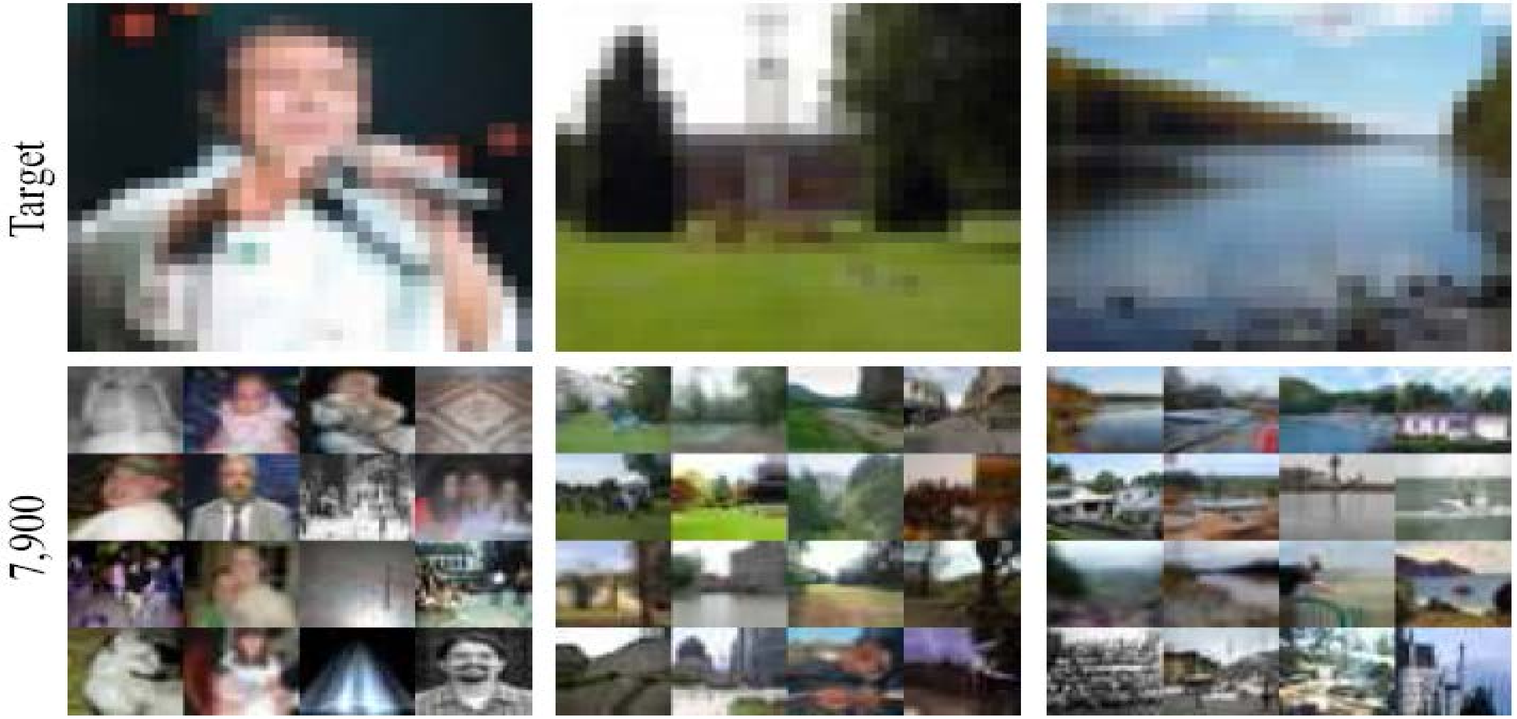
Lots of Tiny Images

- “80 million tiny images: a large dataset for non-parametric object and scene recognition”, Antonio Torralba, Rob Fergus and William T. Freeman. PAMI 2008.

Lots

Of

Images

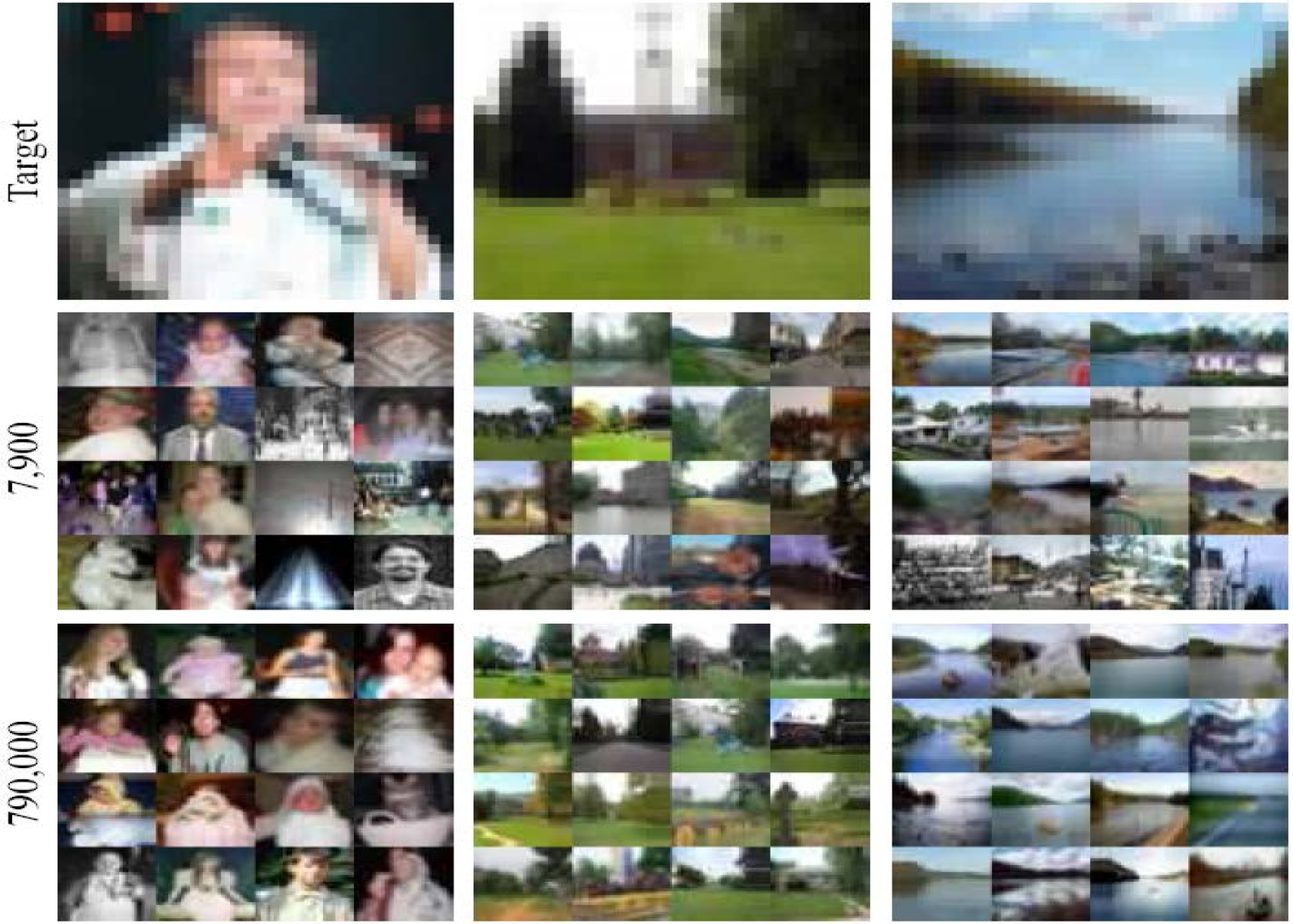


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images

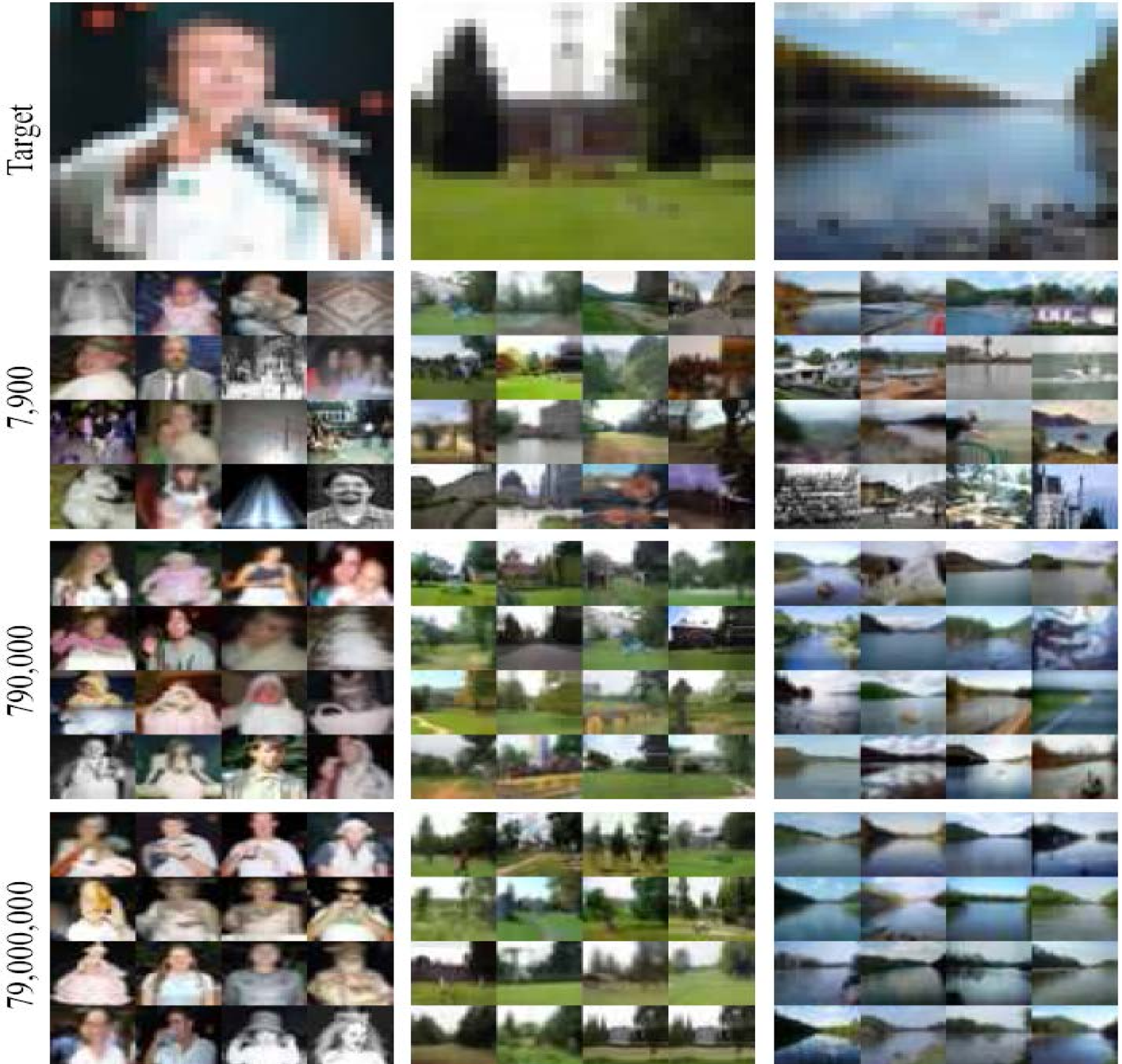


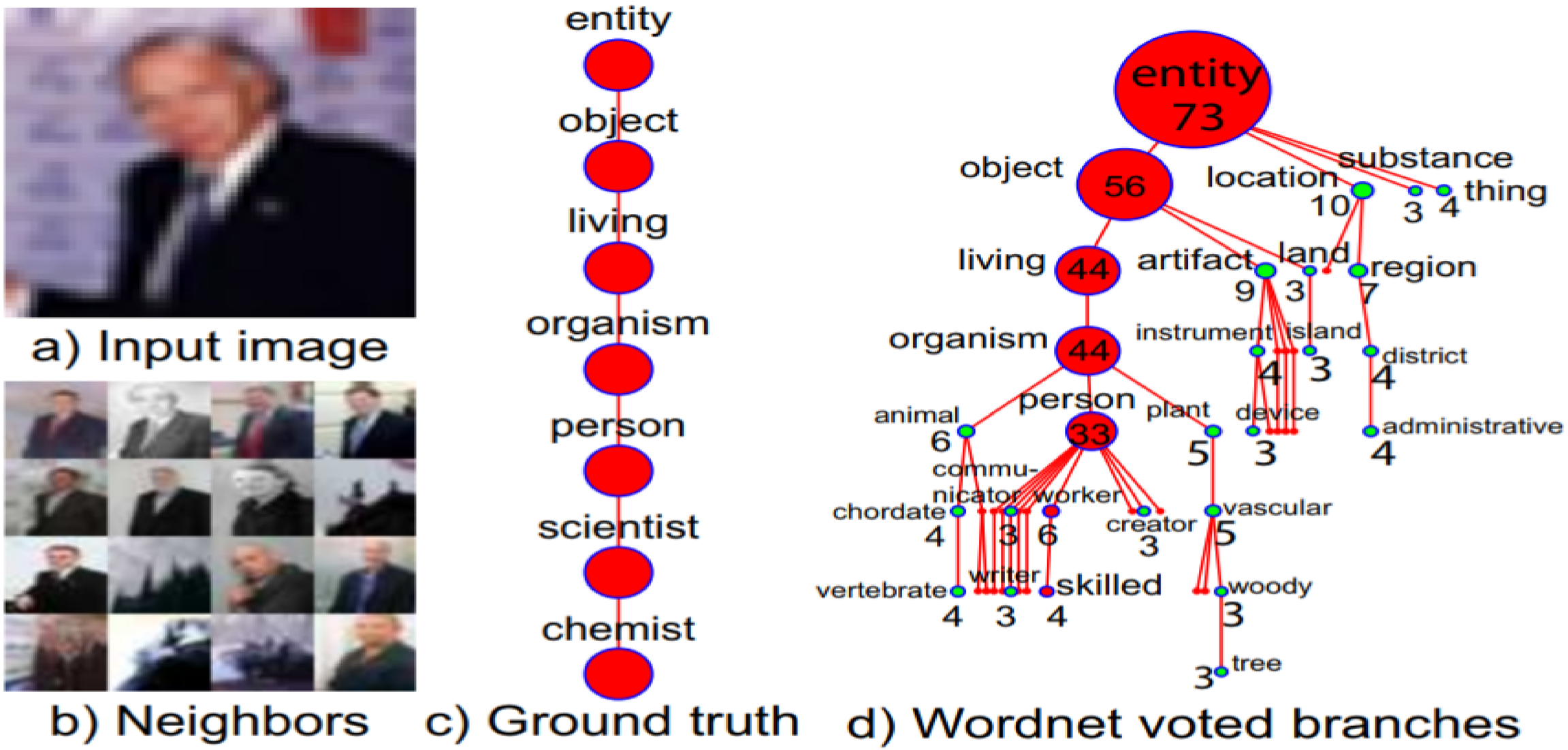
A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images

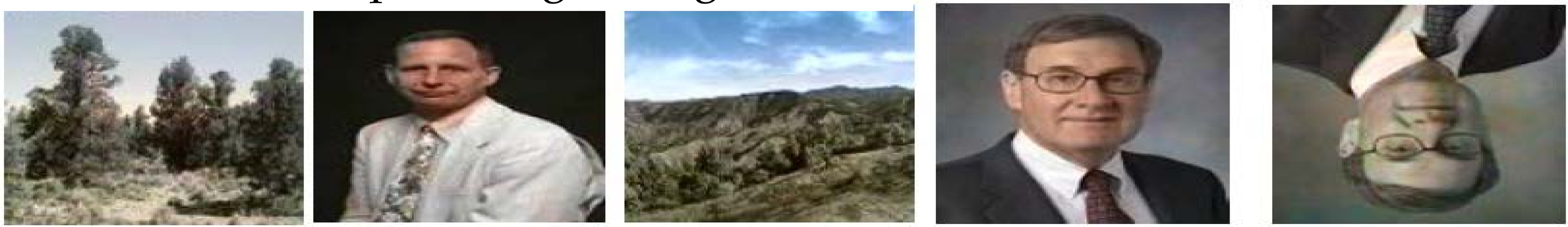



a) Input image
b) Neighbors
c) Ground truth
entity
object
living
organism
person
scientist
chemist
d) Wordnet voted branches
entity 73
object 56
substance
location 10
3
4
thing
living 44
artifact 9
land 3
region 7
organism 44
instrument 4
island 3
district 4
person 33
animal 6
plant 5
device 3
administrative 4
commu-nicator
worker 6
chordate 4
3
creator 3
vascular 5
vertebrate 4
writer 3
skilled 4
woody 3
tree 3

Automatic Colorization

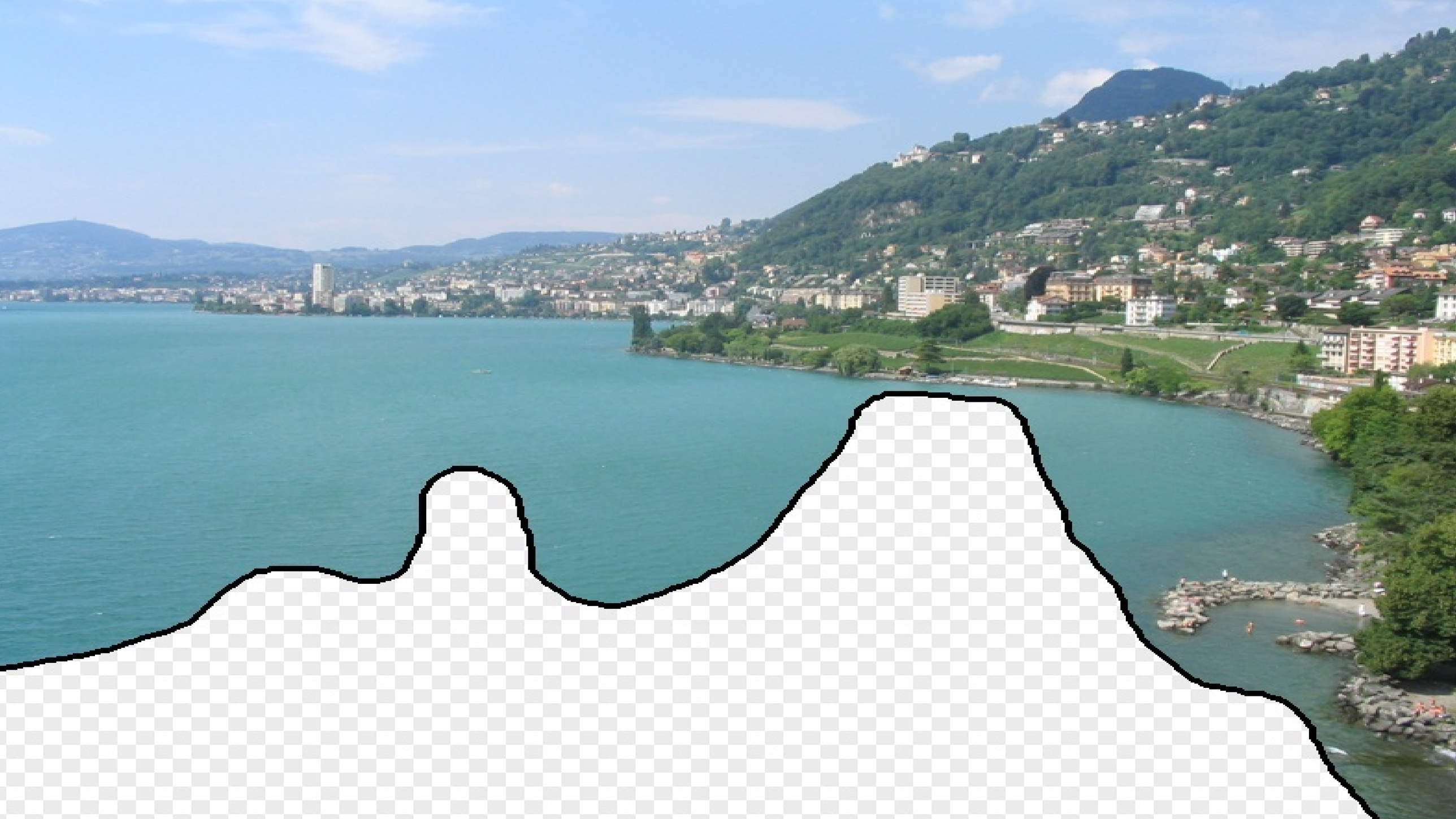
Grayscale input High resolution

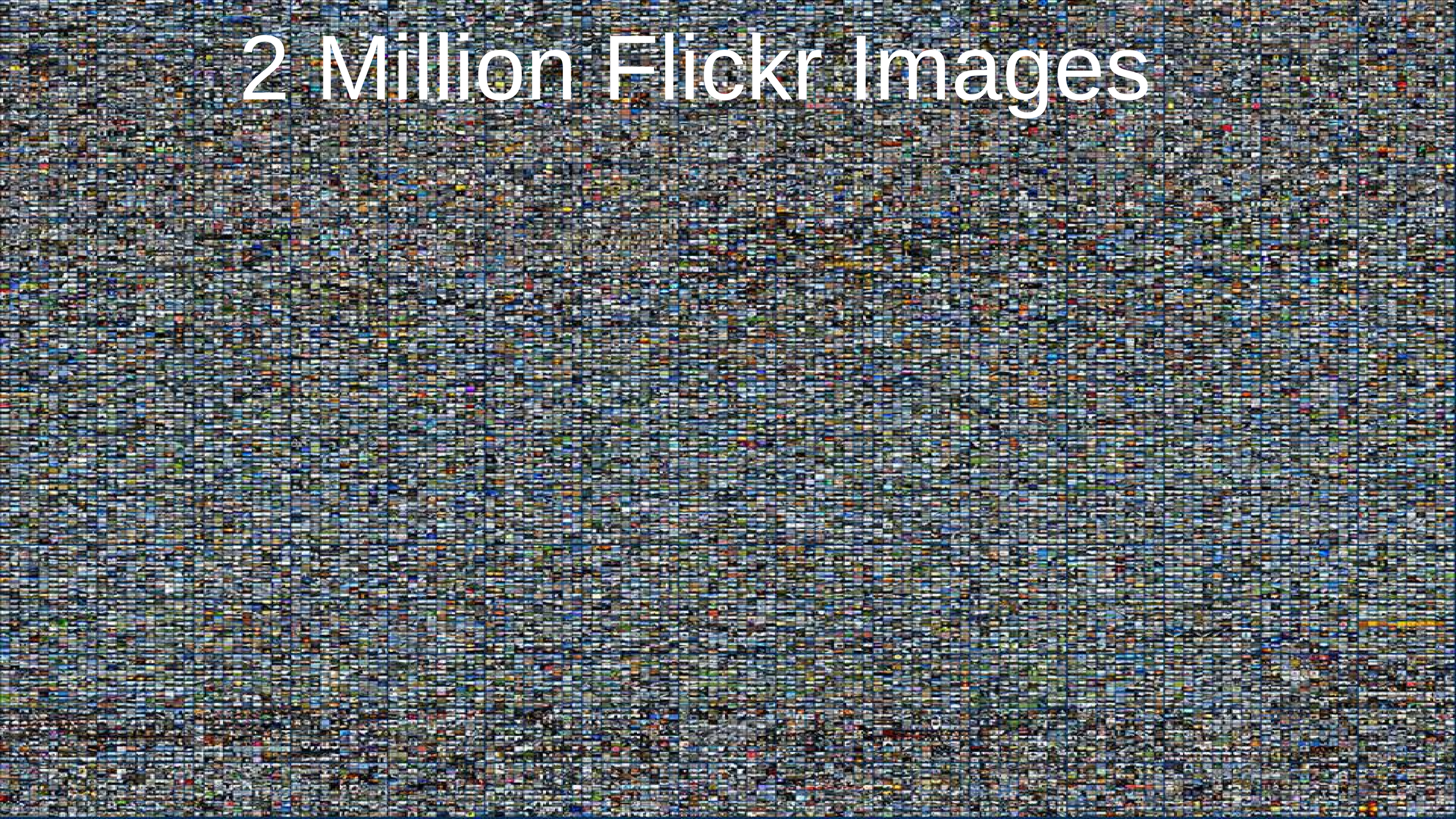
Colorization of input using average

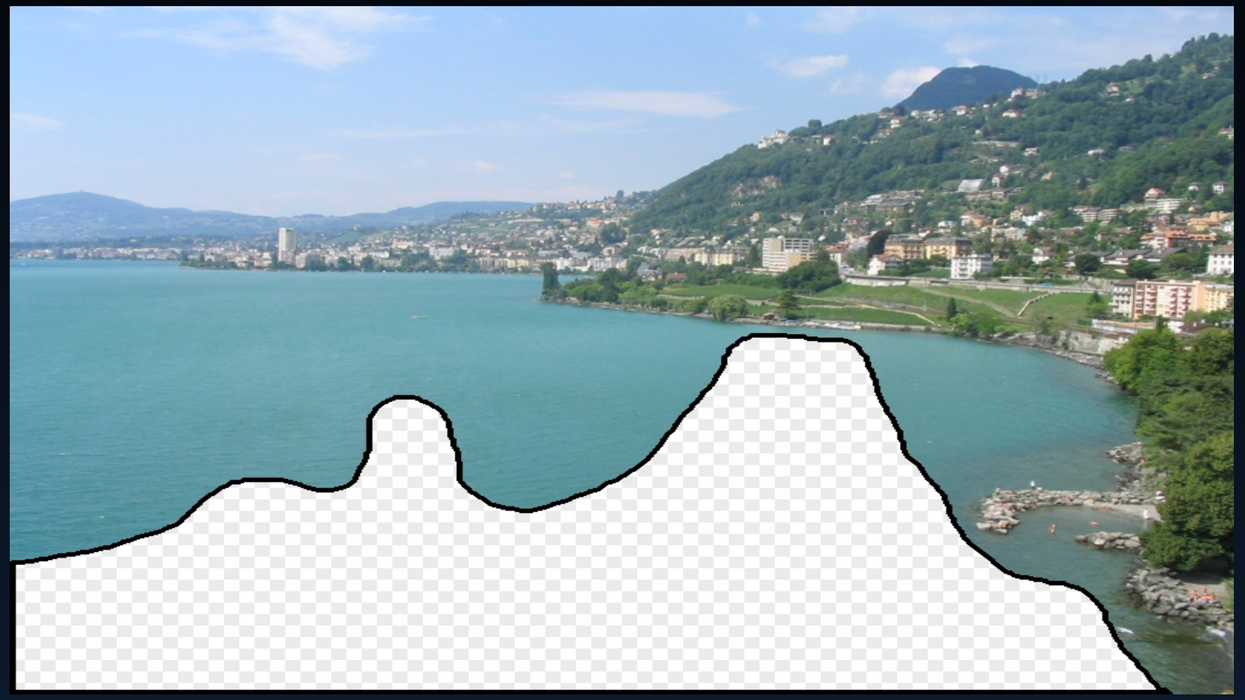


A. Torralba, R. Fergus, W.T.Freeman. 2008

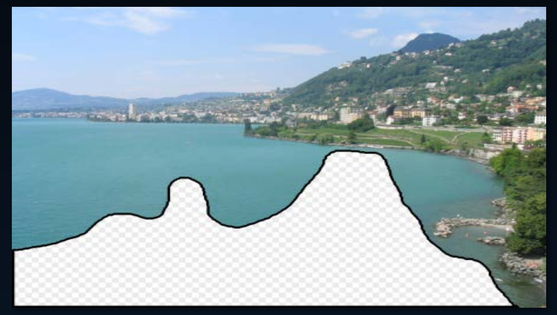
[Hays & Efros, SIGGRAPH’07]

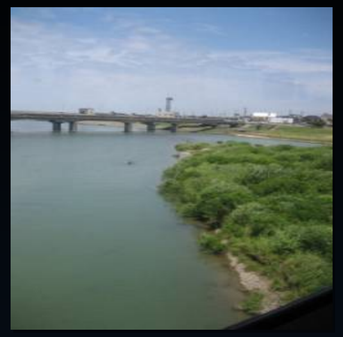



2 Million Flickr Images

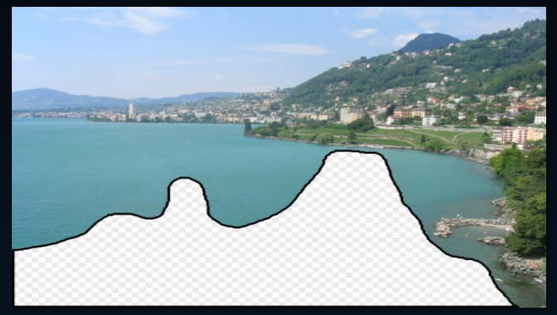


Why does it work?



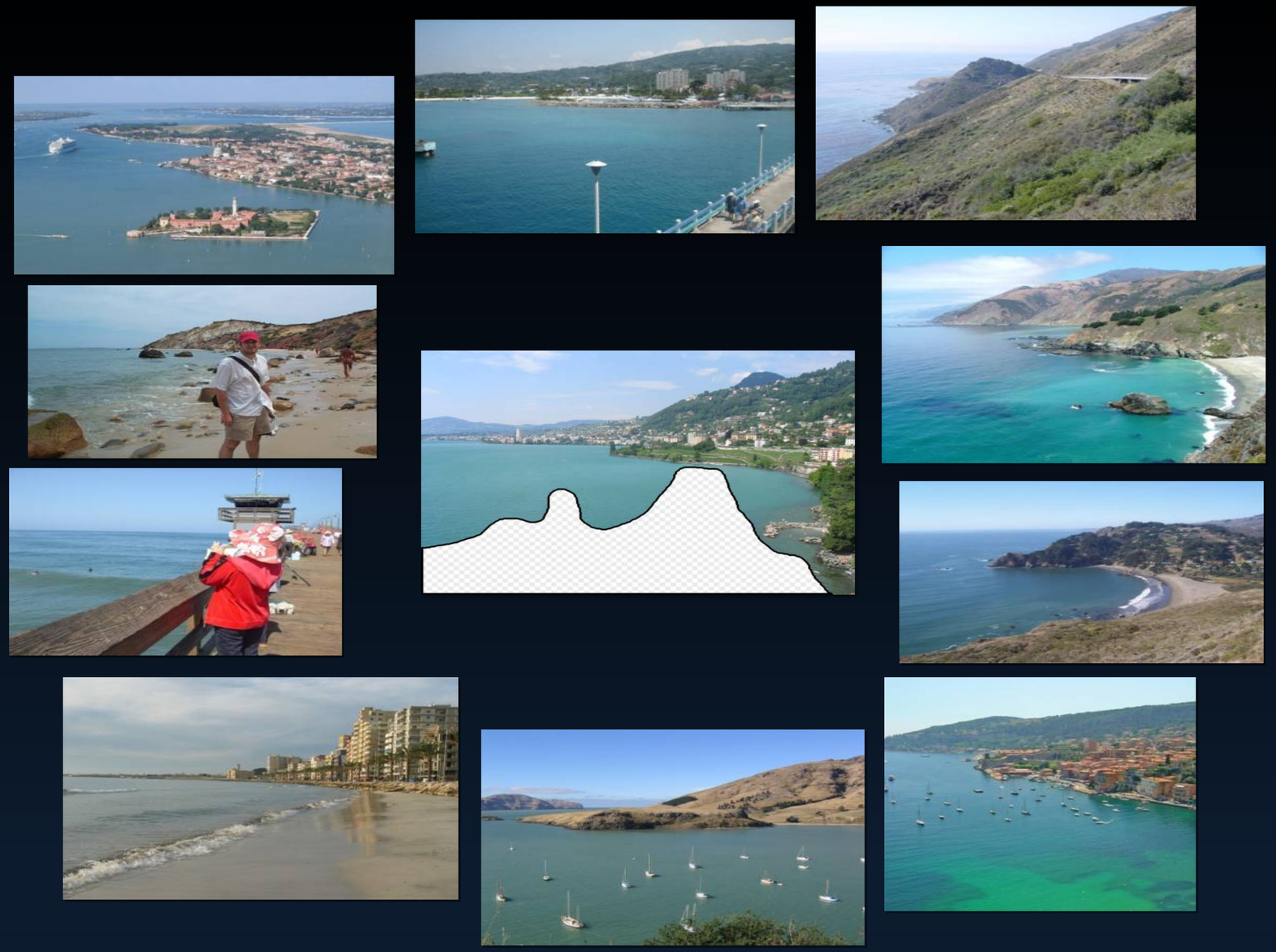







Nearest neighbors from a
collection of **20 thousand images**

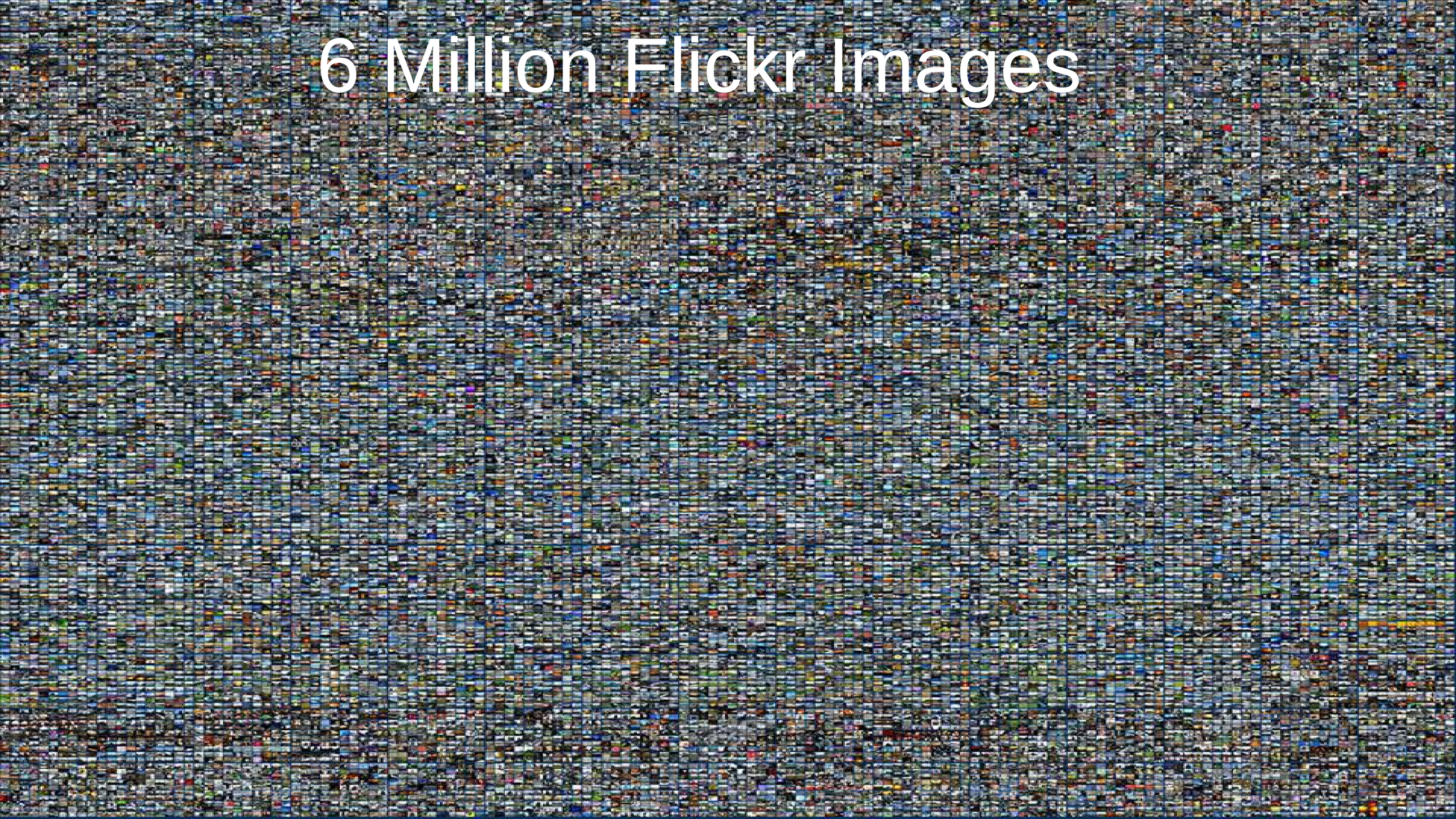


Nearest neighbors from a
collection of **2 million images**

im2GPS
(using 6 million GPS-tagged Flickr images)

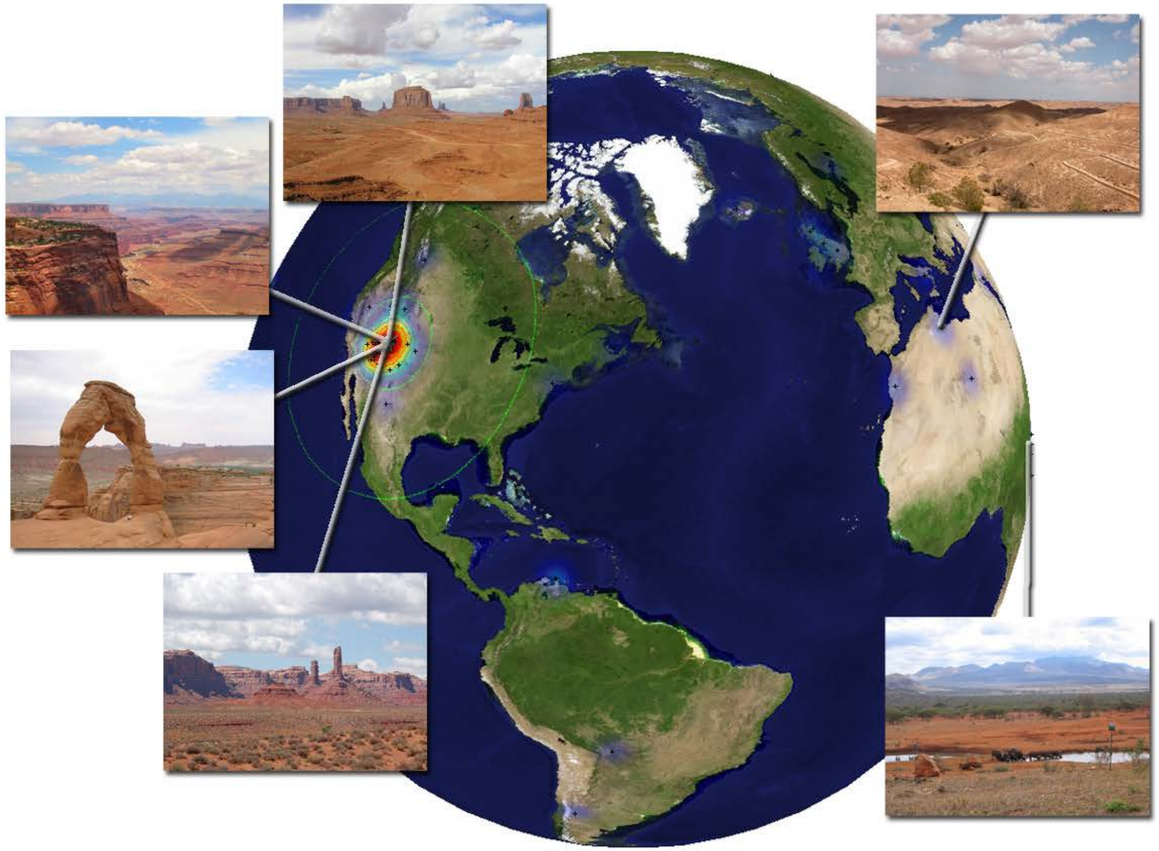
Query Photograph

Im2gps [Hays & Efros, CVPR’08]


6 Million Flickr Images

im2GPS
(using 6 million GPS-tagged Flickr images)

Query Photograph



Visually Similar Scenes

Im2gps [Hays & Efros, CVPR’08]

The Good News

Really stupid algorithms + Lots of Data

= “Unreasonable Effectiveness”

[Halevy, Norvig, Pereira 2009]

But surely the brain can’t remember this much!?

What's the Capacity of Visual Long Term Memory?

What we know…

Standing (1973)

10,000 images

83% Recognition

… people can remember thousands of images

What we don't know…

… what people are remembering for each item?

According to Standing

"Basically, my recollection is that we just separated the pictures into distinct thematic categories: e.g. cars, animals, single-person, 2-people, plants, etc.) Only a few slides were selected which fell into each category, and they were visually distinct."

Dogs
Playing Cards

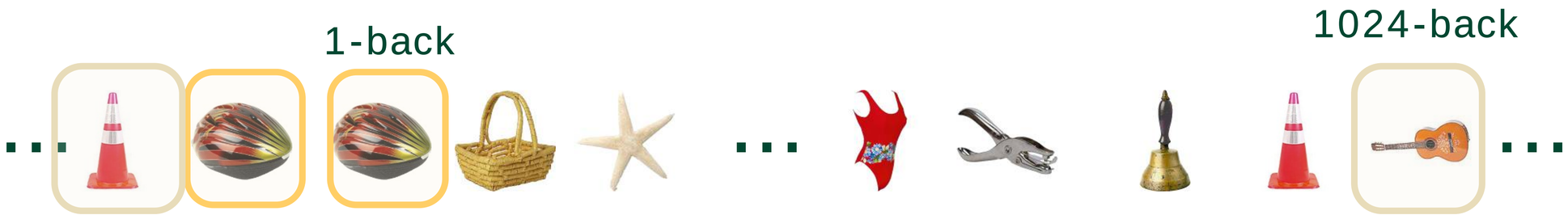
"Gist" Only

Sparse Details

Highly Detailed

Slide by Aude Oliva

Massive Memory I: Methods



Showed 14 observers 2500 **categorically unique objects**

1 at a time, 3 seconds each

800 ms blank between items

Study session lasted about 5.5 hours

Repeat Detection task to maintain focus

Followed by 300 2-alternative forced choice tests

Slide by Aude Oliva

Massive Memory Experiment I

A stream of objects will be presented on the screen for
~ 3 second each.

Your primary task:

Remember them ALL!

afterwards you will be tested with…

Completely different objects...

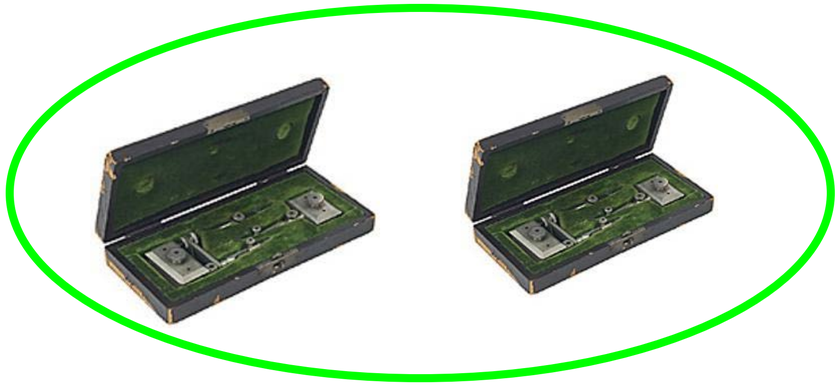
Different exemplars of the same kind of object...

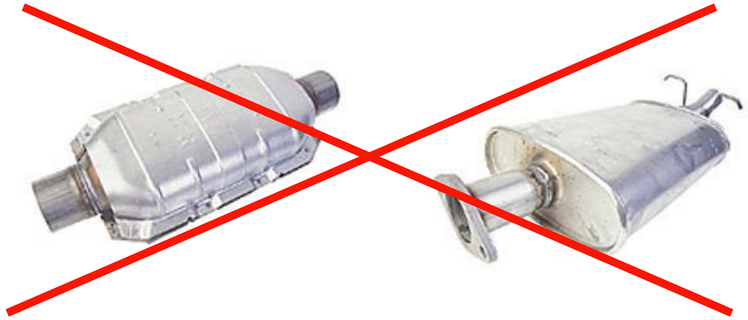
Different states of the same object...

Massive Memory Experiment I

Your other task:

Detect exact repeats
anywhere in the stream





Ready?

(Seriously, get ready to clap. The images go by fast…)

<clap!>

<clap!>

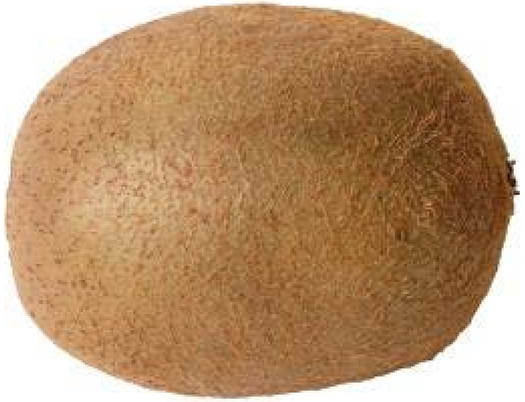
10 Minutes Later...

<clap!>

<clap!>

30 Minutes Later...

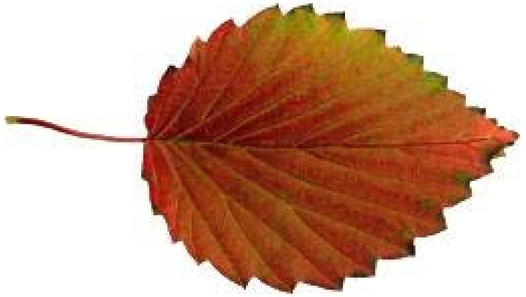
1 Hour Later...




POLICE

<clap!>

2 Hours Later...



<clap!>

4 Hours Later...


PURE
MAPLE SYRUP
540 mL


FM
SW
STANDARD
PROGRAM

<clap!>

5:30 Hours Later...

Which one did you see?

(go ahead and shout out your answer)

-A-

-B-

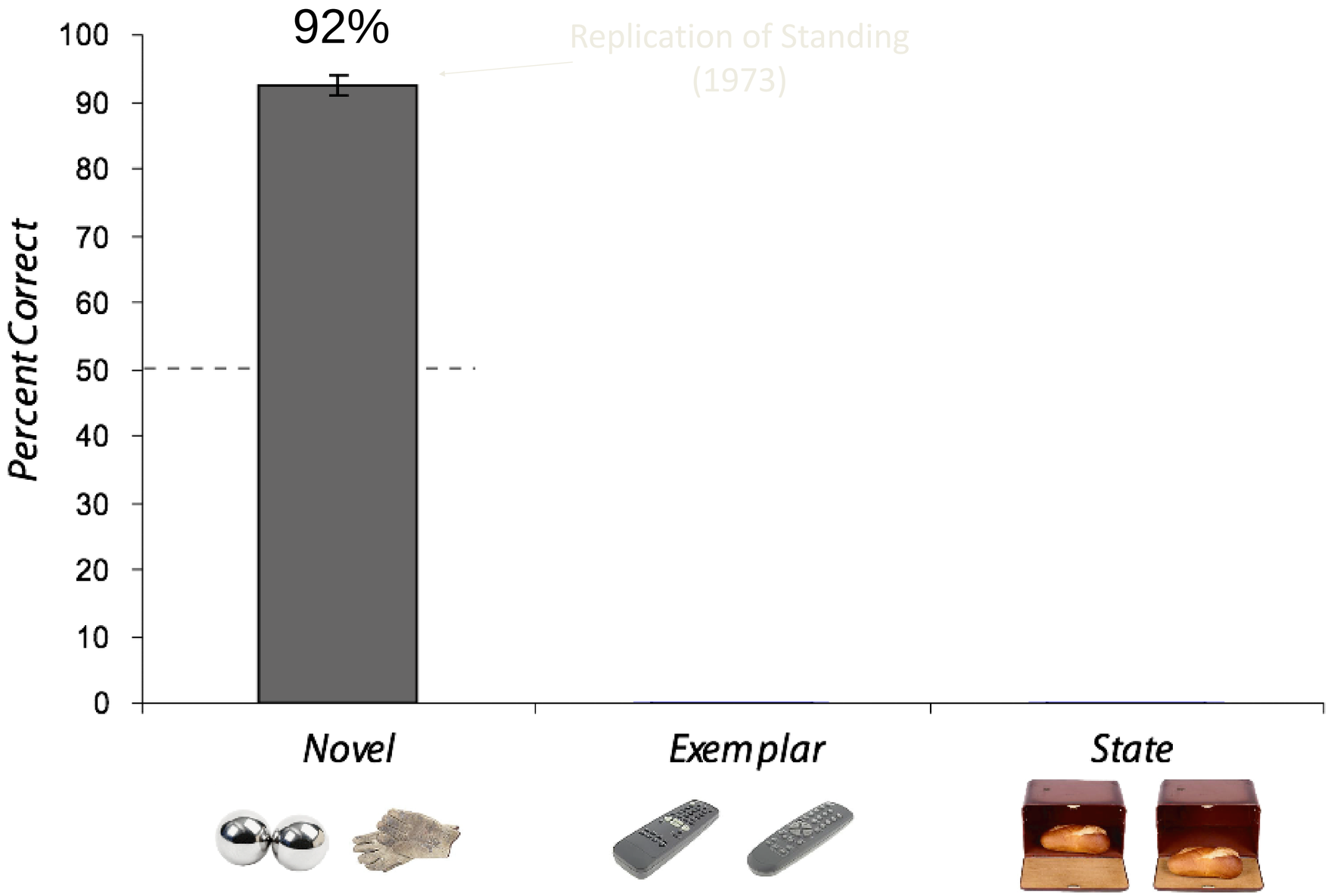
-A-

-B-

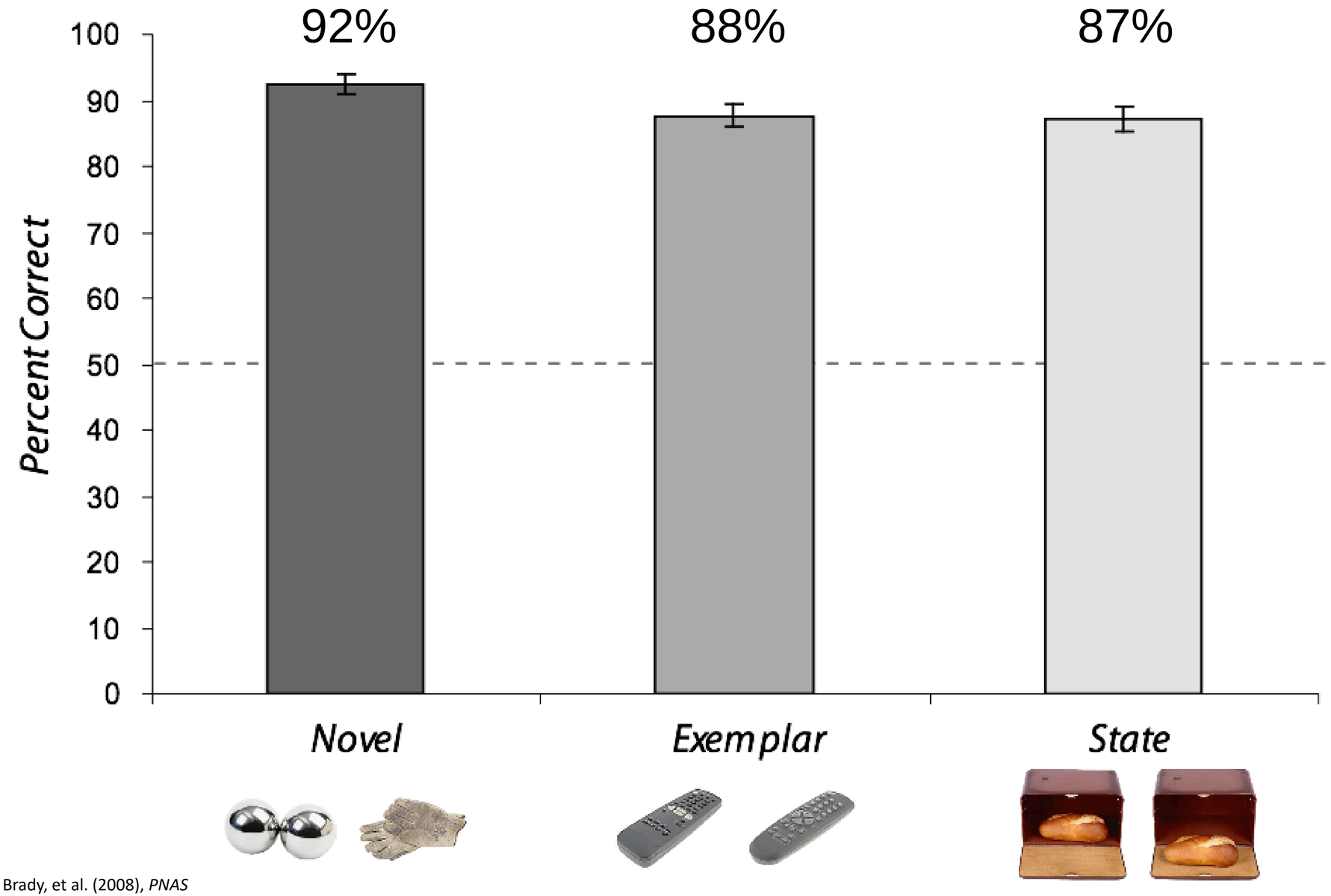
-A-

-B-

Recognition Memory Results



Recognition Memory Results



Brady, et al. (2008), *PNAS*

So, do humans have
“photographic memory”?

Clap your hands when
you see an image repeat

Ready?

<clap!>

<clap!>


BET NOW AT BLUESQUARE.COM
WEMBLEY STADIUM

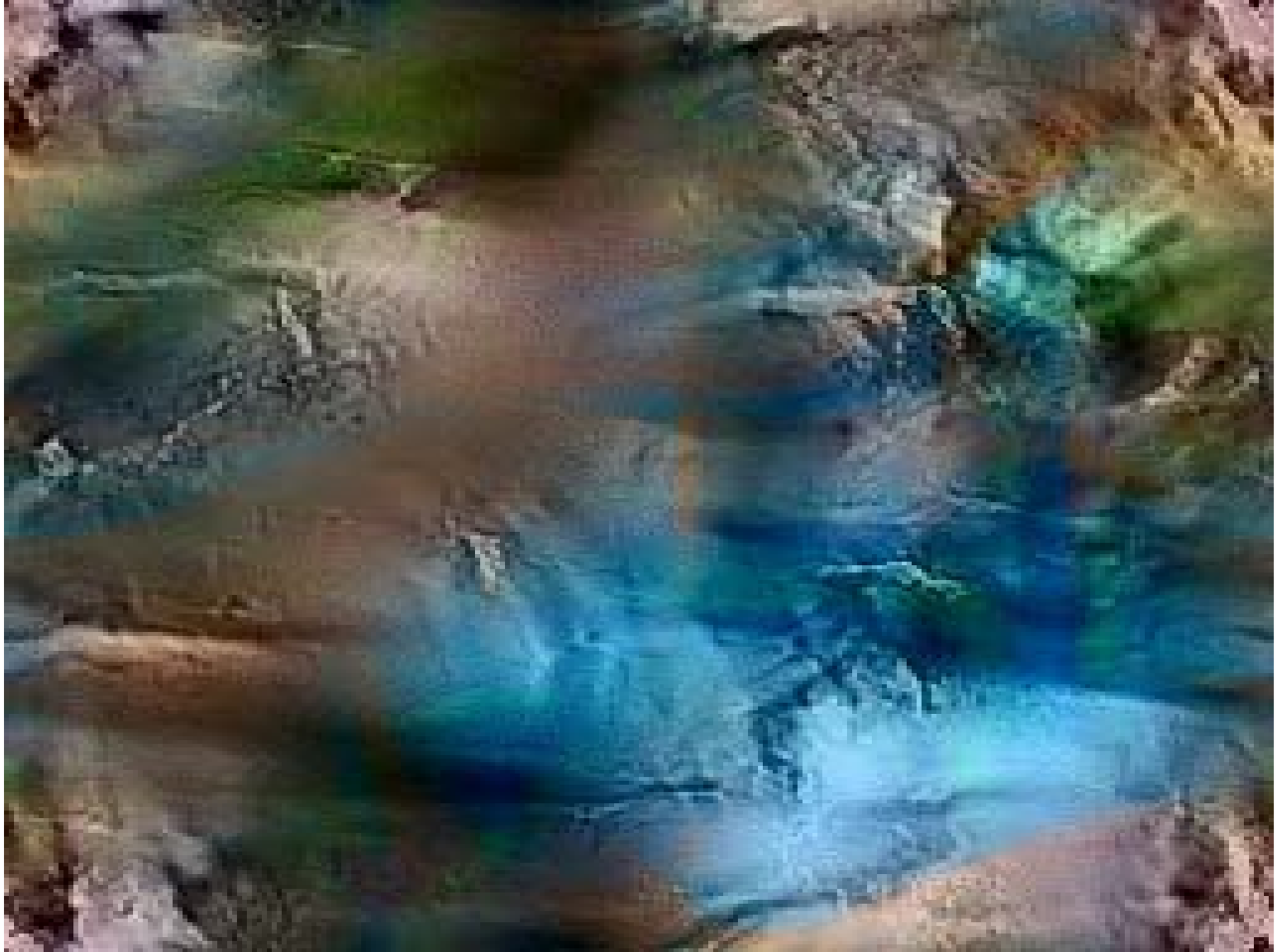
<clap!>


FURLA
SPORTSTAFF


FURLA
SPORTSTAFF

<clap!>

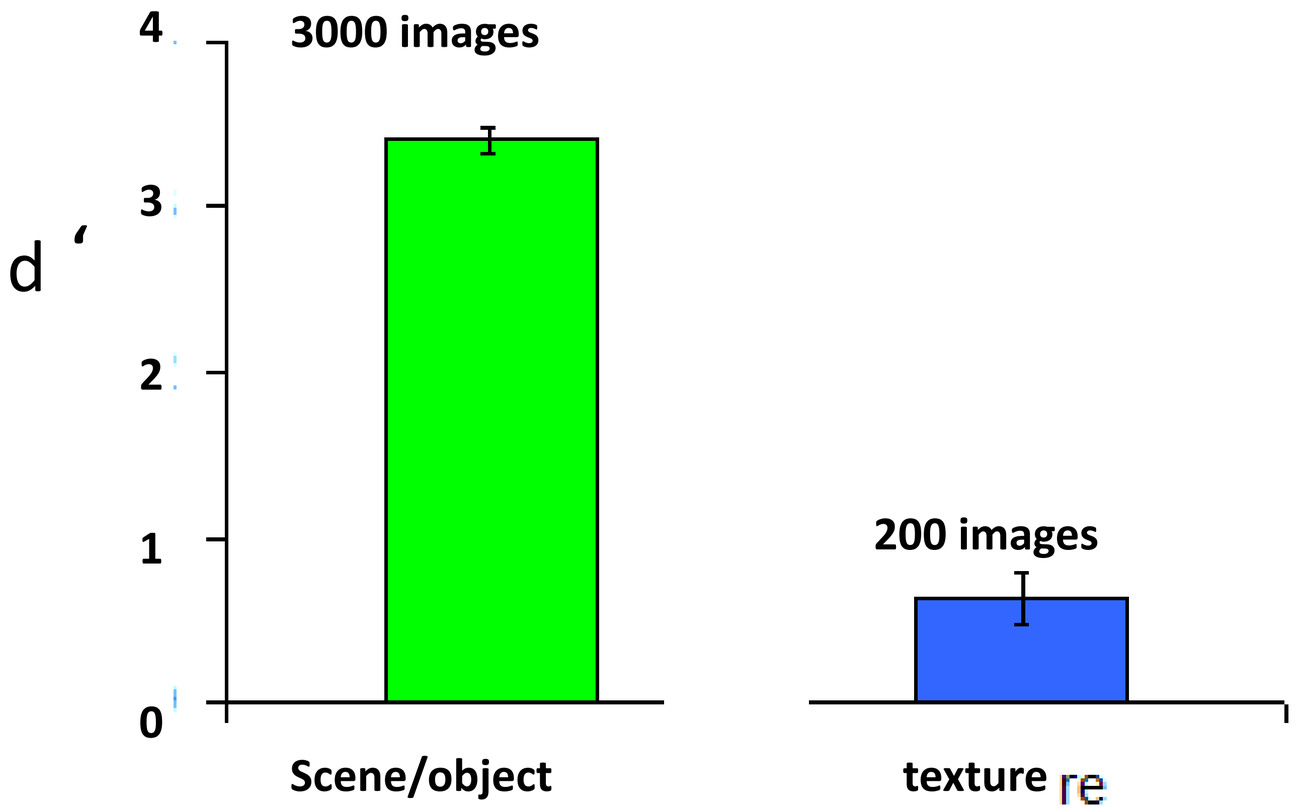
<clap!>



<clap!>


BET NOW AT BLUESQUARE.COM
WEMBLEY STADIUM

<clap!>



Oliva, unpublished

This is where ConvNets come in

- Huge capacity
  - effectively unlimited?
- Learned feature space
  - focuses on what’s “important”

With huge capacity, huge temptations…

- ConvNets can memorize anything
  - e.g. “Rethinking generalization” [Zhang et al, 2017]

- ConvNets love to cheat…

- Semantic supervision might be to blame

classifiers love to cheat

image X

Convolutional Neural Network

“Shetland Sheepdog”

label Y

classifiers love to cheat

image *X*

Convolutional Neural Network

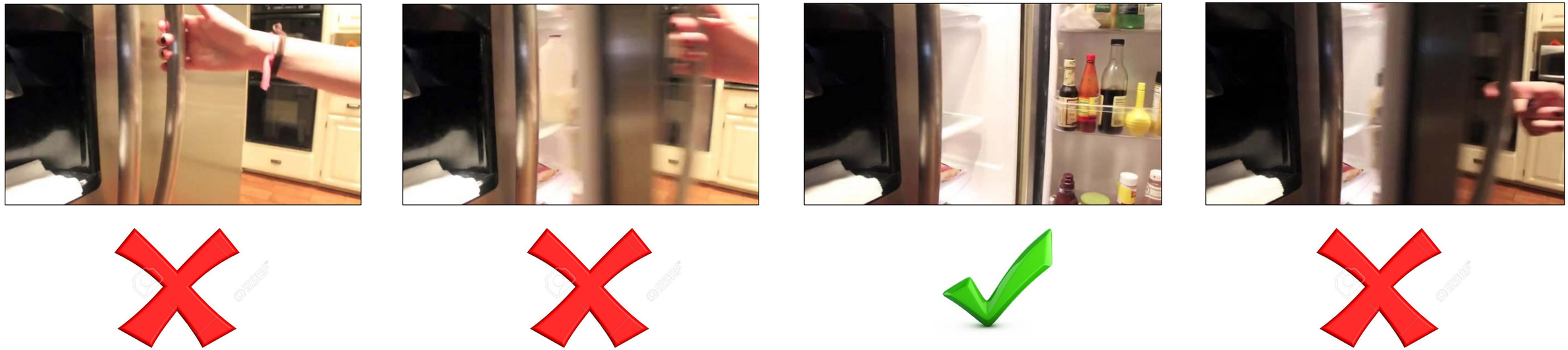
“Shetland Sheepdog”

label *Y*

Gatys et al, 2017

example: action recognition in video

– input video:



– output: class label

- Picking up cup
- Slicing bread
- **Opening fridge**
- etc

example by David Fouhey

semantic supervision == memorization

We are raising a generation of algorithms who can only “cram for the test” (set)

Why do we have vision?

- “To see what is where by looking”
  - Aristotle, Marr, etc
- .
- .
- .
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

Why do we have vision?

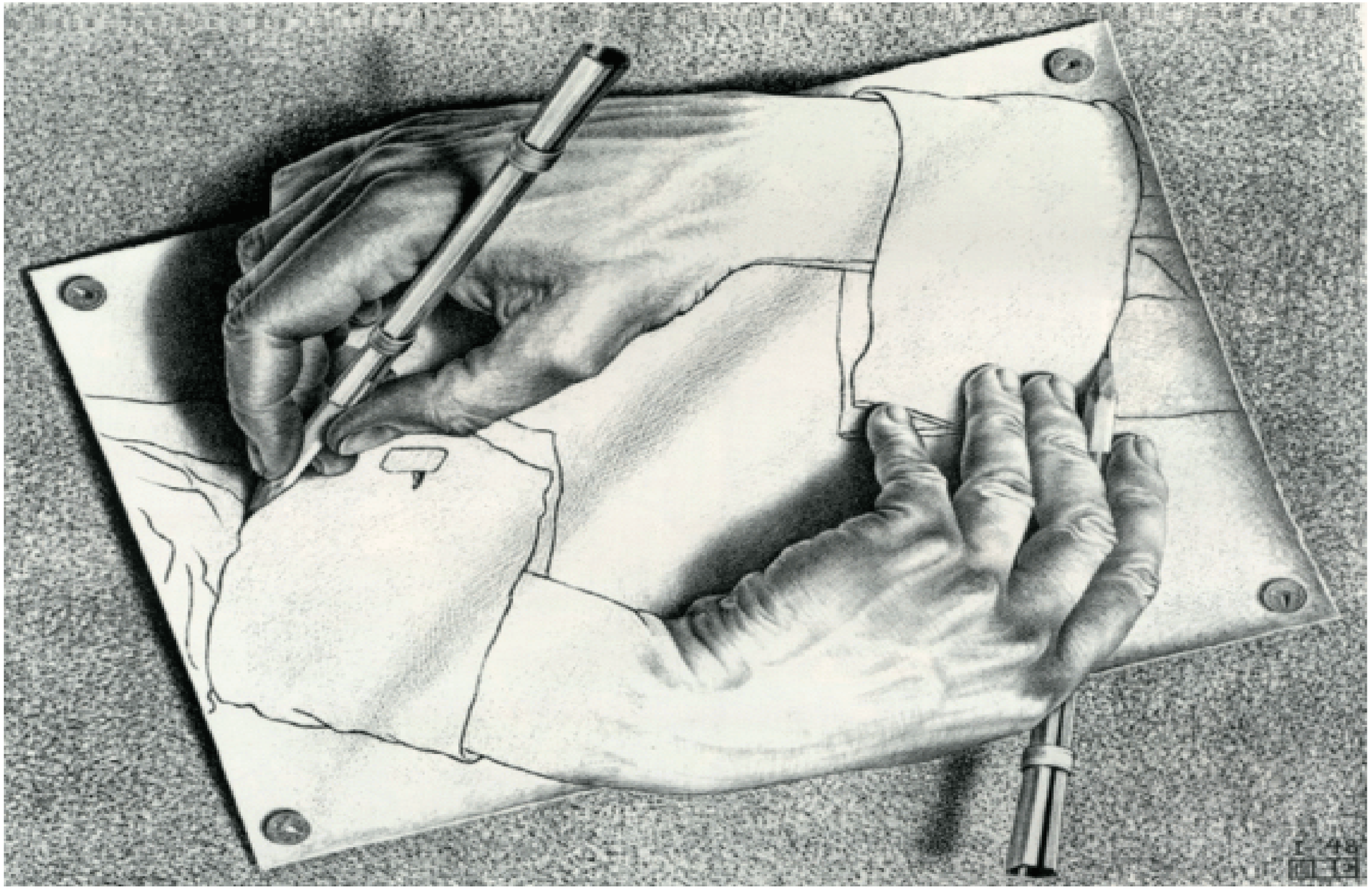
- “To see what is where by looking”
  - Aristotle, Marr, etc.
- .
- “To predict the world”
  - Moshe Bar, Jan Koenderink, etc.
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

The world as supervision

Try to predict some aspect of the world that we interact with / have effect on:

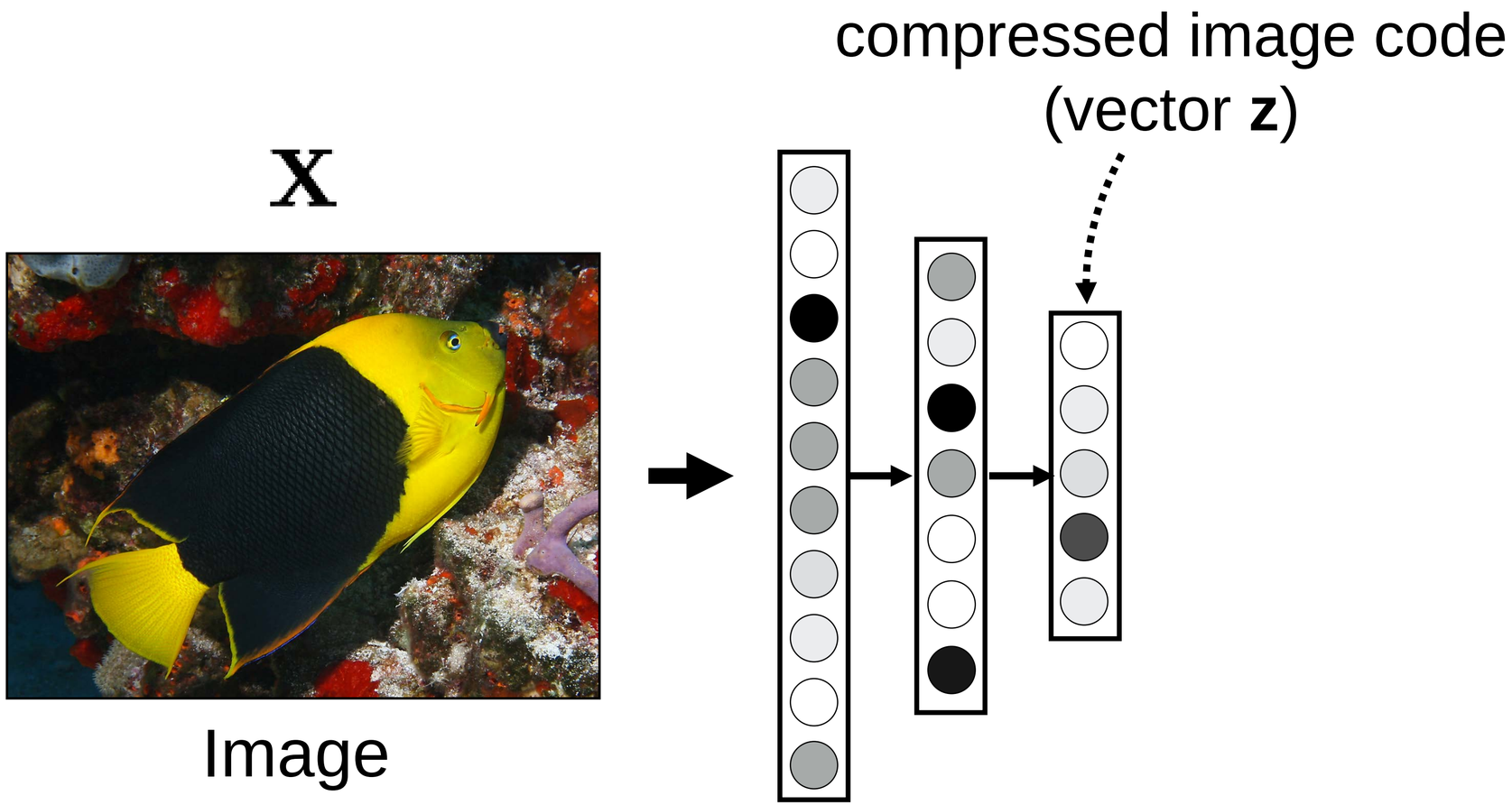
- What’s gonna happen next?
- What’s to my left?
- What can I touch?
- What will make a sound?
- Etc.

Self-Supervision

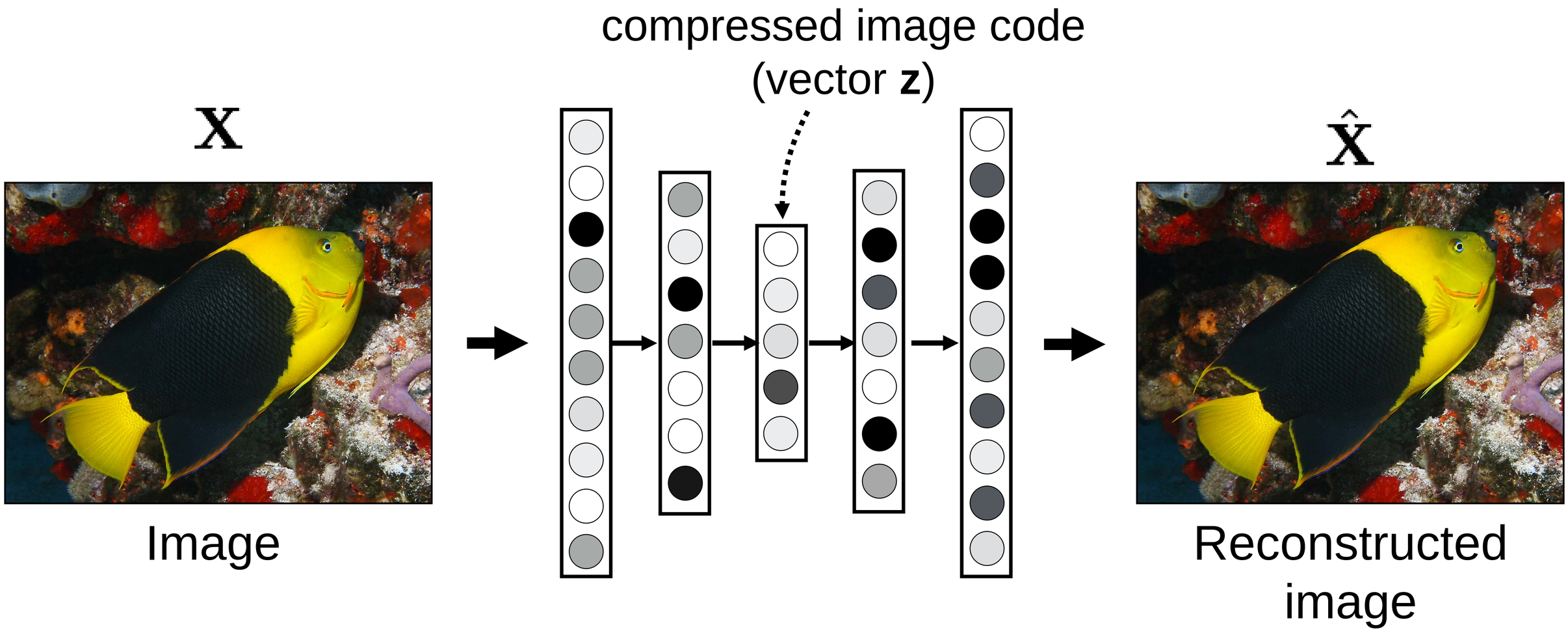


Drawing Hands, M.C. Escher, 1948

Auto-encoders

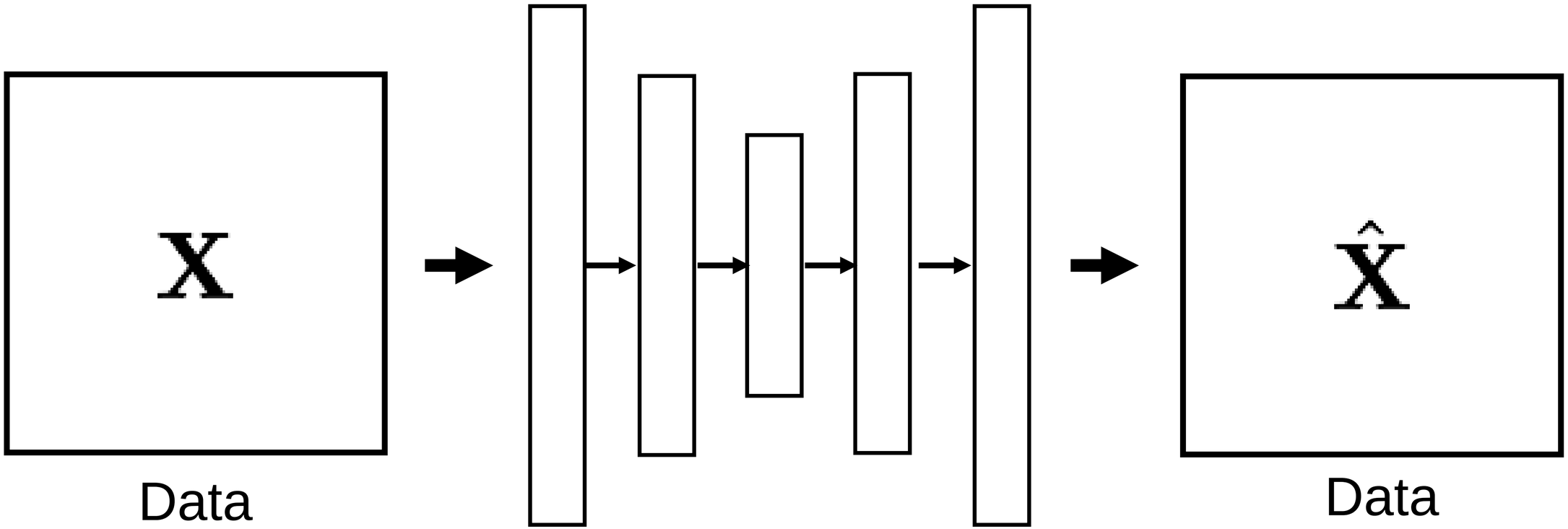


Auto-encoders



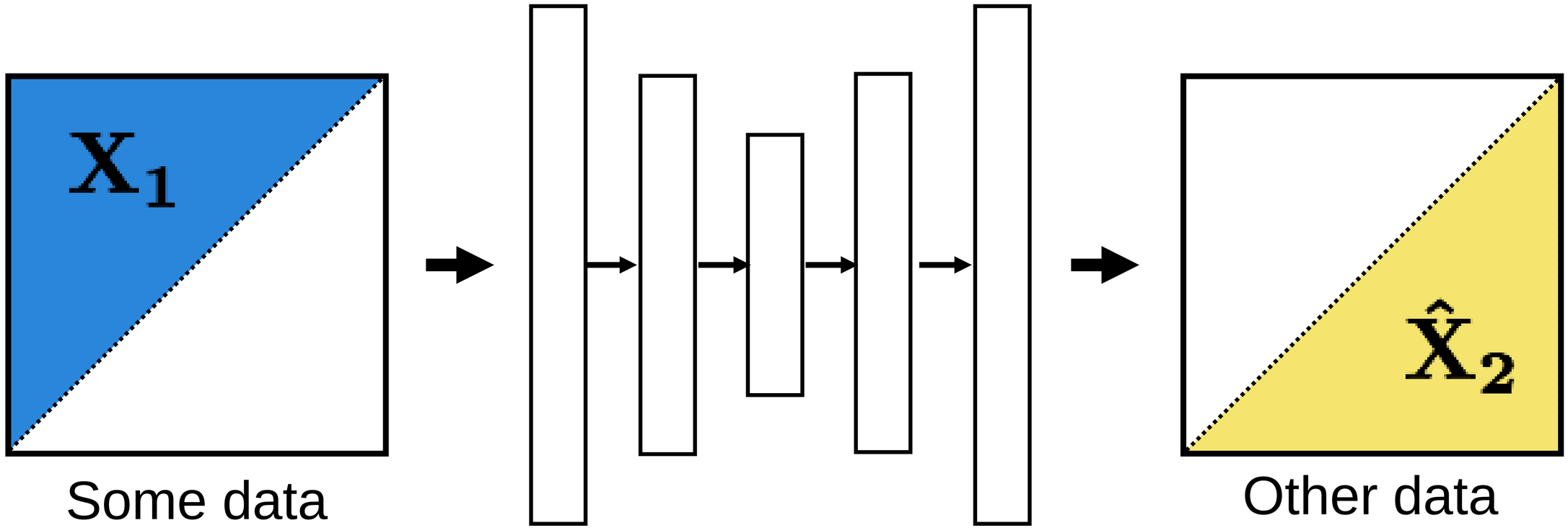
[e.g., Hinton & Salakhutdinov, Science 2006]

Data compression

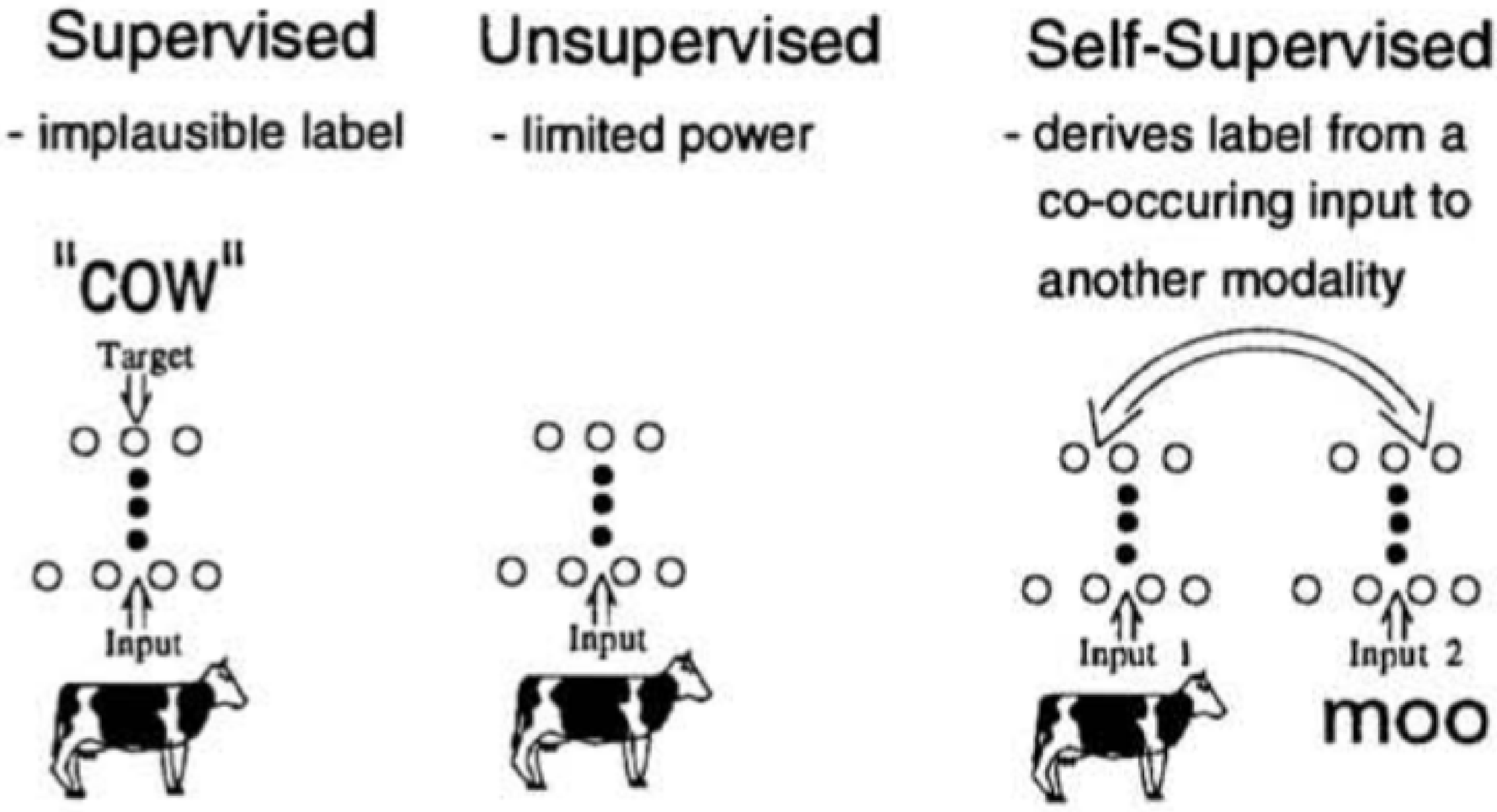


[Hinton & Salakhutdinov, Science 2009]

Data *prediction*

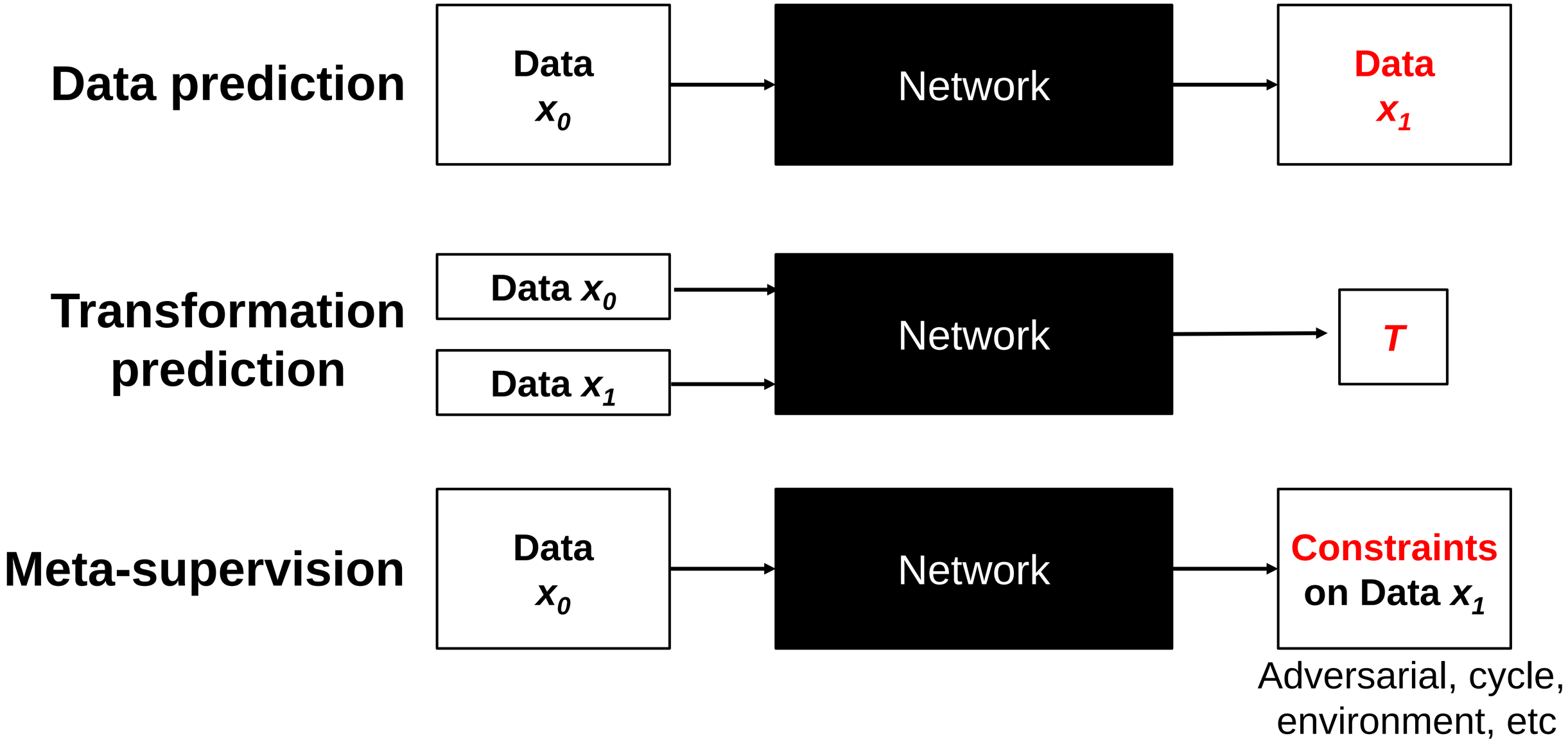


Self-Supervision in Multisensory Learning



Virginia De Sa, “Learning classification from unlabelled data”, NIPS 1994

(Partial) Taxonomy of Self-Supervision



Self-Supervision as Data Prediction

Data prediction

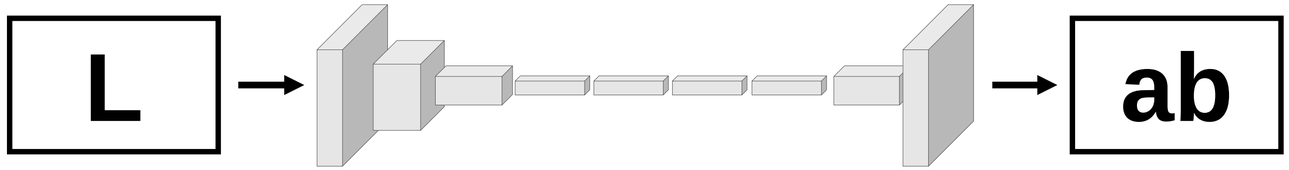
$\mathcal{F}$

Grayscale image: L channel

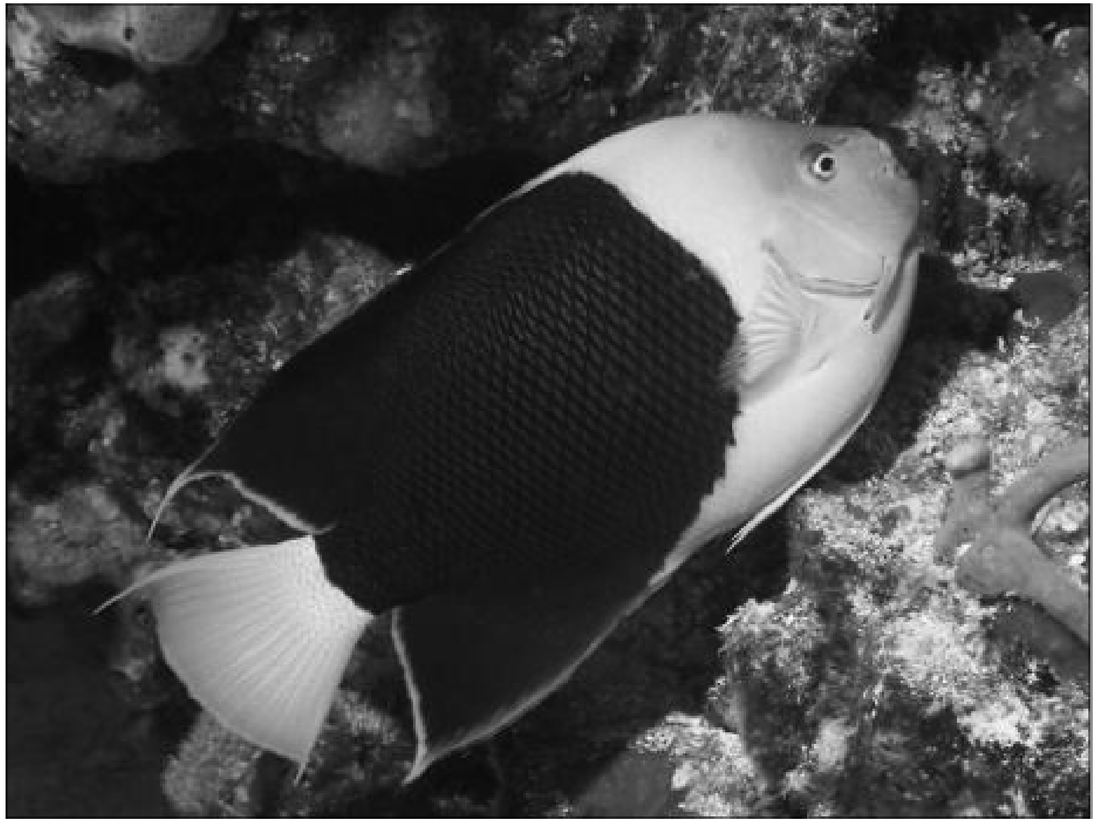
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: ab channels

$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



[Zhang, Isola, Efros, ECCV 2016]



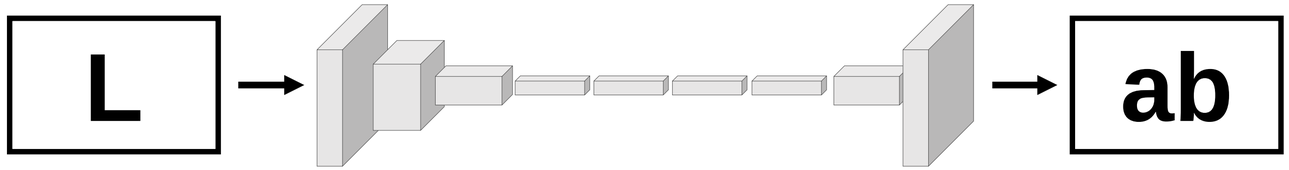
$\mathcal{F}$

Grayscale image: L channel

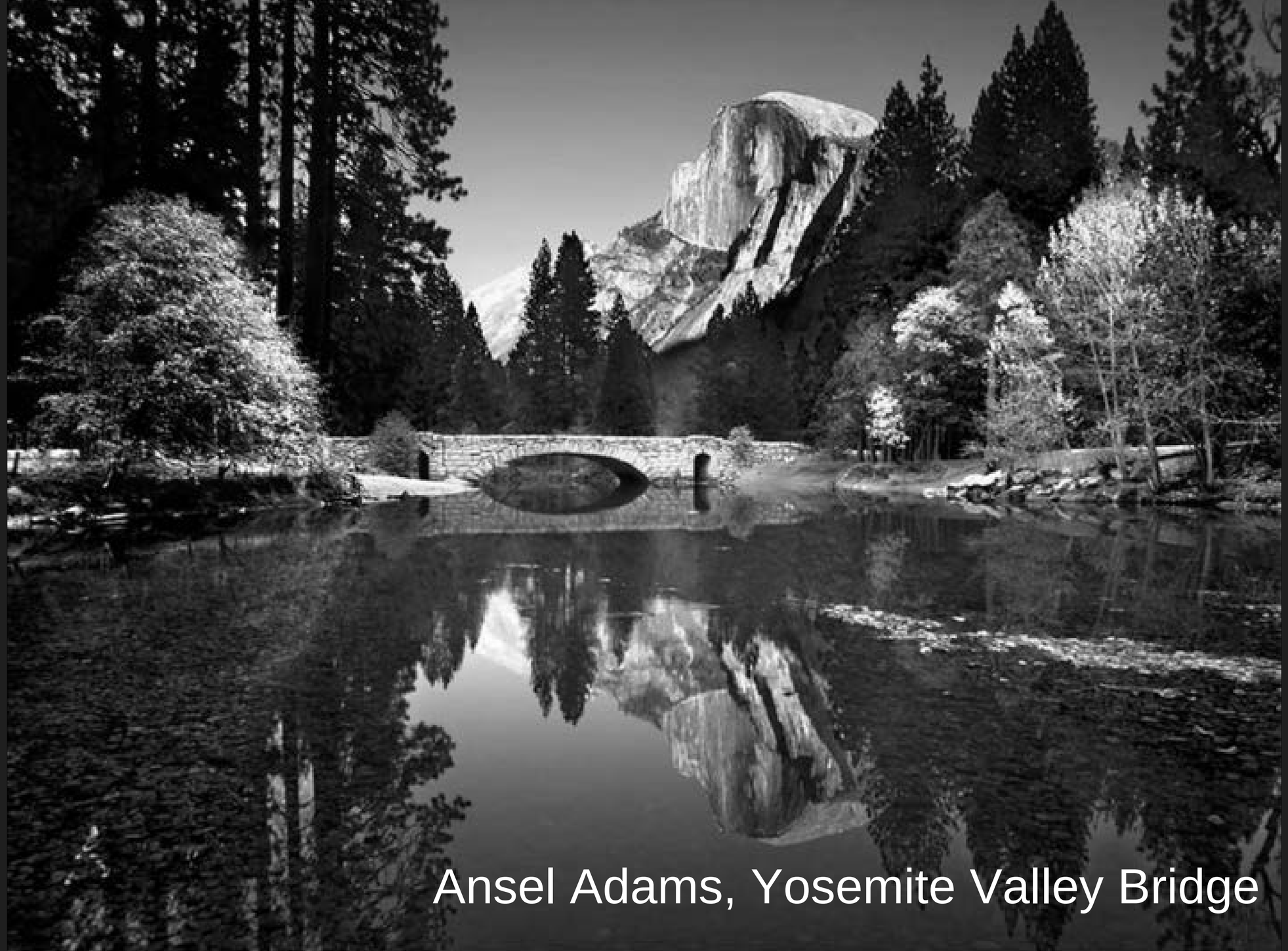
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: ab channels

$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$

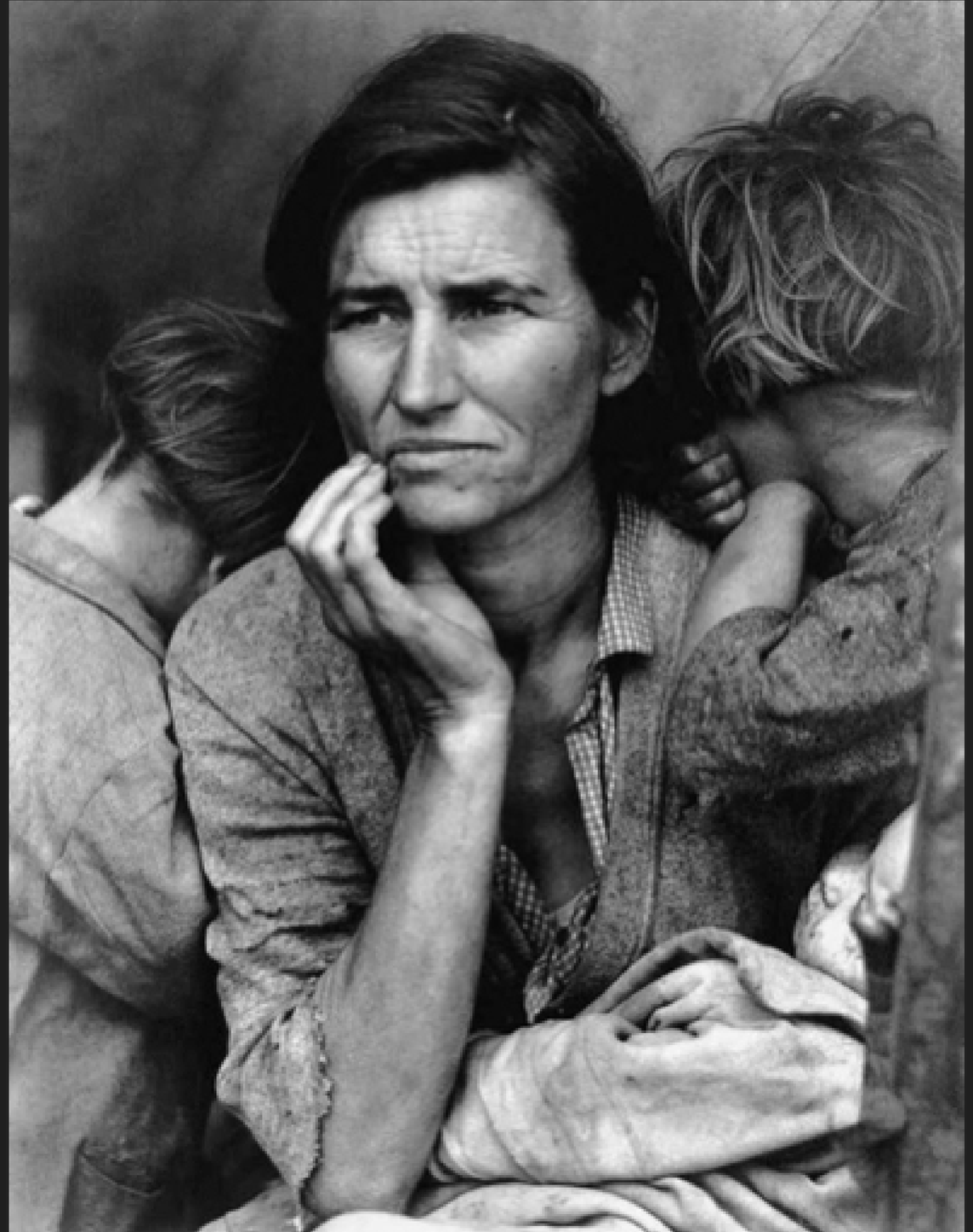


[Zhang, Isola, Efros, ECCV 2016]

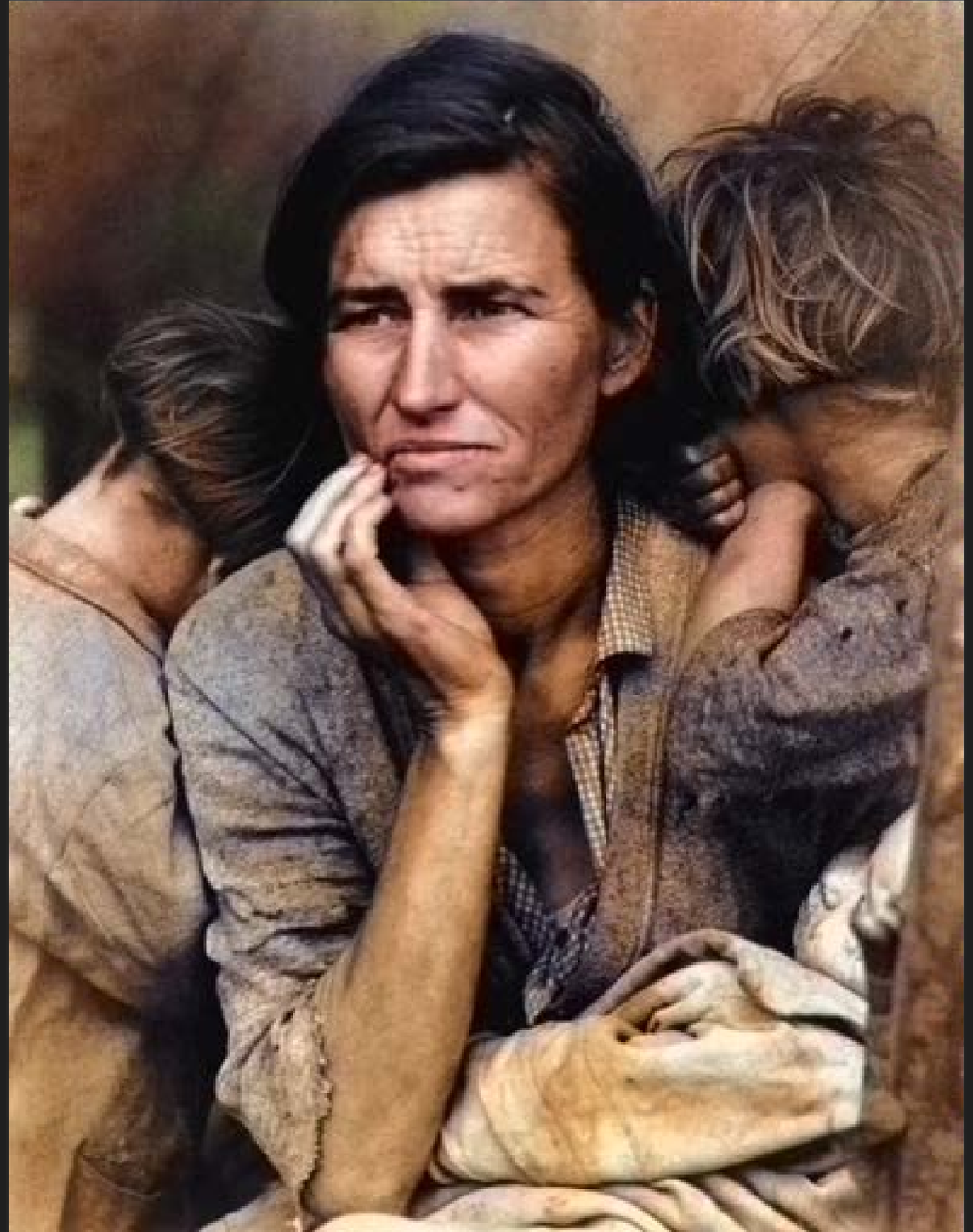


Ansel Adams, Yosemite Valley Bridge


Our result



Migrant Mother
Dorothea Lange
1936



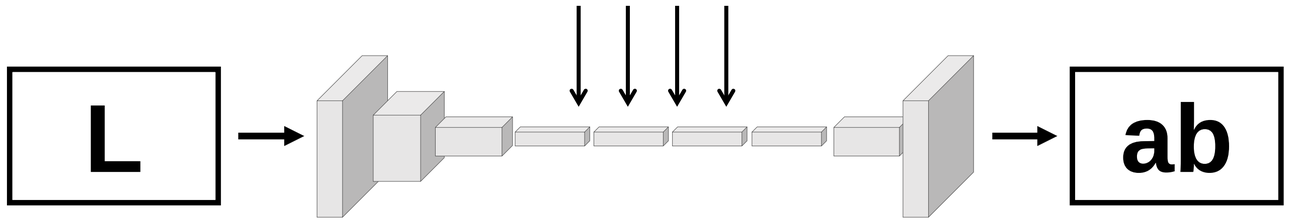
Our result

$\mathcal{F}$

Grayscale image: L cha information: ab channels

$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$ $\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$

Semantics? Higher-level abstraction?

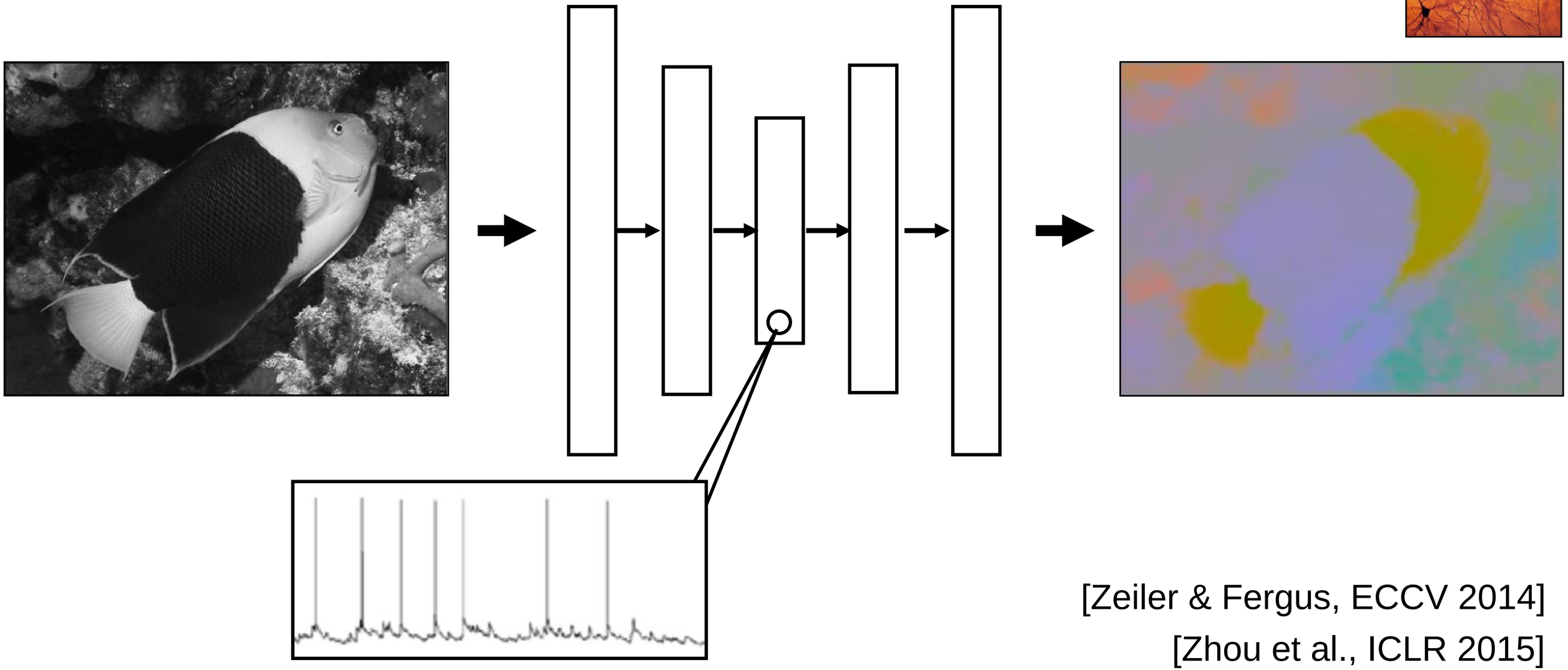


[Zhang, Isola, Efros, ECCV 2016]

Instructive failure

Instructive failure

Deep Net “Electrophysiology”



[Zeiler & Fergus, ECCV 2014]

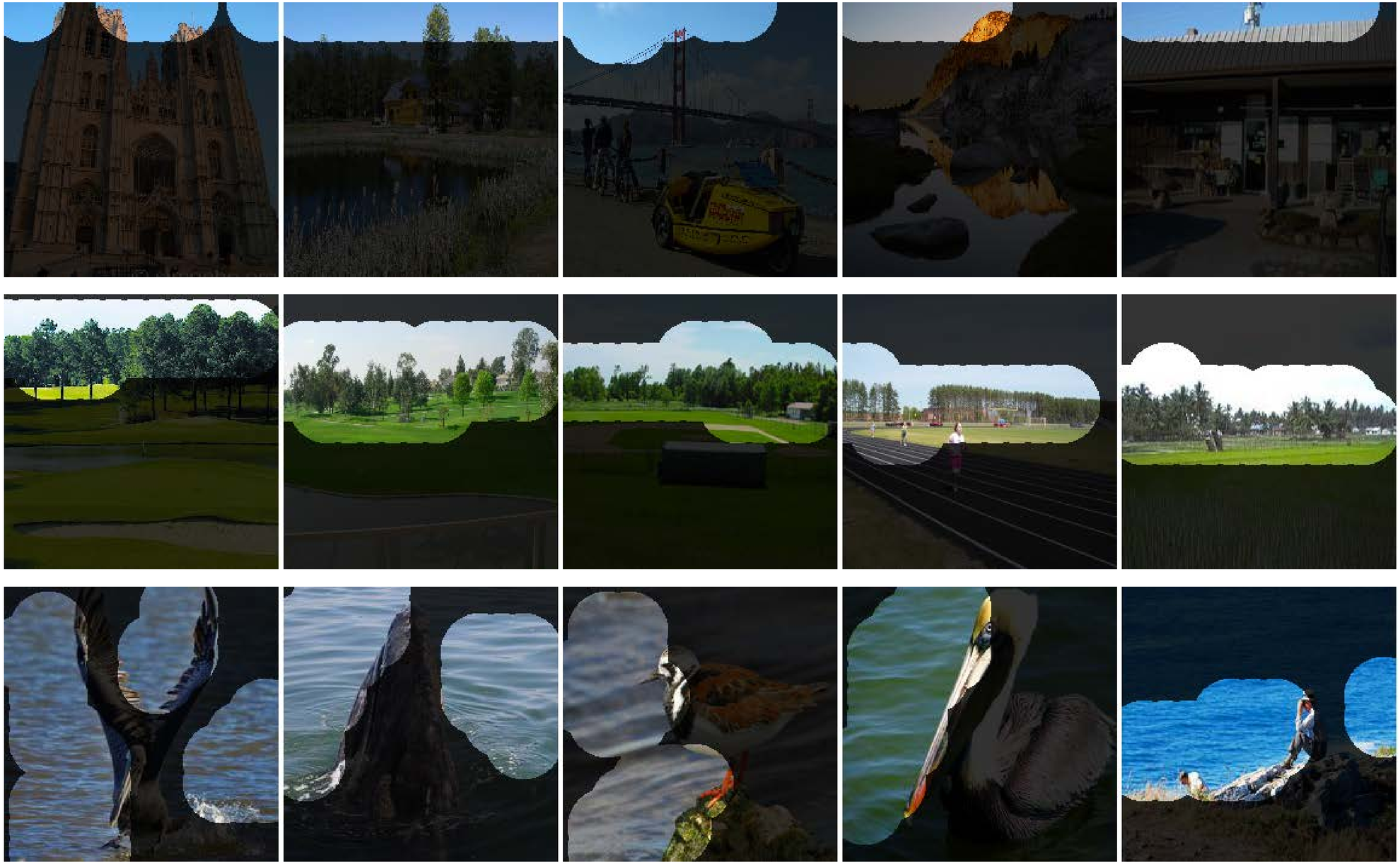
[Zhou et al., ICLR 2015]

Hidden Unit (conv5) Activations

sky

trees

water

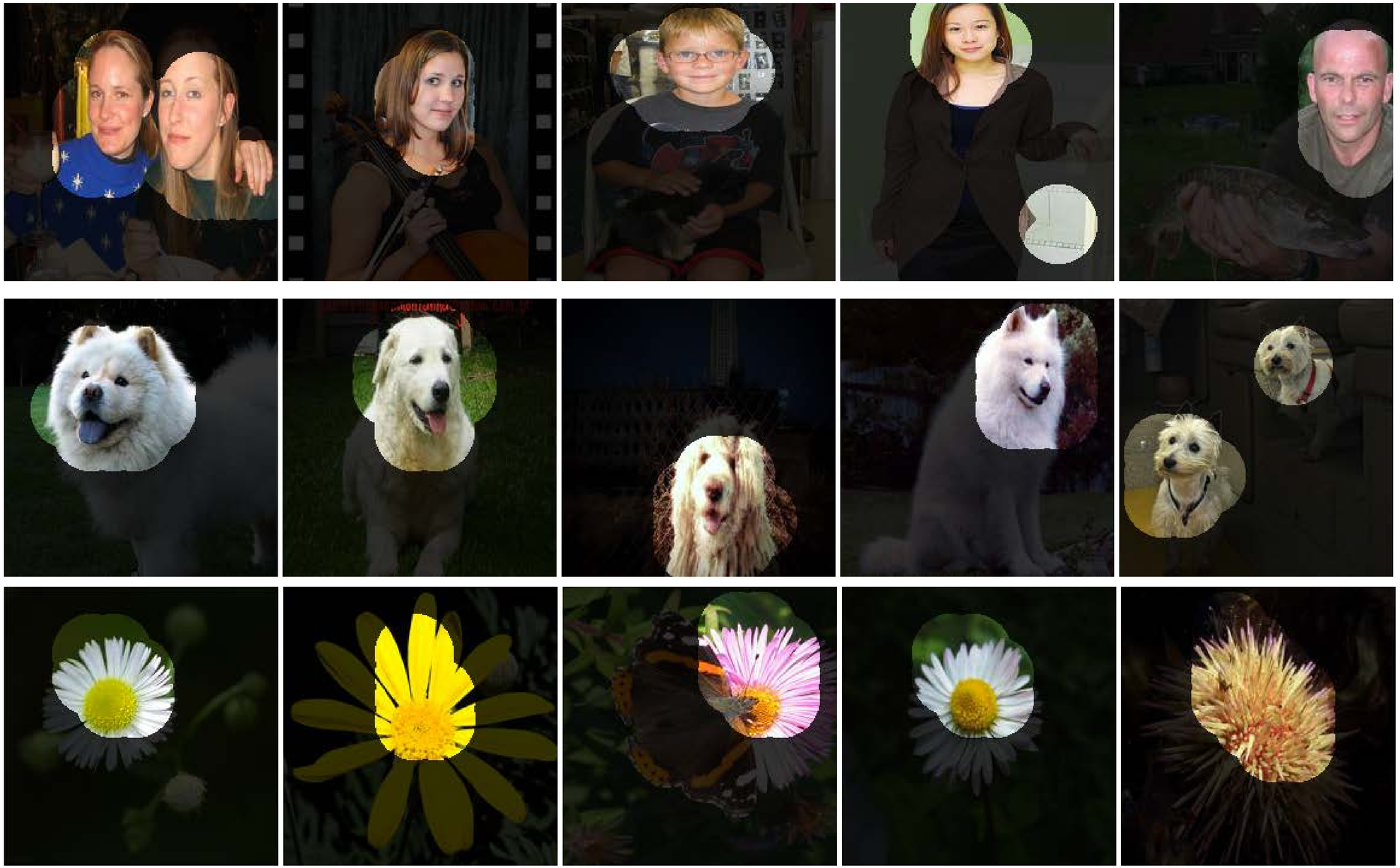


Hidden Unit (conv5) Activations

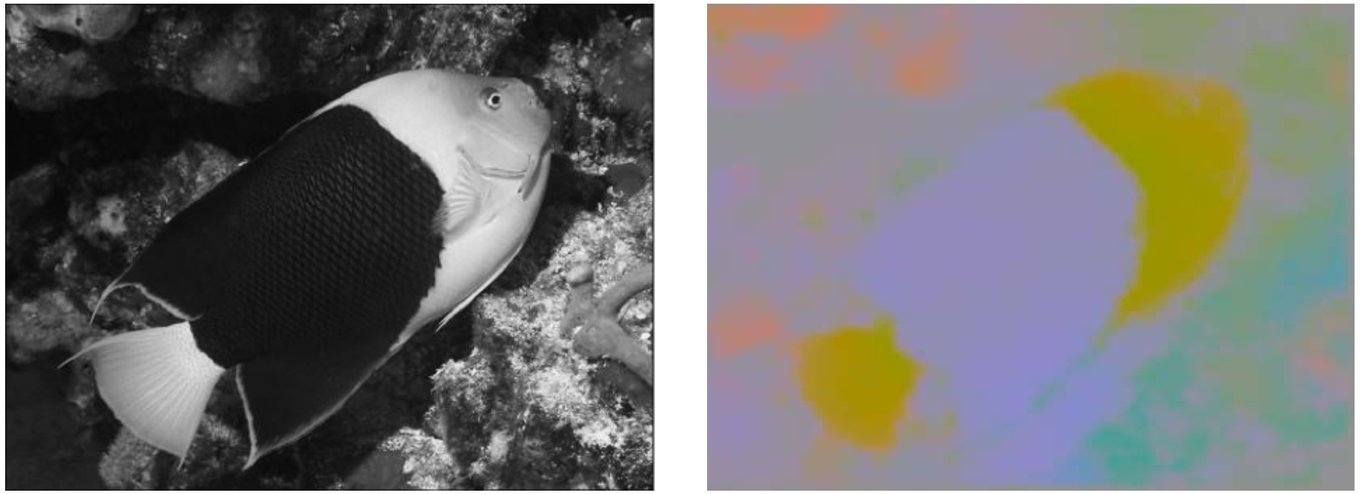
faces

dog faces

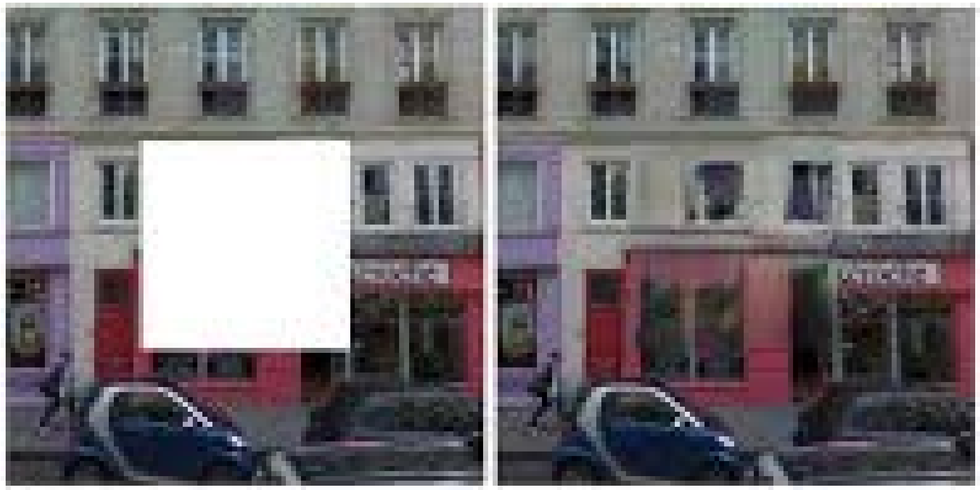
flowers



Self-supervision as data prediction

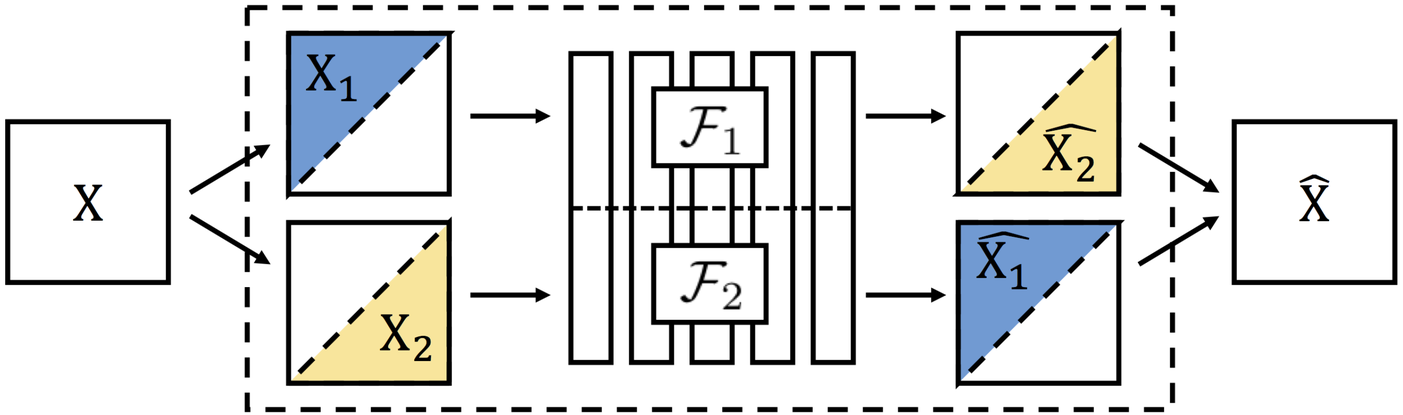


Zhang et al, ECCV 2016



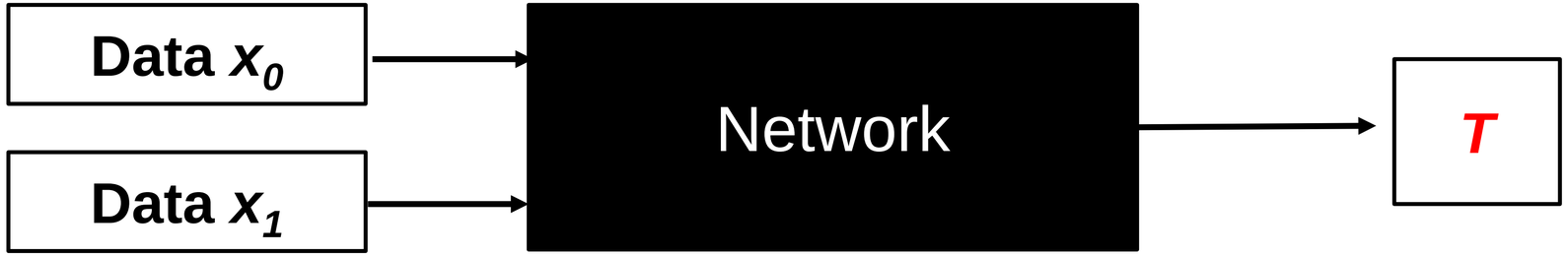
Pathak et al. CVPR 2016

Split-brain Auto-encoder

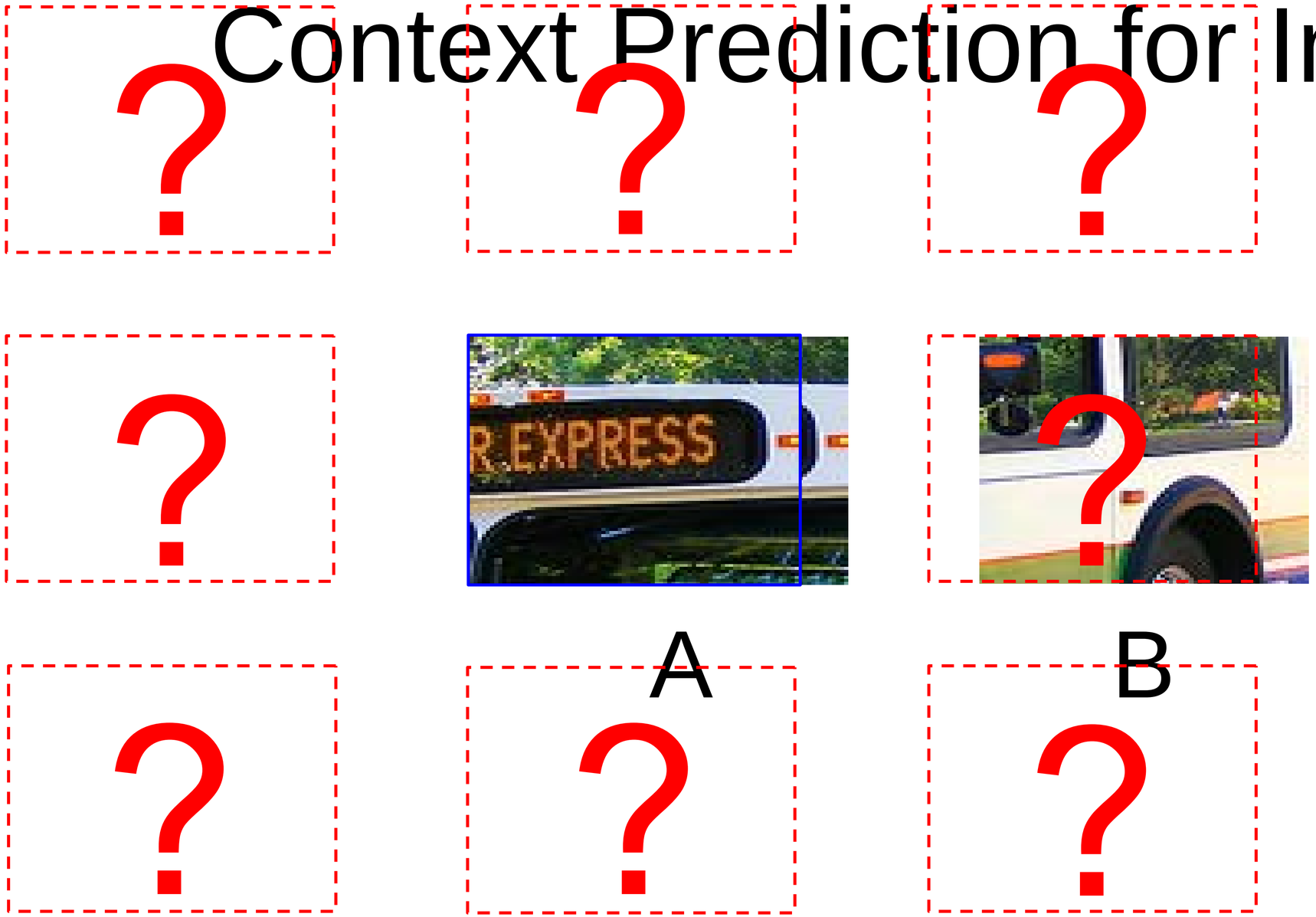


Zhang et al, CVPR 2017

Self-supervision as transformation predication

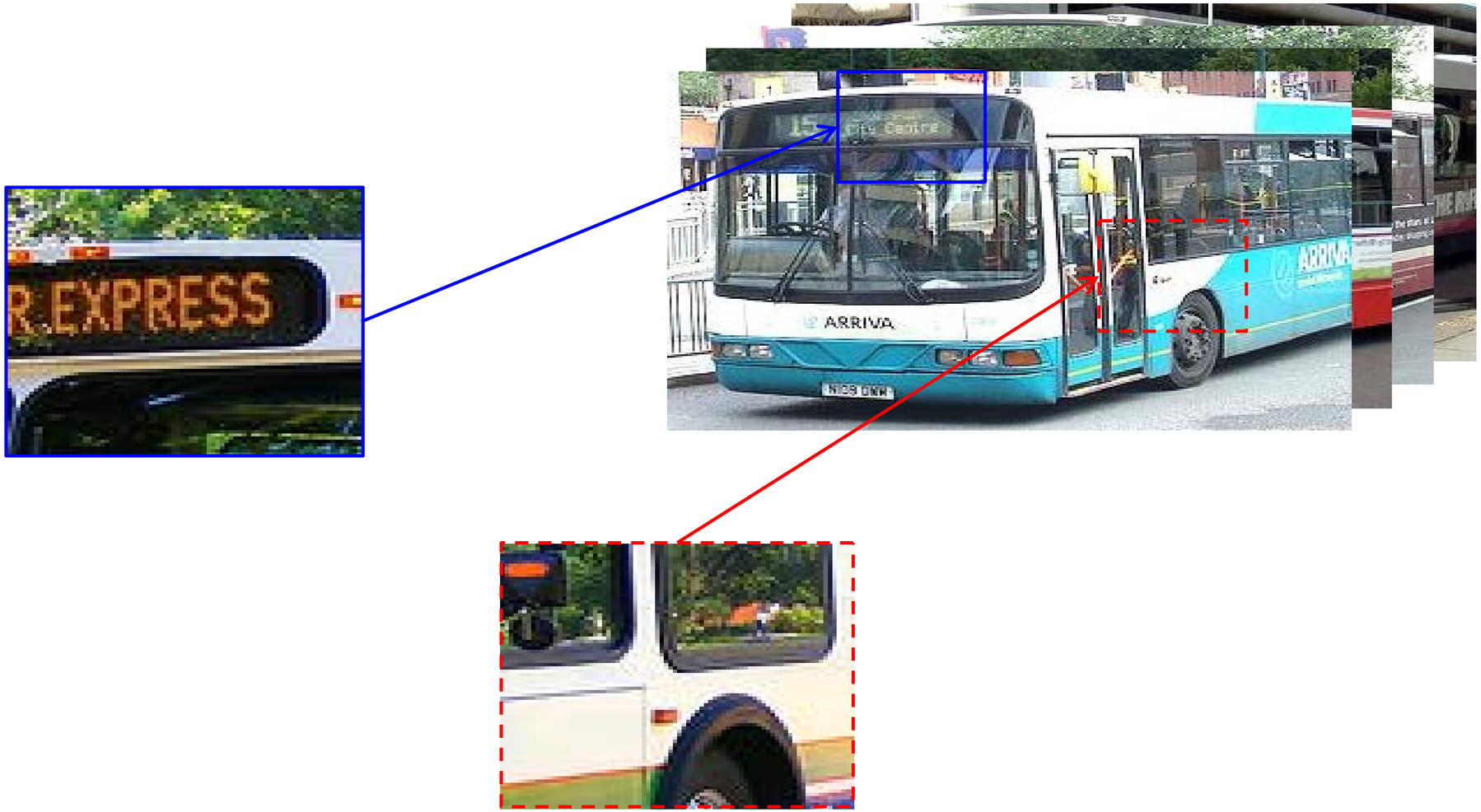


Context Prediction for Images

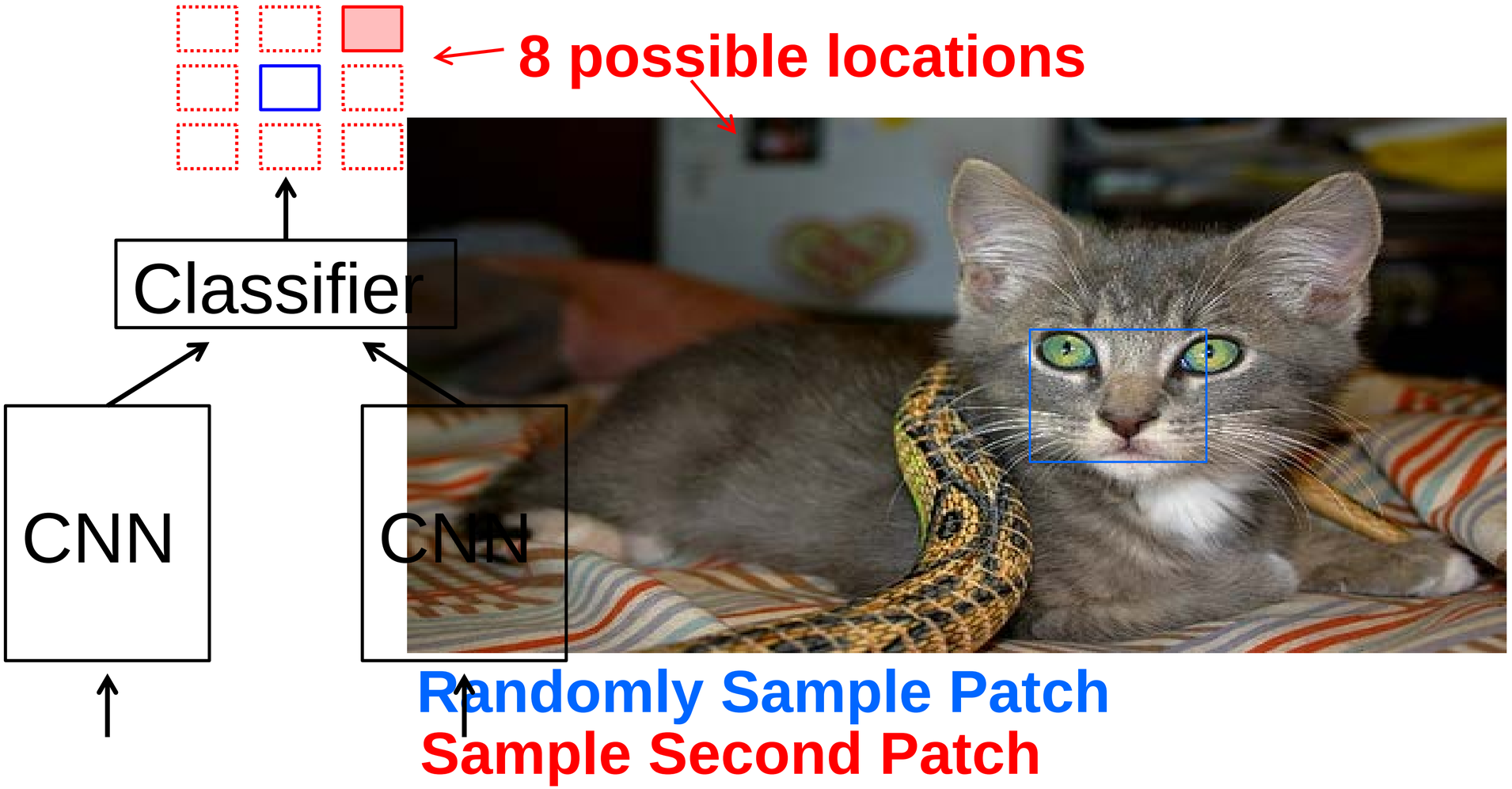


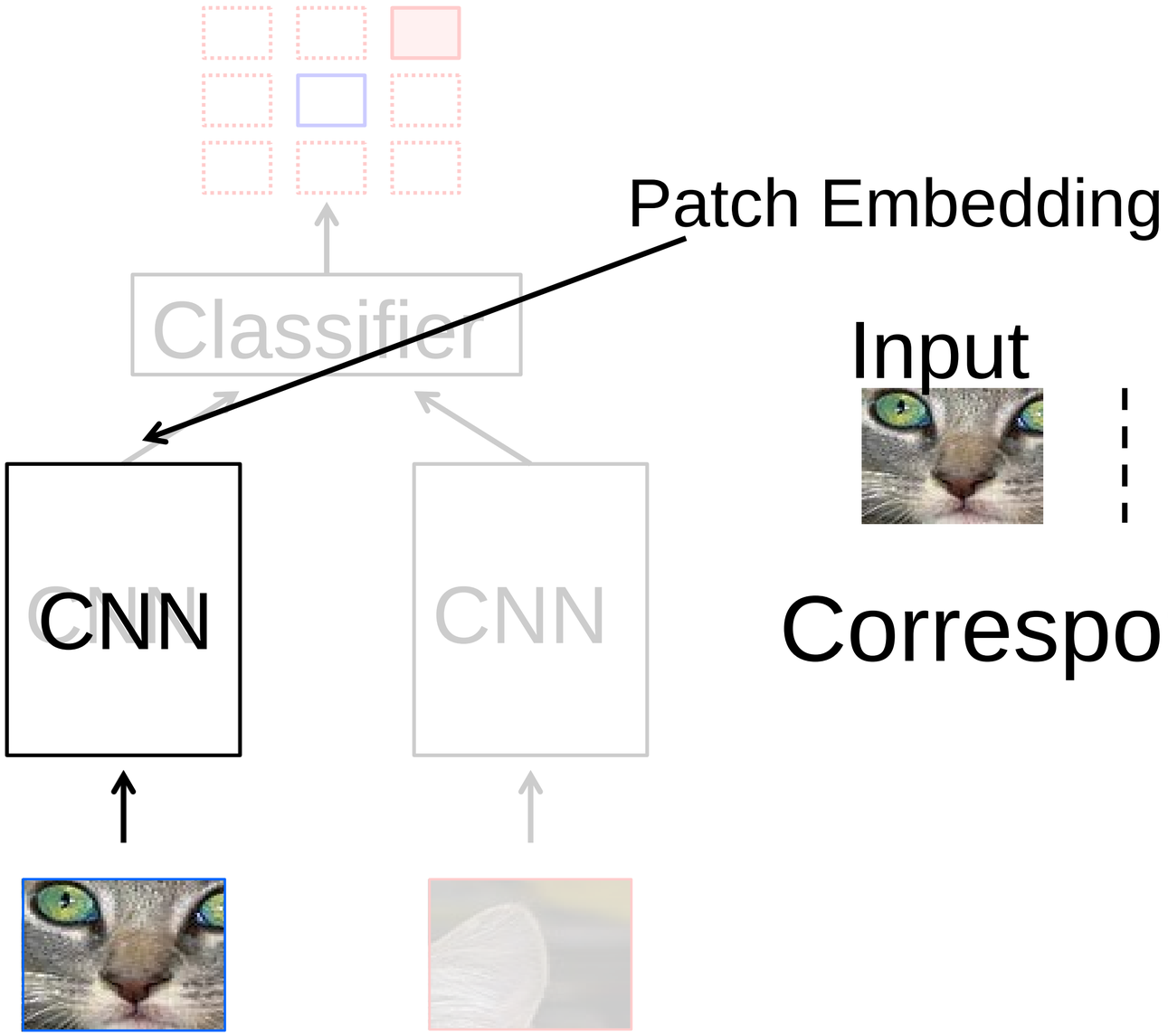
Deorsch, Gupta, Efros ICCV 2015

Semantics from a non-semantic task



Relative Position Task





Input

Nearest Neighbors

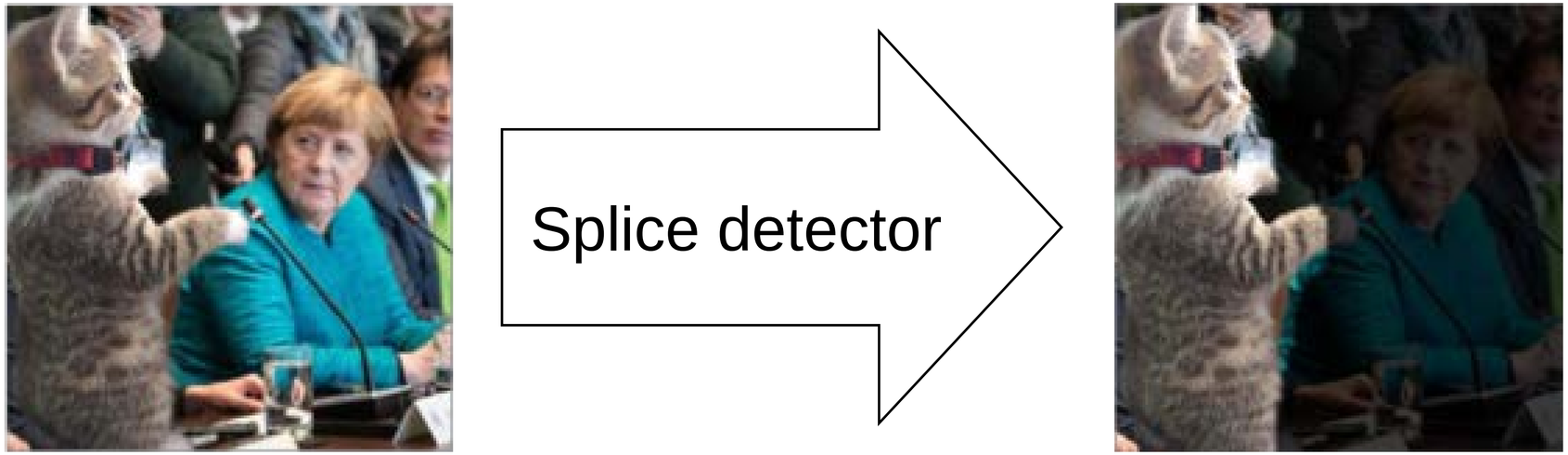
Correspondence ***across*** instances!

ECCV 2018

Image Splice Detection via Learned Self-Consistency

Minyoung Huh*[1,2] **Andrew Liu*[1]** Andrew Owens[1] Alexei A. Efros[1]

UC Berkeley[1] Carnegie Mellon University[2]

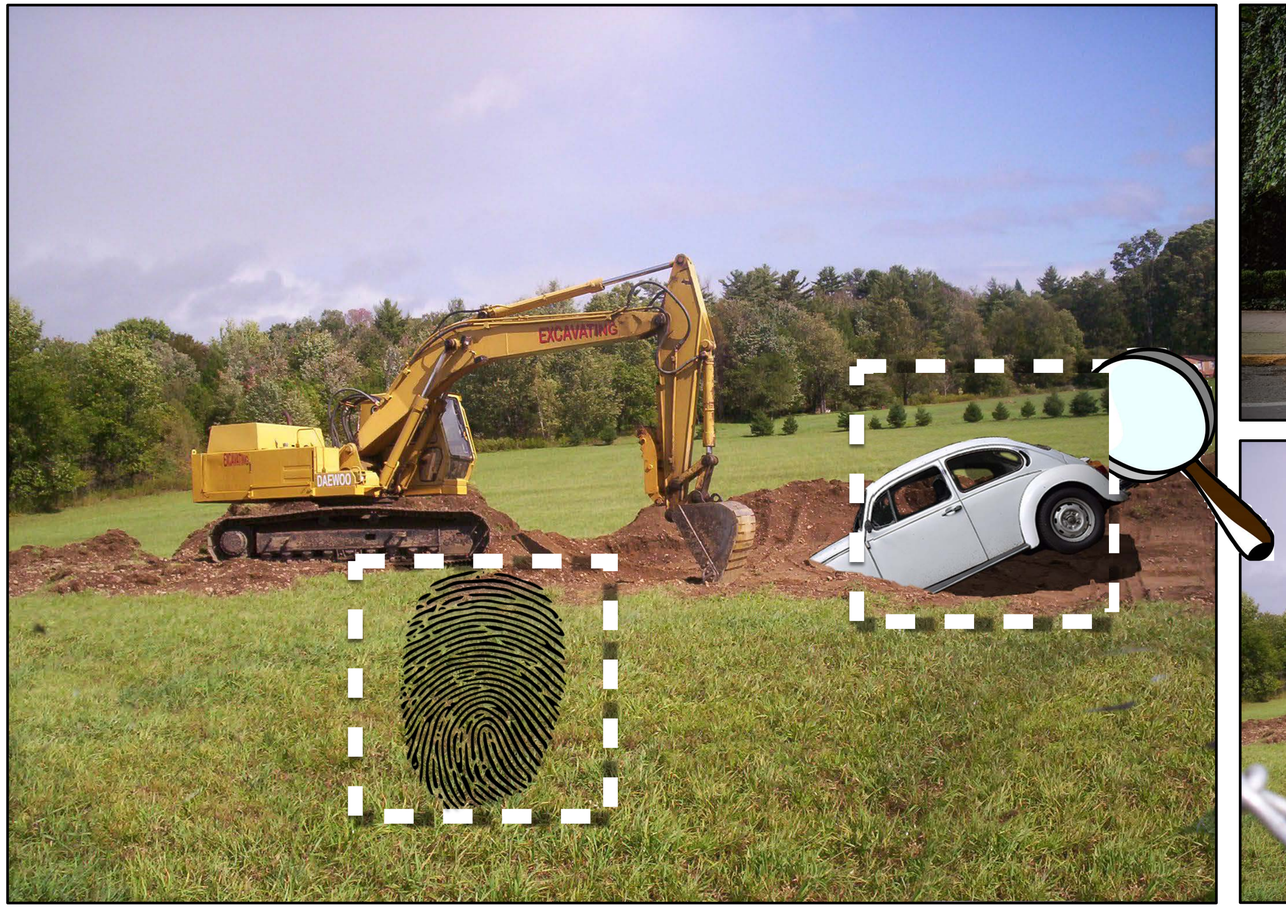


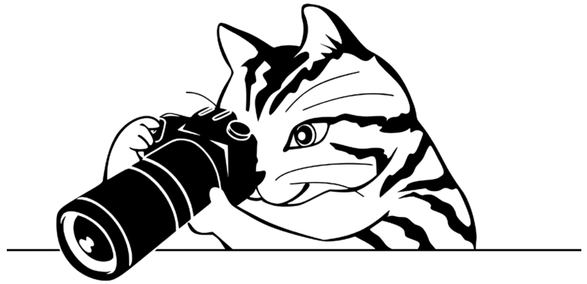
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH


EXCAVATING
DAEWOO


EXCAVATING
DAEWOO


EXCAVATING
DAEWOO

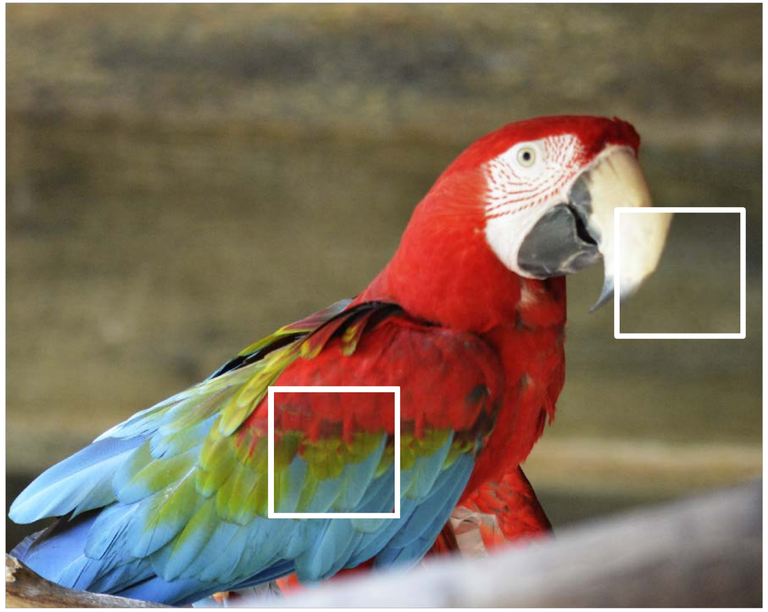


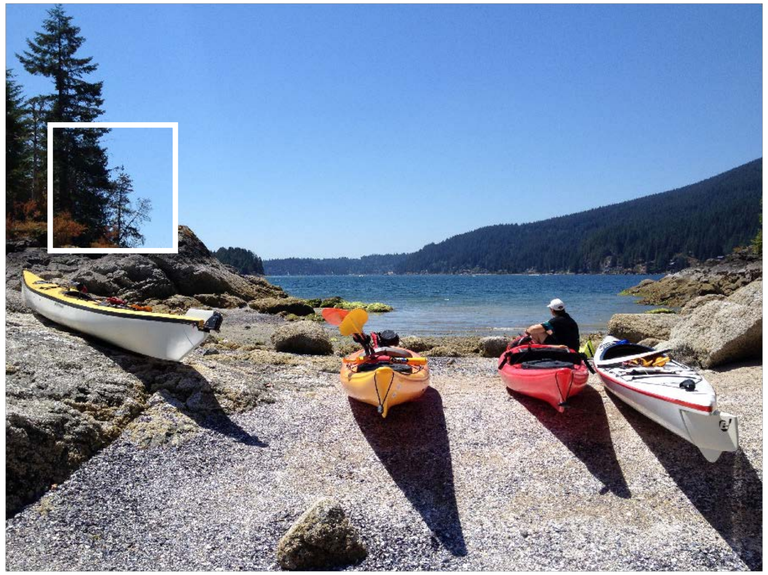


Inconsistent

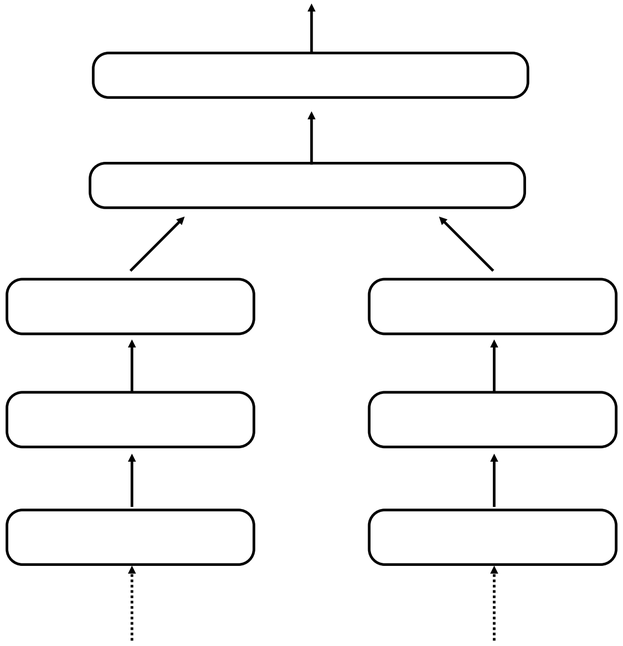
Consistent

Learning same-image consistency





Same Imaging Pipeline?

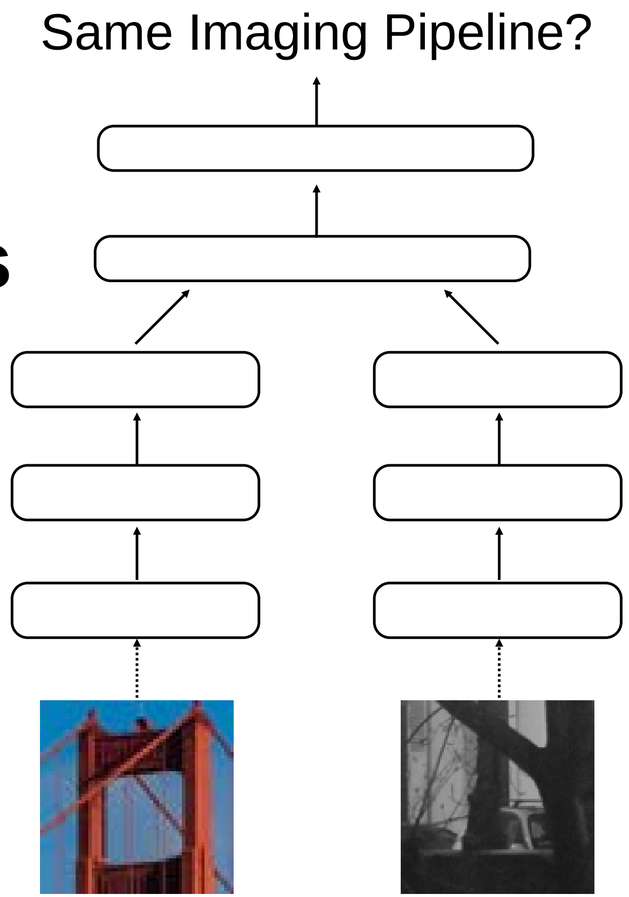


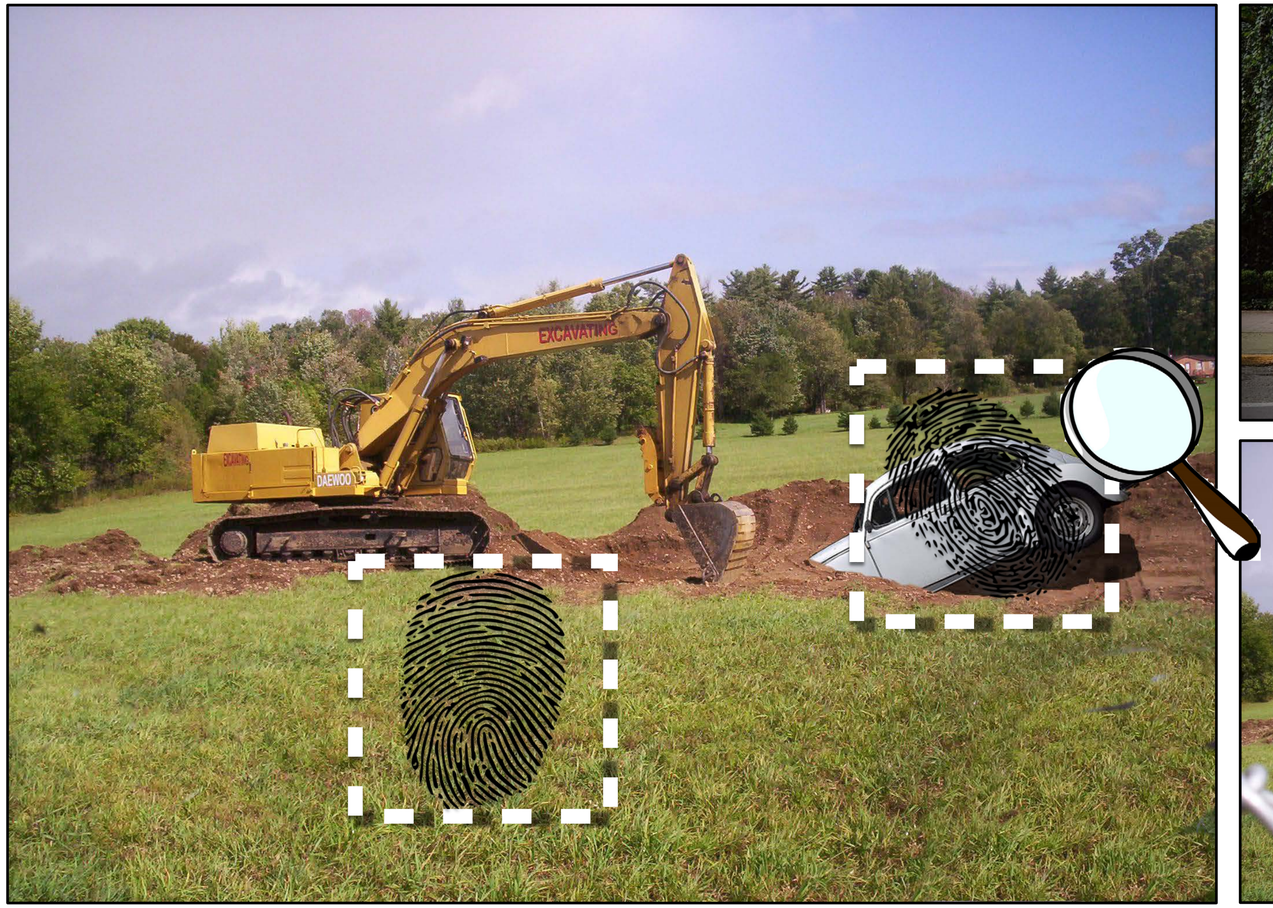
Consistent

Inconsistent

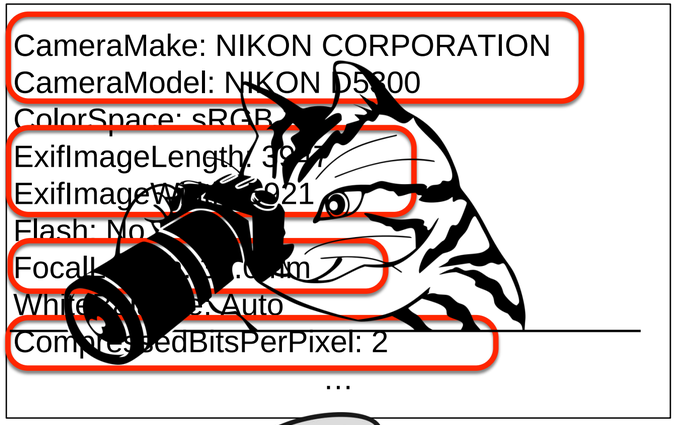
Might be too slow…

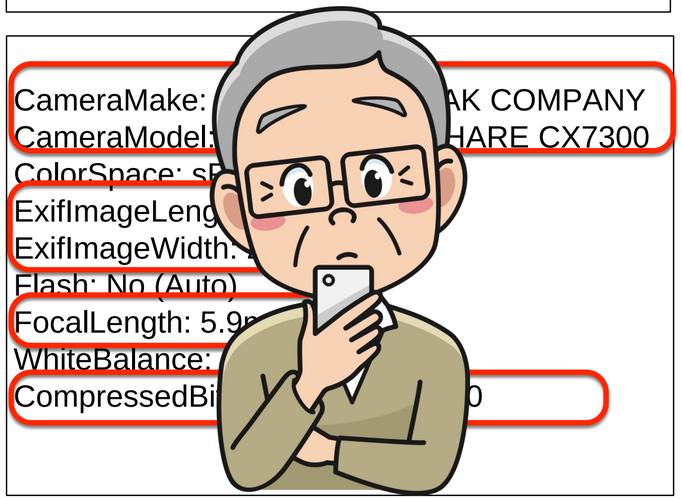
Negative examples are too easy!



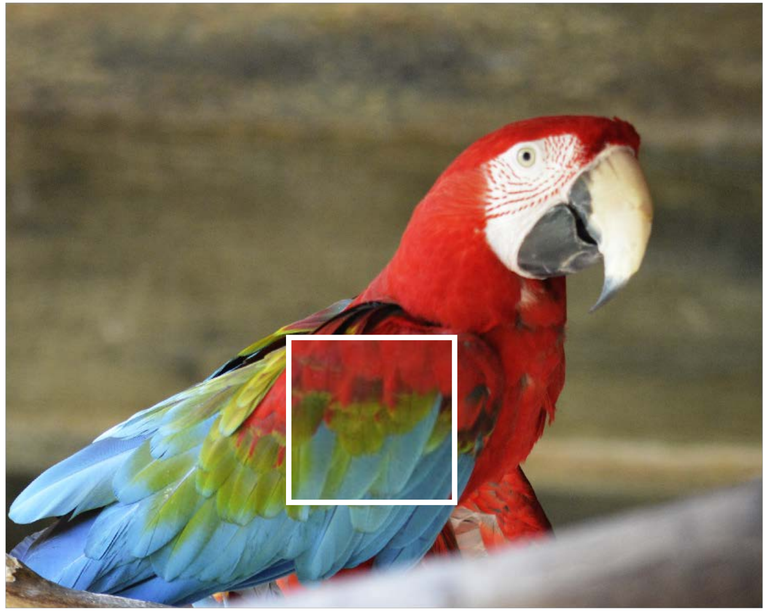


EXIF metadata

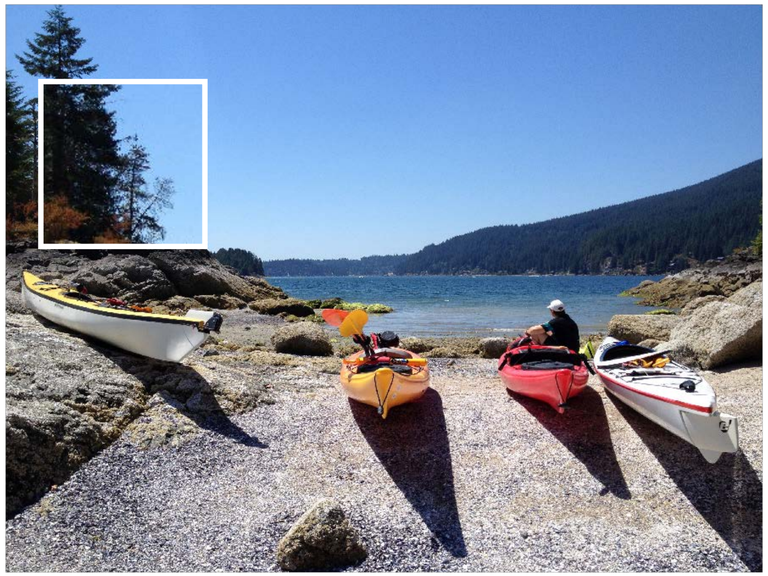




Learning EXIF self-consistency

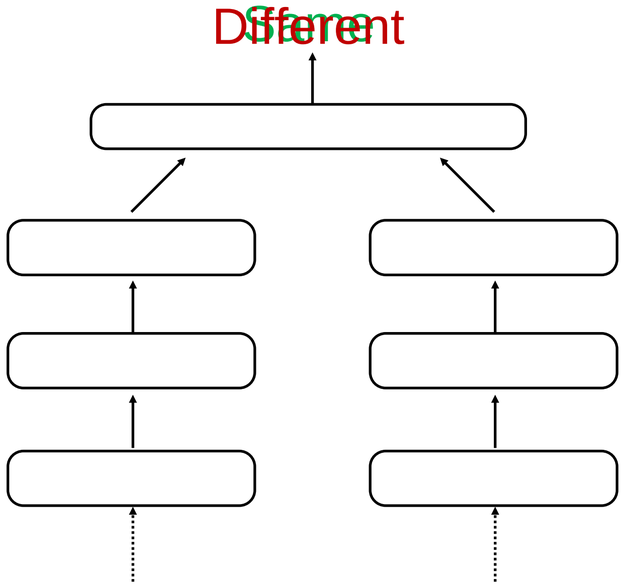


CameraMake: NIKON CORPORATION
CameraModel: NIKON D3200
ColorSpace: Uncalibrated
ExifImageLength: 2472
ExifImageWidth: 3091
Flash: Flash did not fire
FocalLength: 90
WhiteBalance: Auto
ExposureTime: 1/100
…



CameraMake: Apple
CameraModel: iPhone 4s
ColorSpace: sRGB
ExifImageLength: 2448
ExifImageWidth: 3264
Flash: Flash did not fire
FocalLength: 107/2
WhiteBalance: Auto
ExposureTime: 1/2208
…

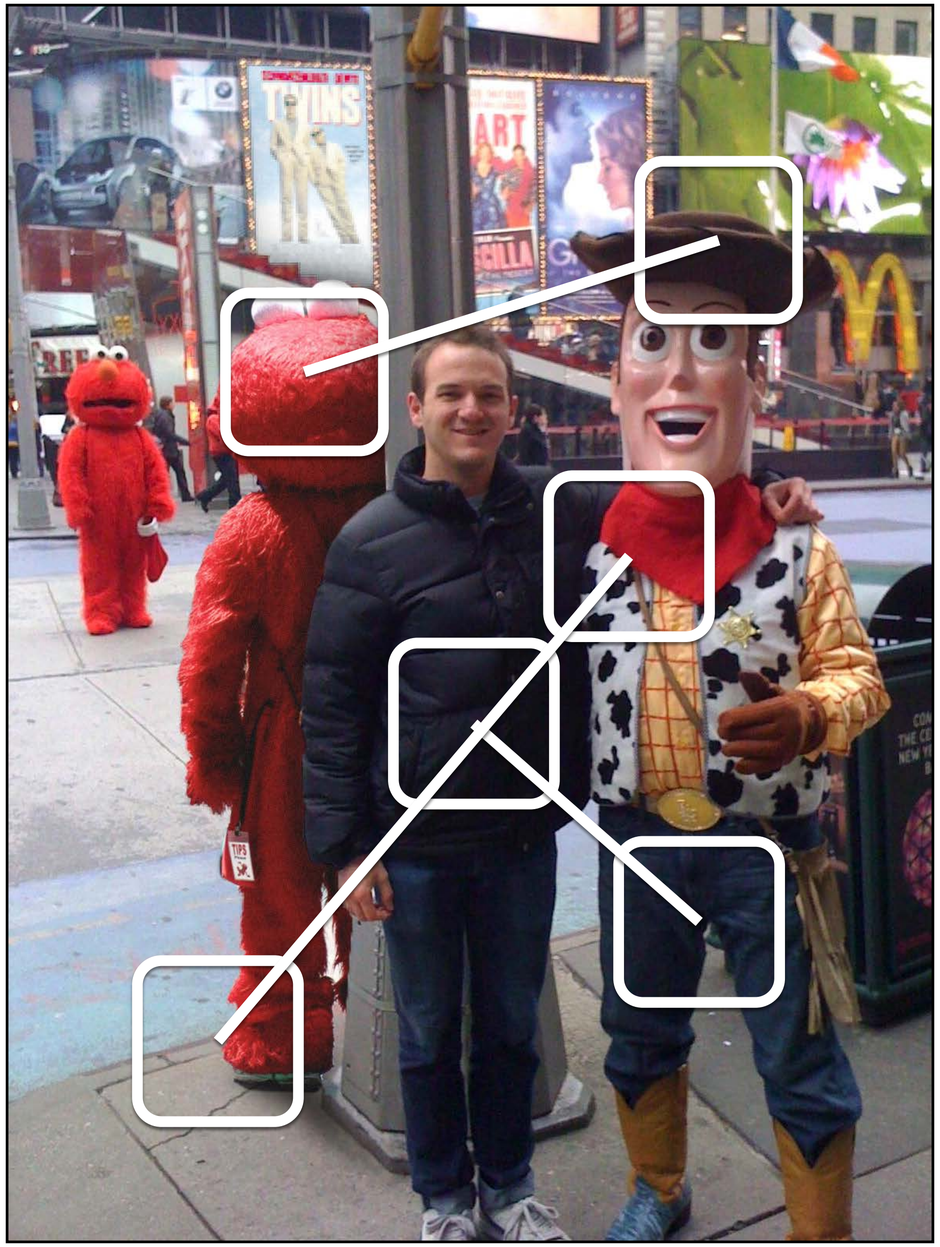
Same WhiteBalance?
Same vs. different EXIF attributes



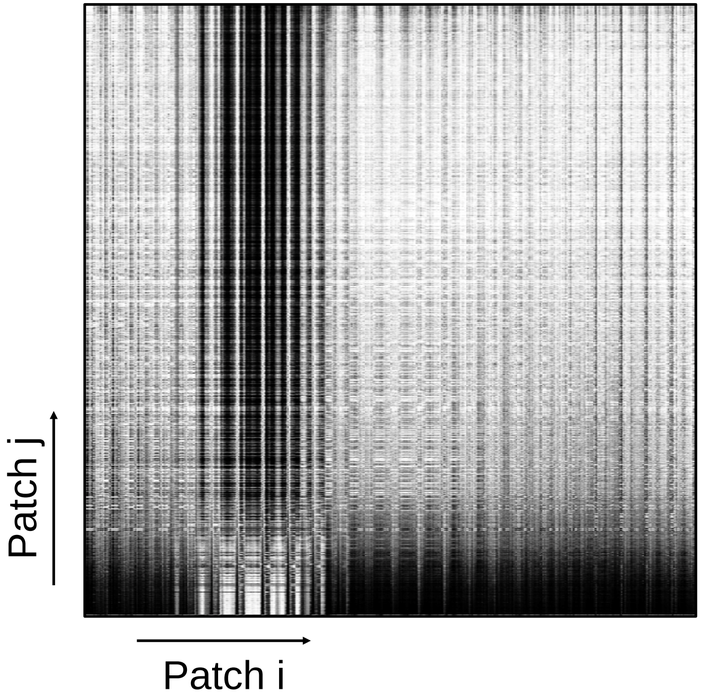
Unmanipulated Image

Manipulated Image

Manipulated Image

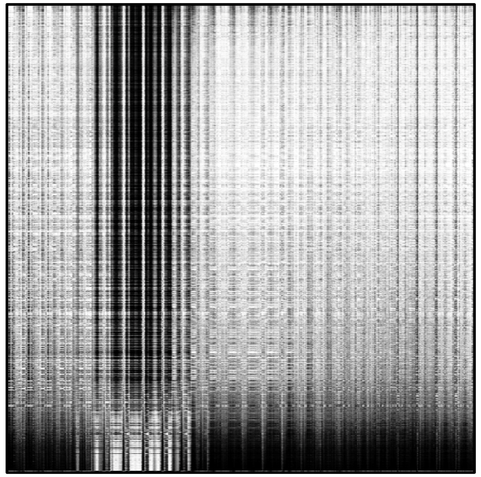


Affinity Matrix



Manipulated Image

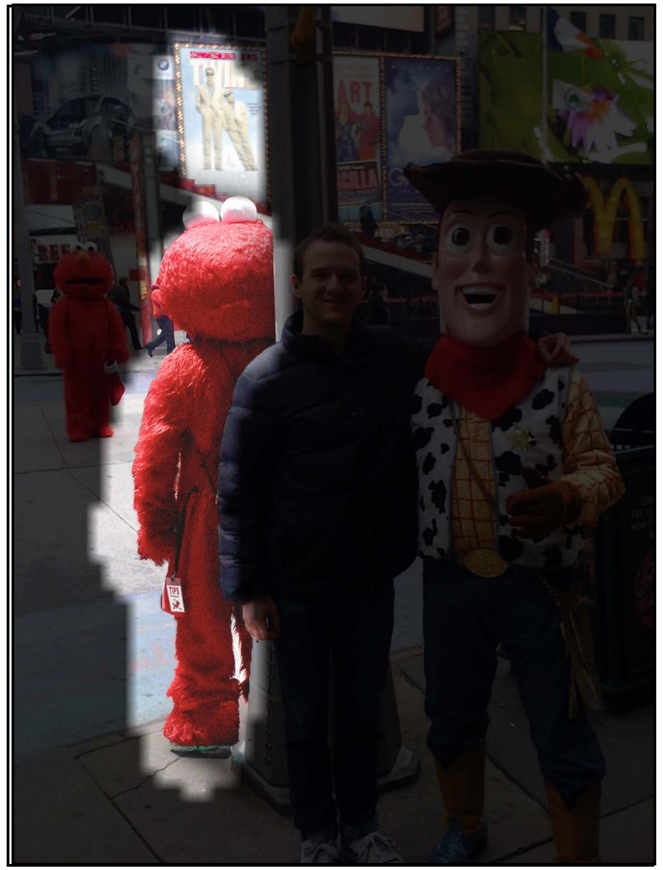
Affinity Matrix



Normalized Cuts

Mean Shift

Manipulated Region



Input Image

Manipulated Region

Our detection

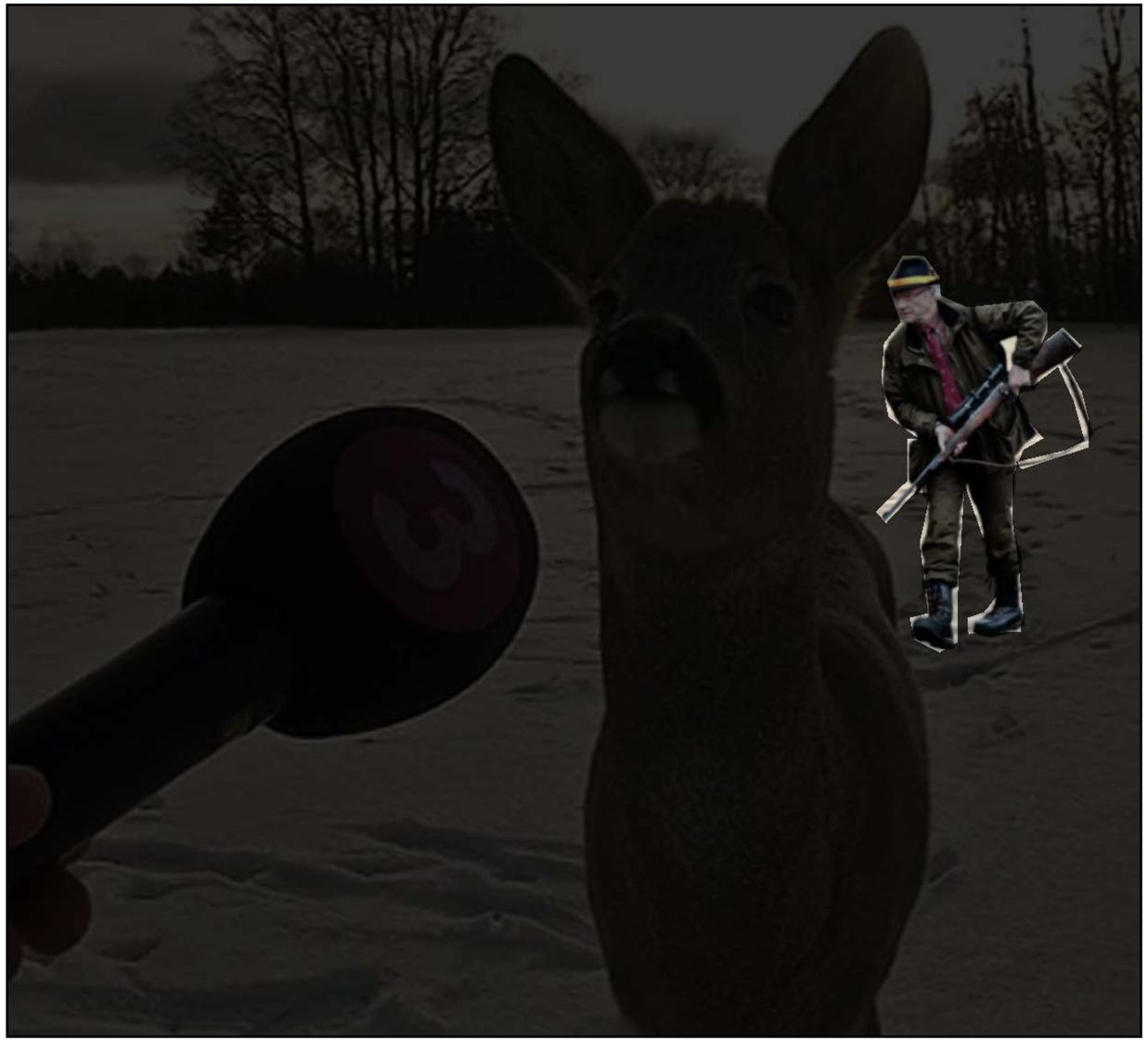
Ground Truth

Input Image

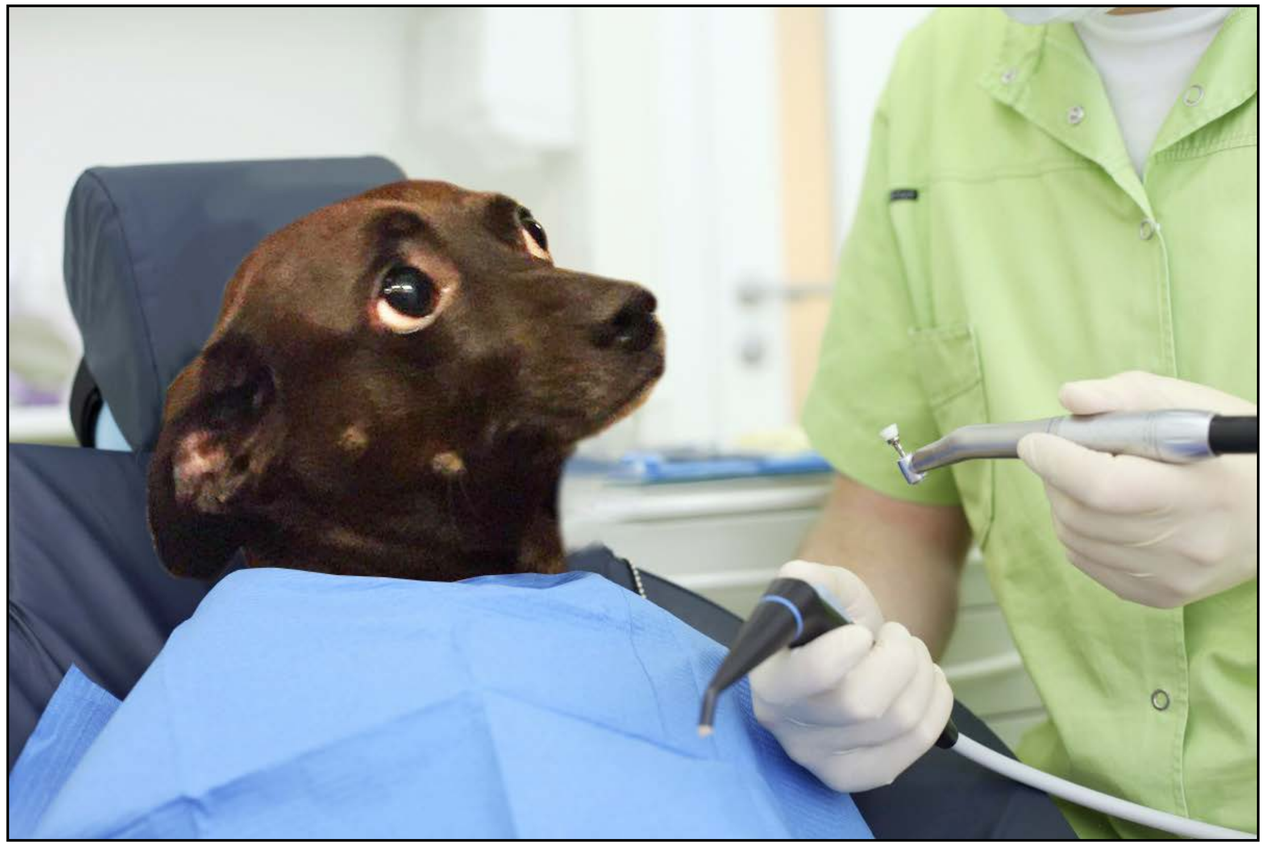
Manipulated Region

Our detection

Ground Truth

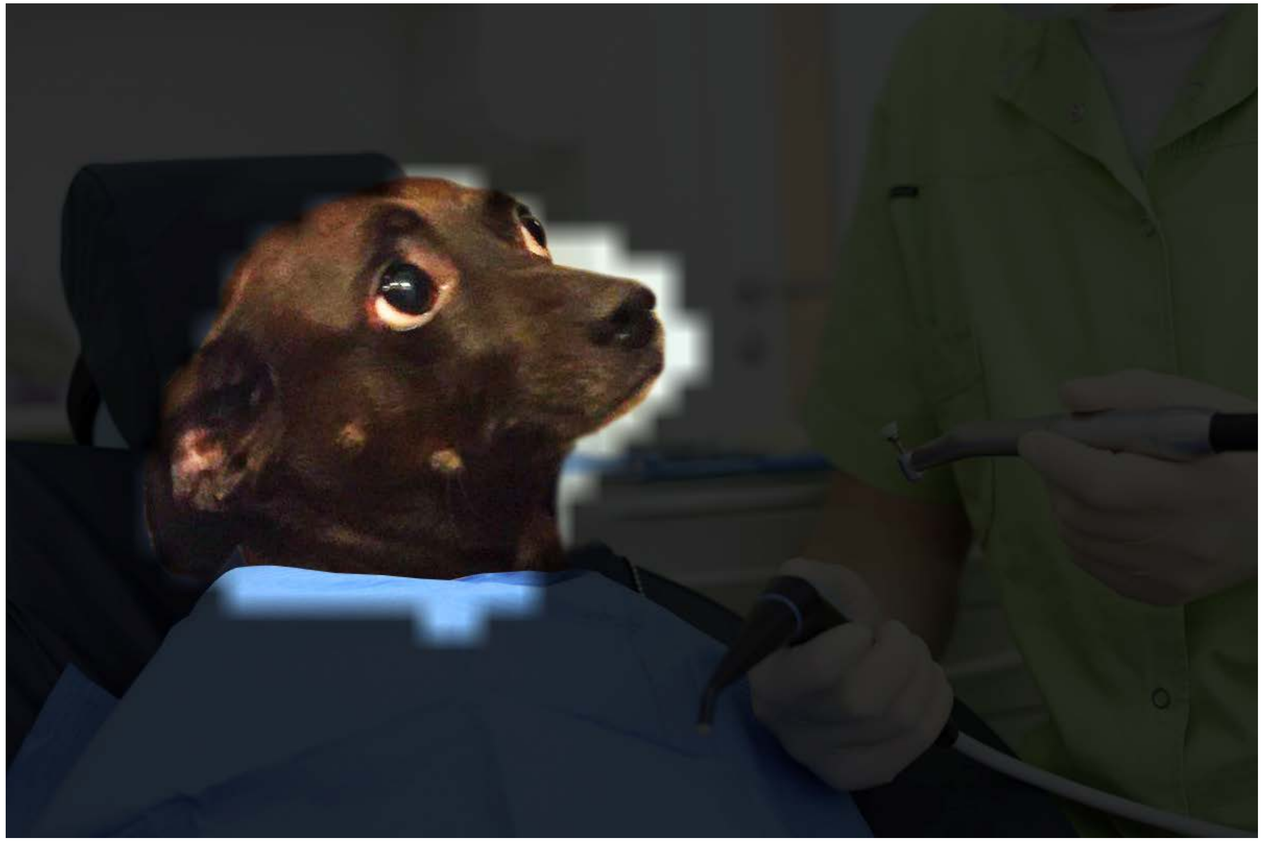


Input Image

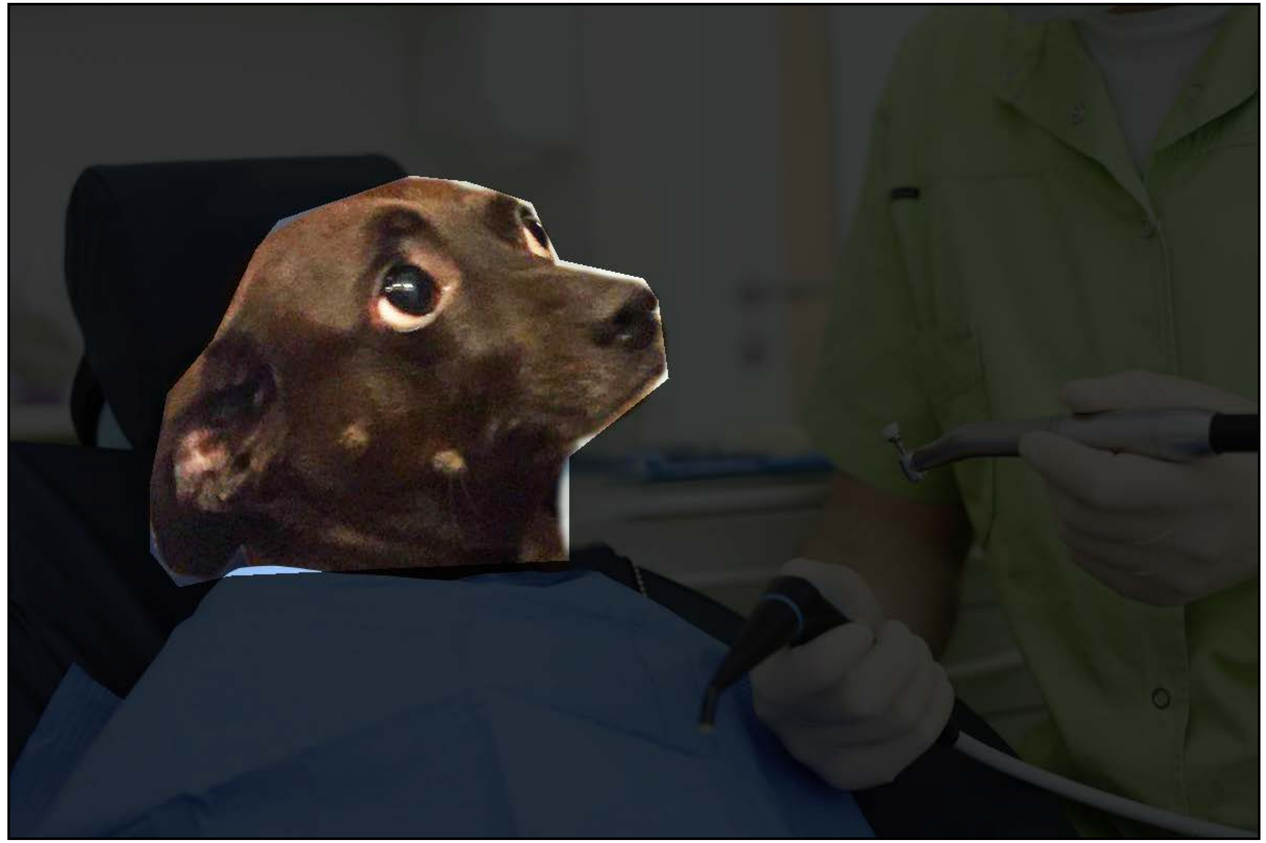


Manipulated Region

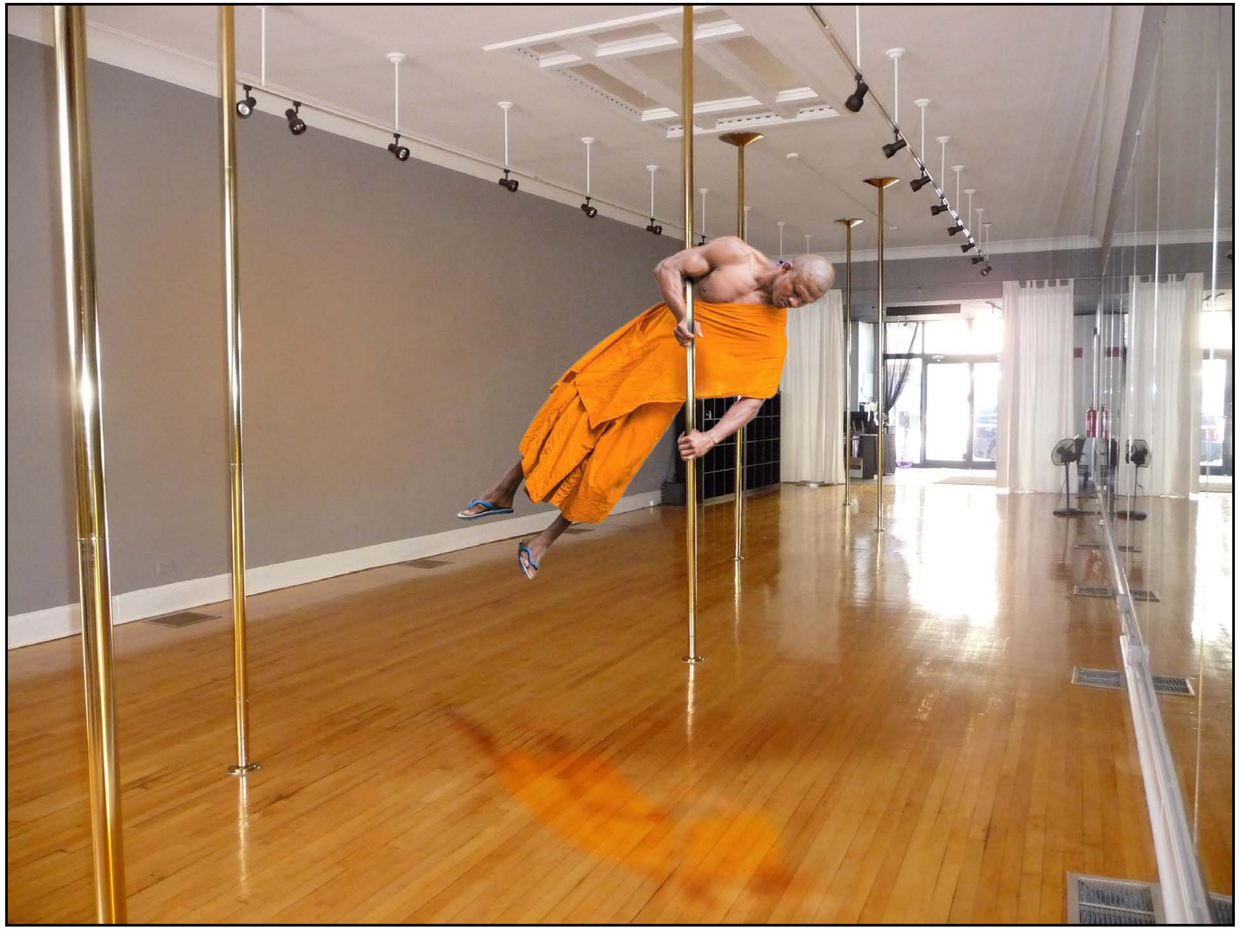
Our detection



Ground Truth



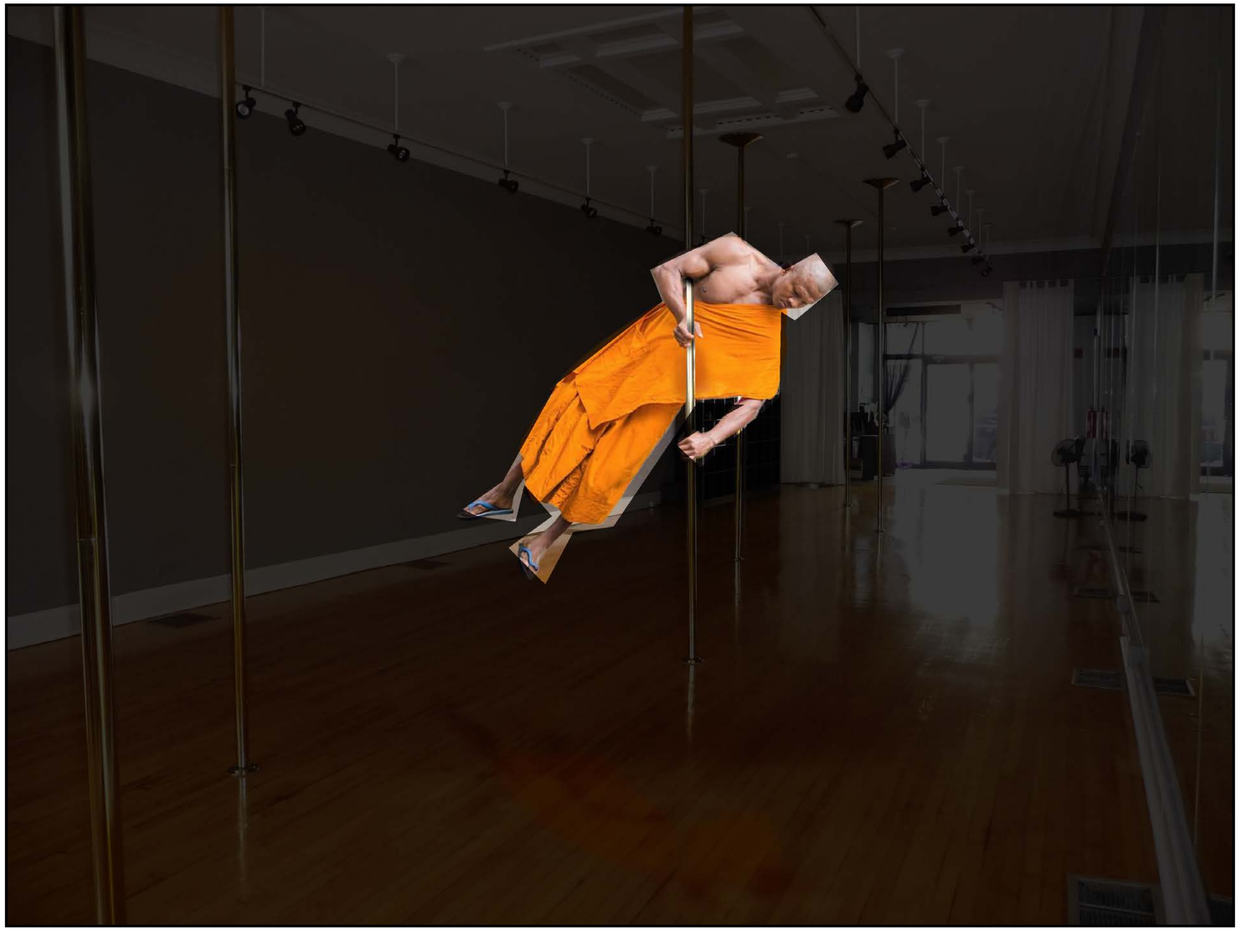
Input Image



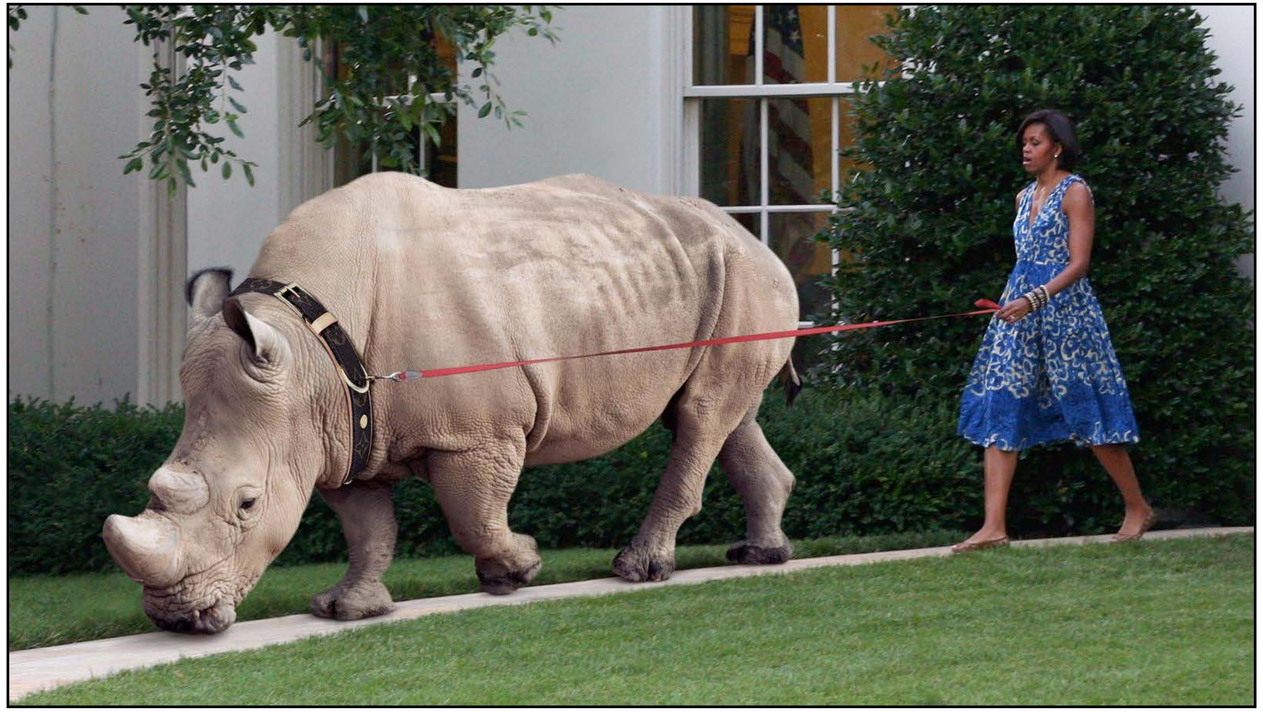
Manipulated Region

Our detection

Ground Truth

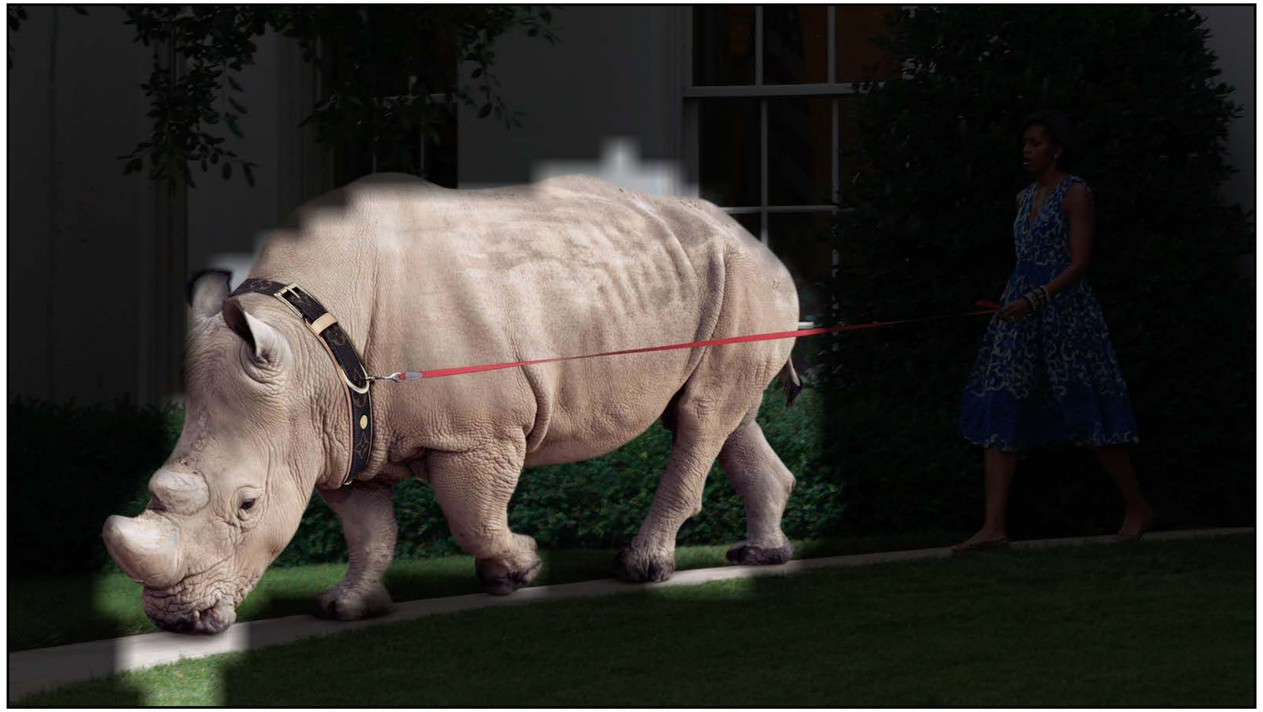


Input Image

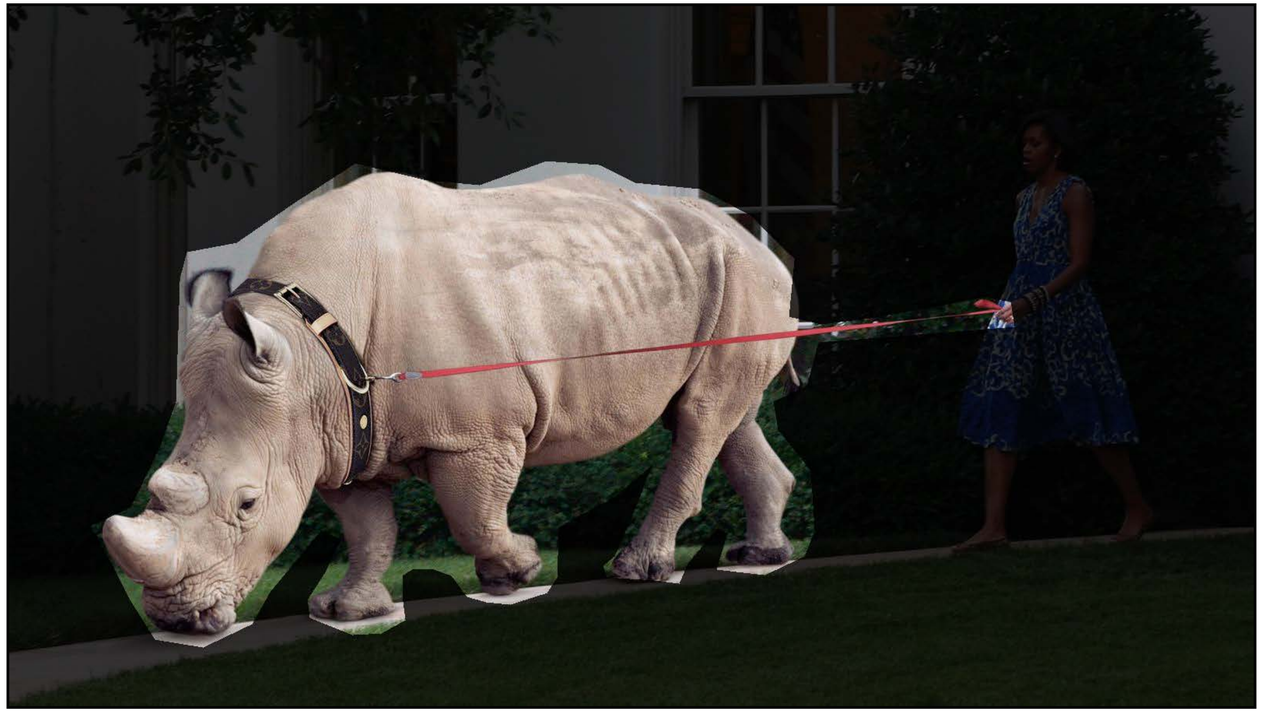


Manipulated Region

Our detection



Ground Truth



Input Image

Hays, Efros SIGGRAPH 2007

Manipulated Region

Our detection

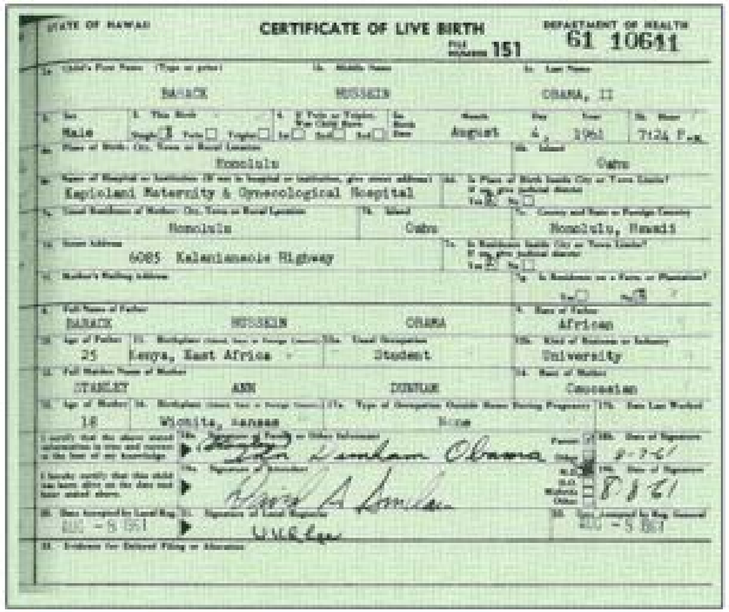
Ground Truth

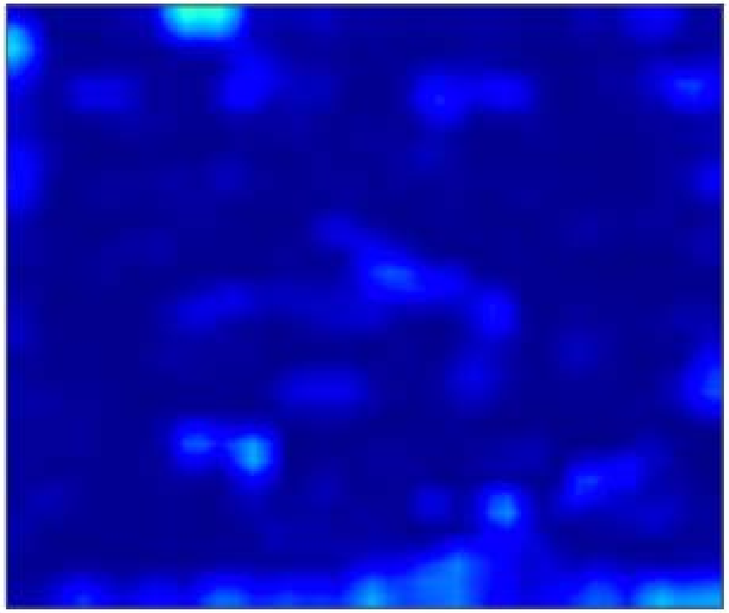
Hays, Efros SIGGRAPH 2007

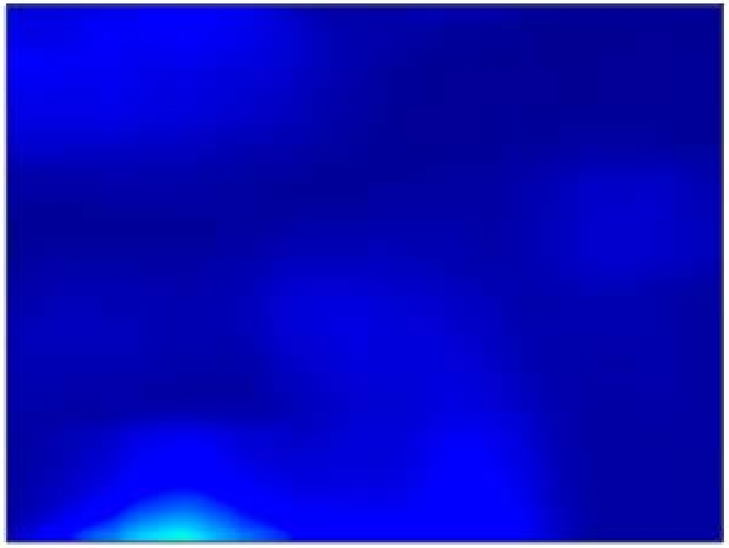
Authentic Images

Image

Consistency Map







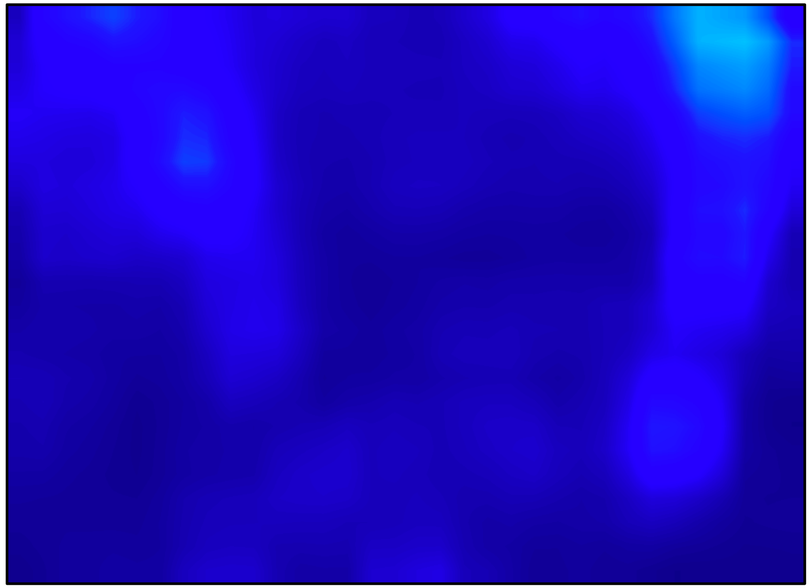
Splice Localization

Failures

Image

Ground Truth

Consistency Map



Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

ECCV 2018

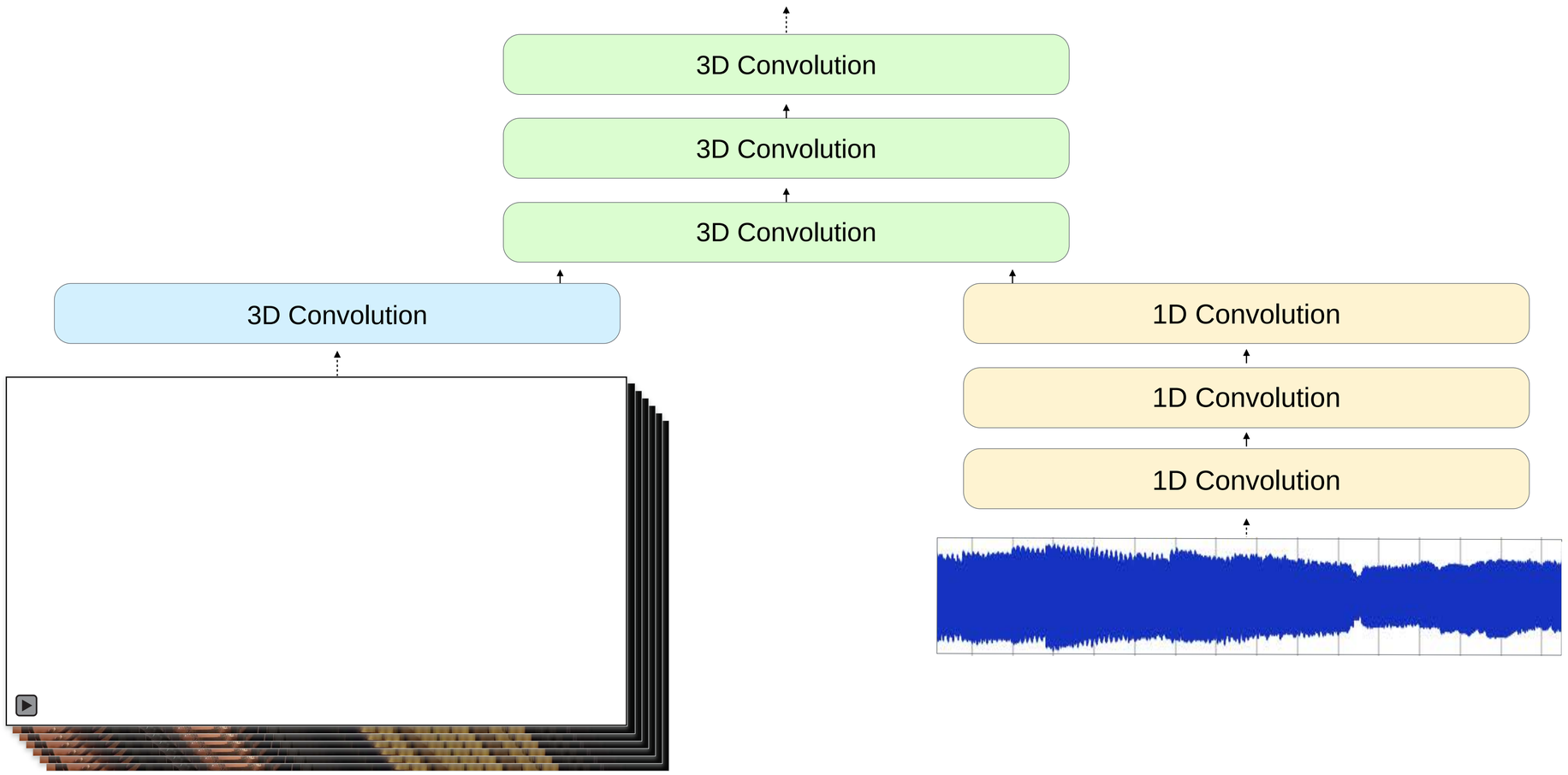
Andrew Owens Alexei A. Efros

UC Berkeley

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

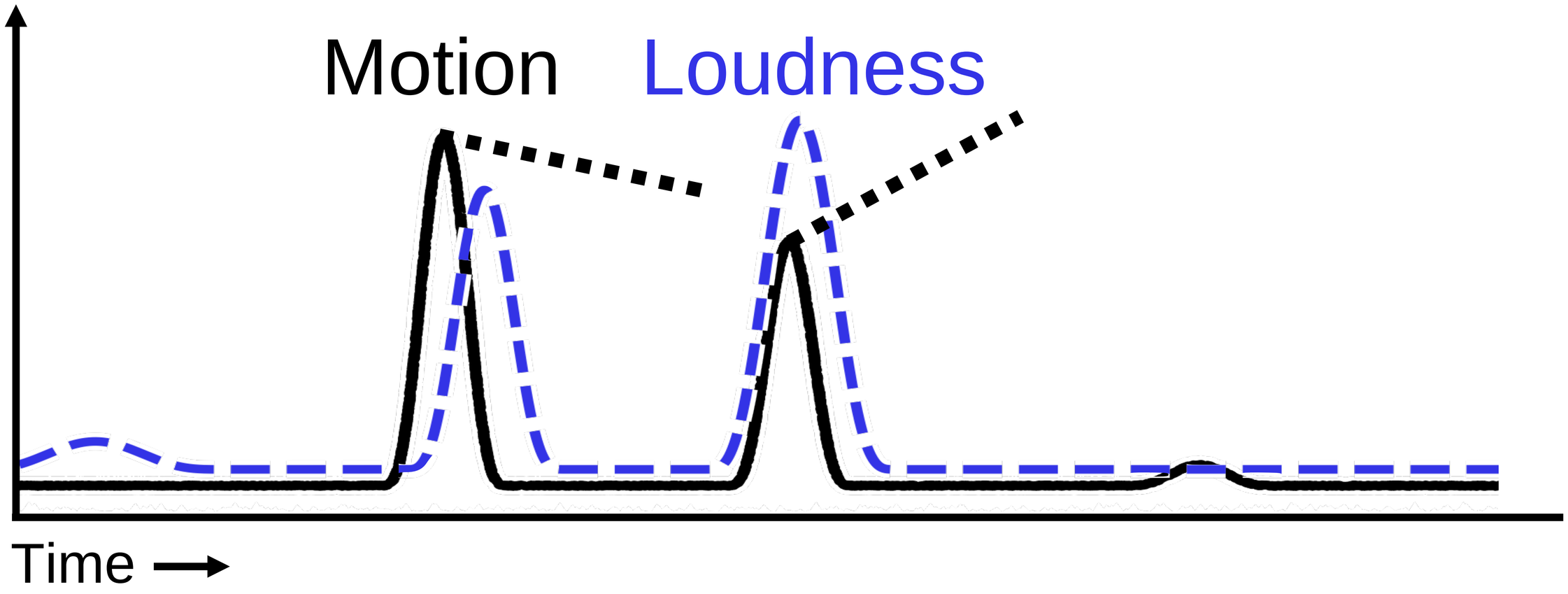
Multisensory representation



Idea #1: random pairs

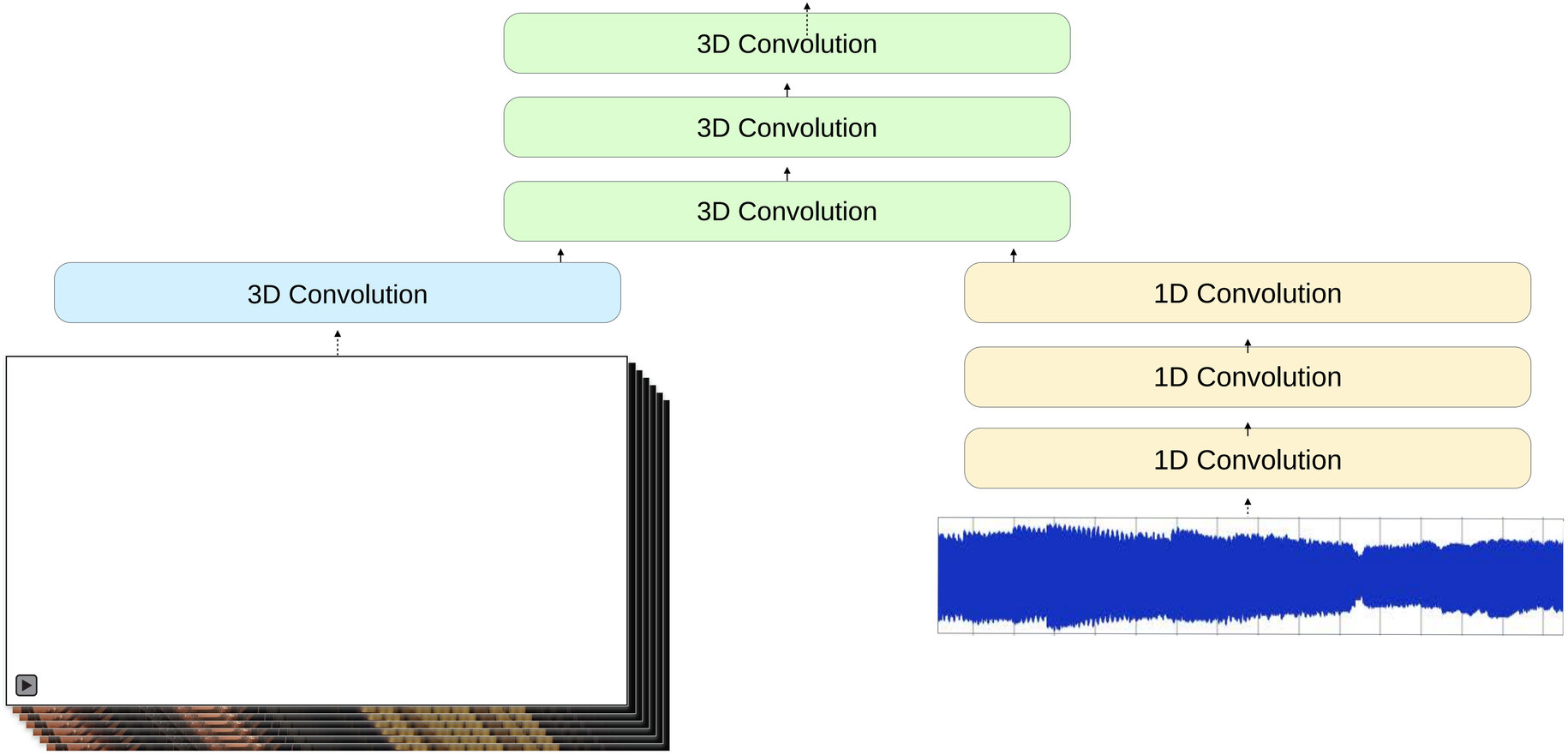
Idea #2: time-shifted pairs

Idea #2: time-shifted pairs

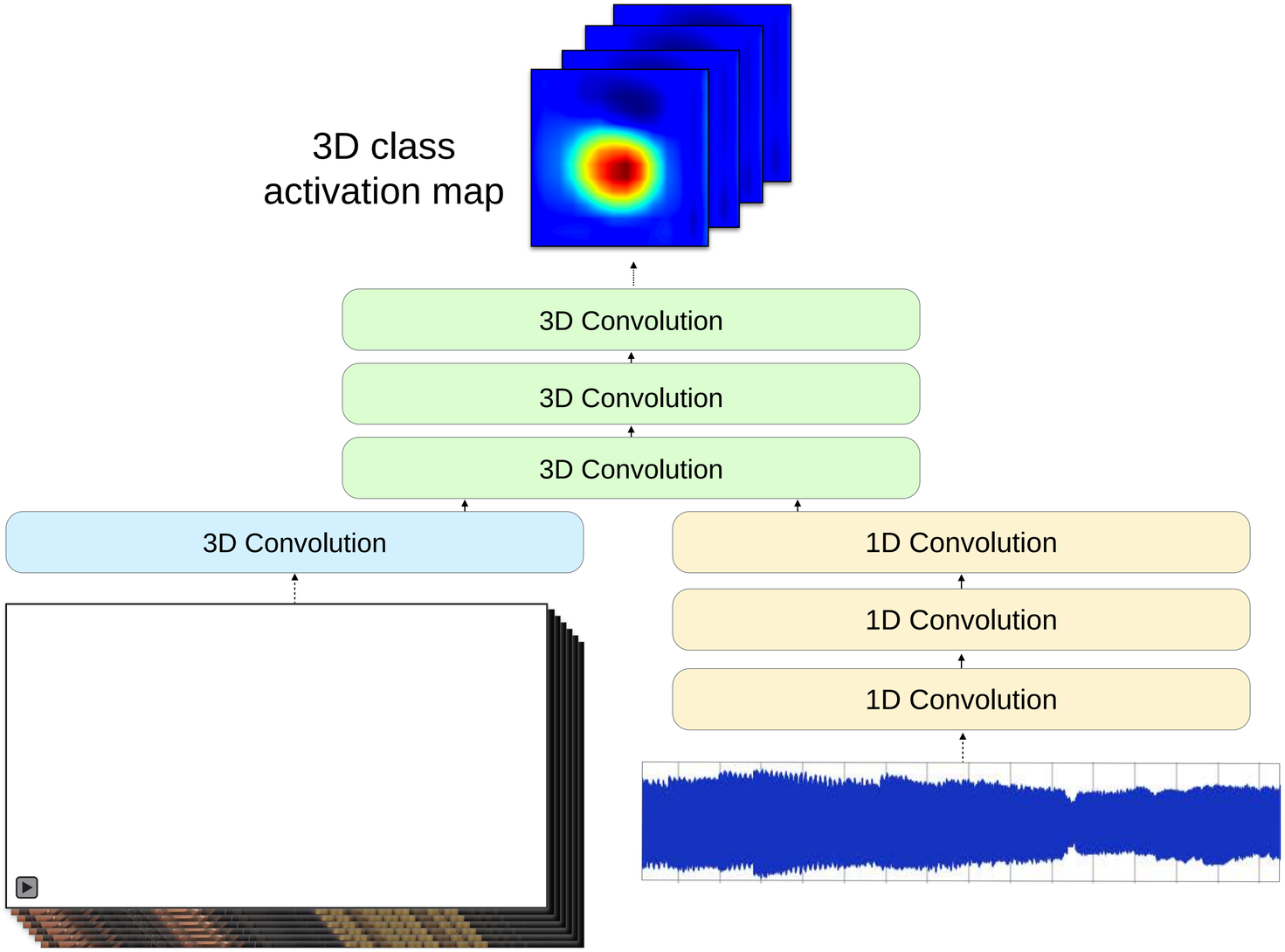


Self-supervised Training

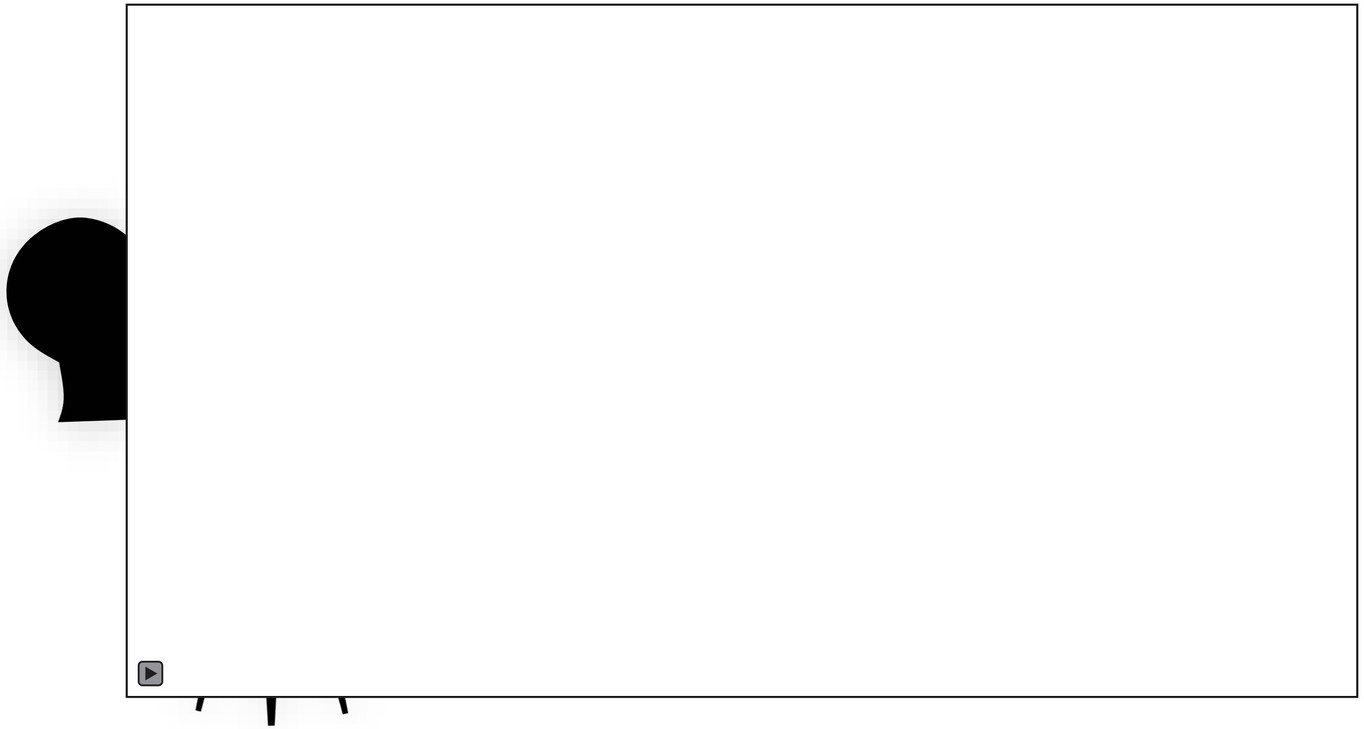
aligned vs. not-aligned



Visualizing the location of sound sources



On/off-screen source separation



On/off-screen source separation

Meta-supervision

Meta-supervision

$F(x) = y$

- direct supervision

$F(x) \in Y$

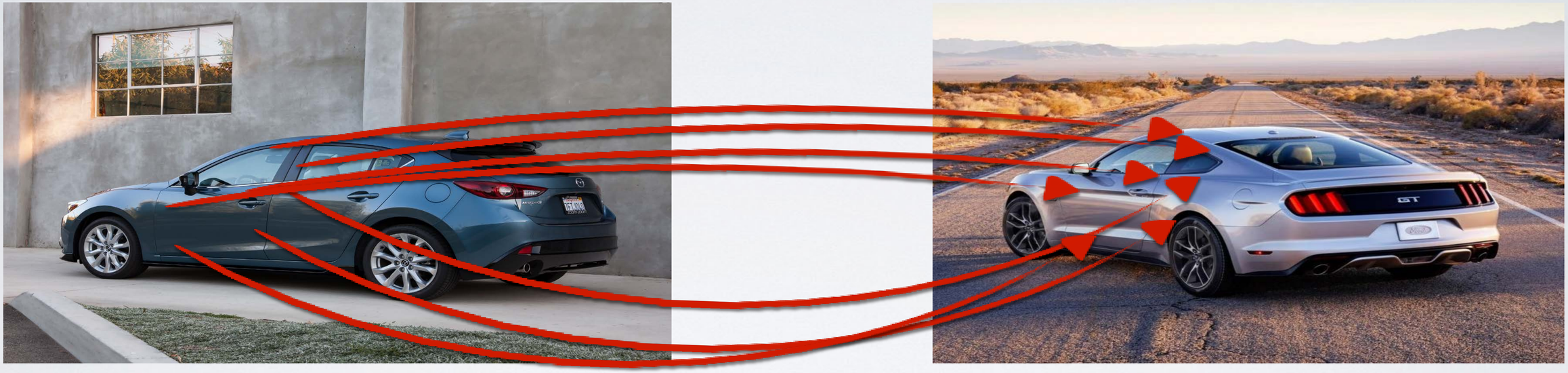
- GANs

$G(F(x)) = x$

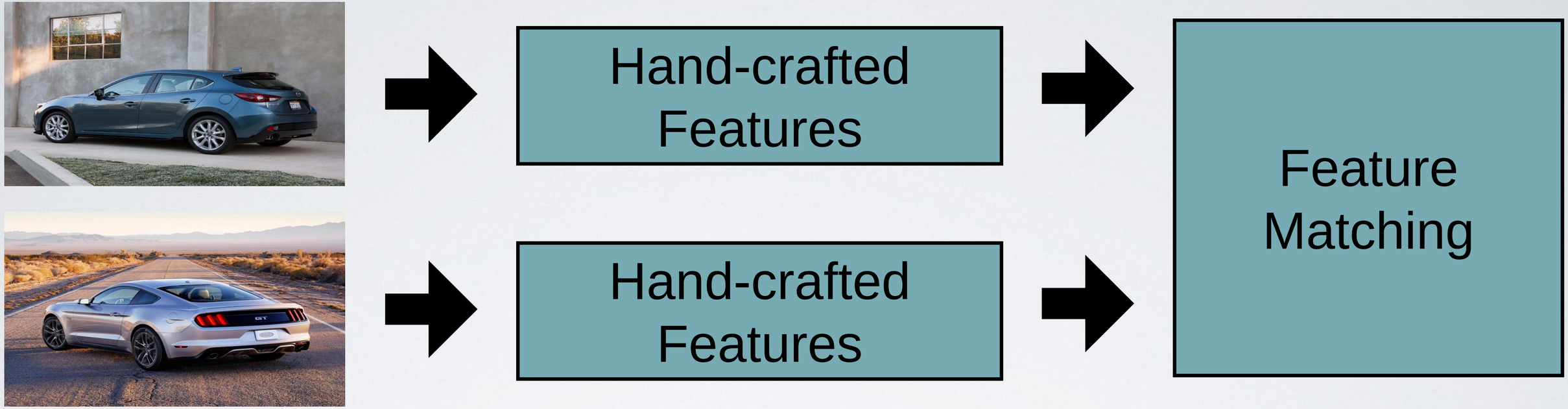
- cycle-consistency

- …

Dense Semantic Correspondence



Traditional Pairwise Methods



- SIFT flow: Liu *et al.*, ECCV 2008
- Generalized PatchMatch: Barnes *et al.*, ECCV 2010
- Deformable Spatial Pyramid: Kim *et al.*, CVPR 2013

Collection Correspondence

- Congealing: Learned-Miller, PAMI 2006
- Collection Flow: Kramelmacher-Shlizerman *et al.*, CVPR 2012
- Object discovery and segmentation: Rubinstein *et al.*, CVPR 2013
- Compositional Image Model: Mobahi *et al.*, CVPR 2014
- Object discovery and localization: Cho *et al.*, CVPR 2015
- FlowWeb: T. Zhou *et al.*, CVPR 2015
- Multi-image Matching: X. Zhou *et al.*, ICCV 2015

Labels for CNN Training?

Learning Dense Correspondence via 3D-guided Cycle Consistency

Tinghui Zhou[1], Philipp Krähenbühl[1], Mathieu Aubry[2],
Qixing Huang[3], Alexei A. Efros[1]

UC Berkeley[1], ENPC ParisTech[2], TTI-Chicago[3]

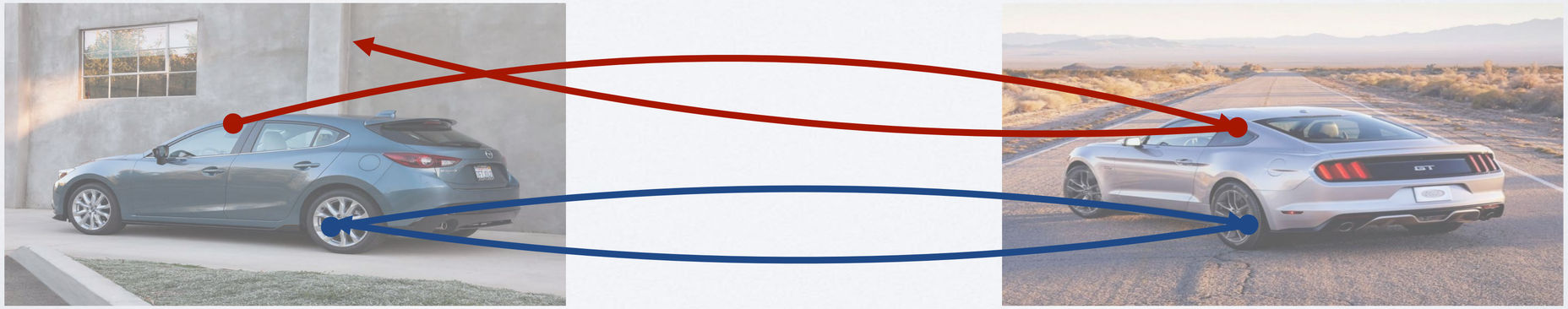
CVPR 2016

Cycle-consistency as Supervision

- Composite flows along a cycle should be zero

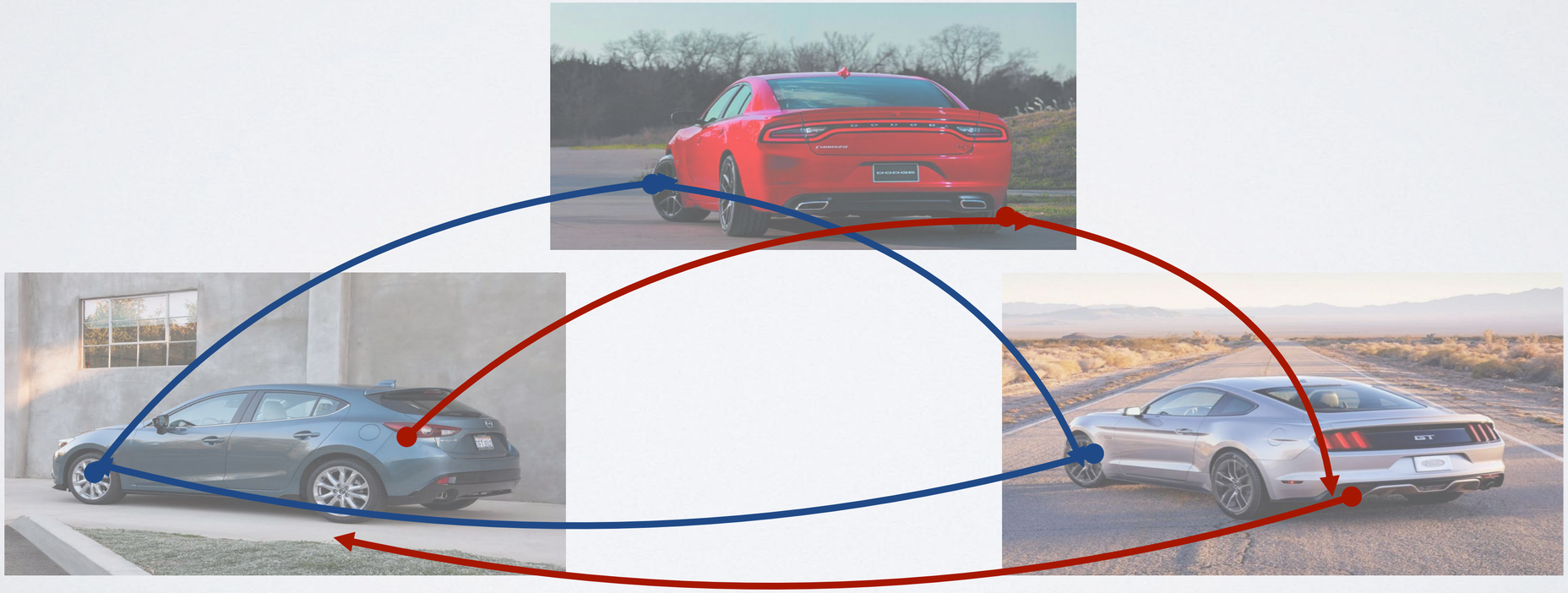
Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$



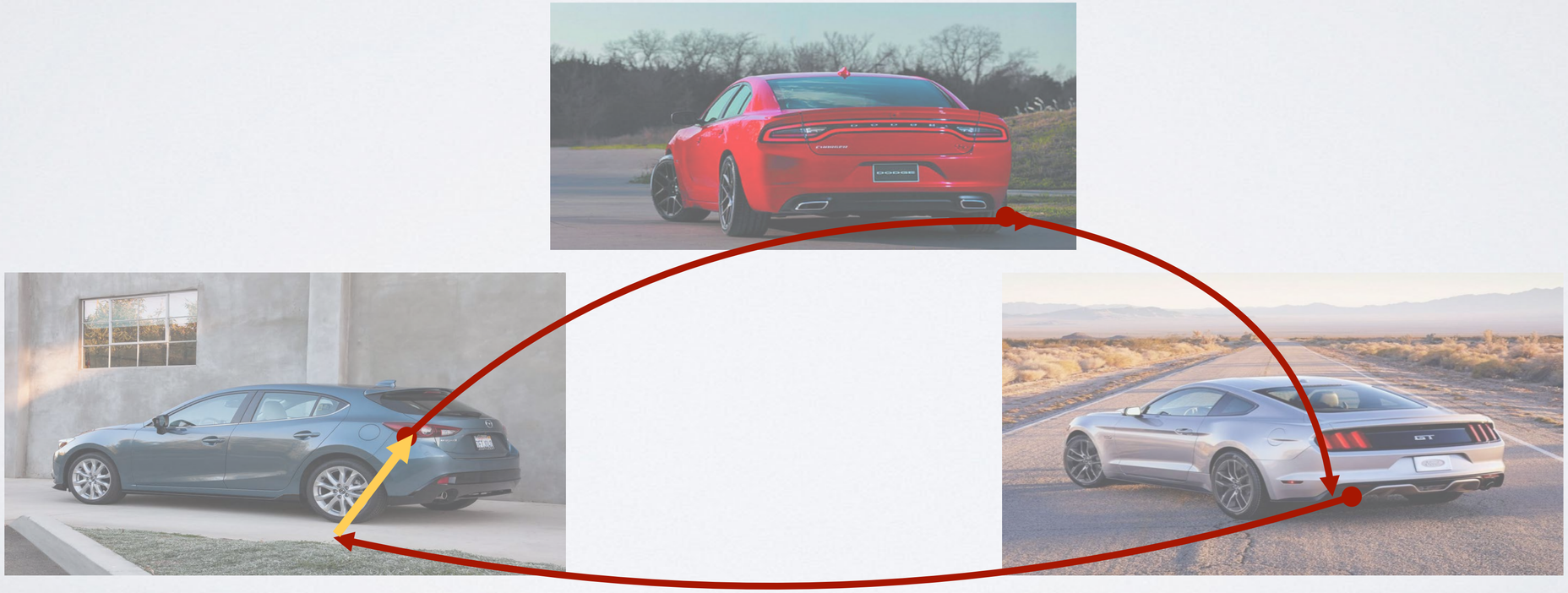
Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$



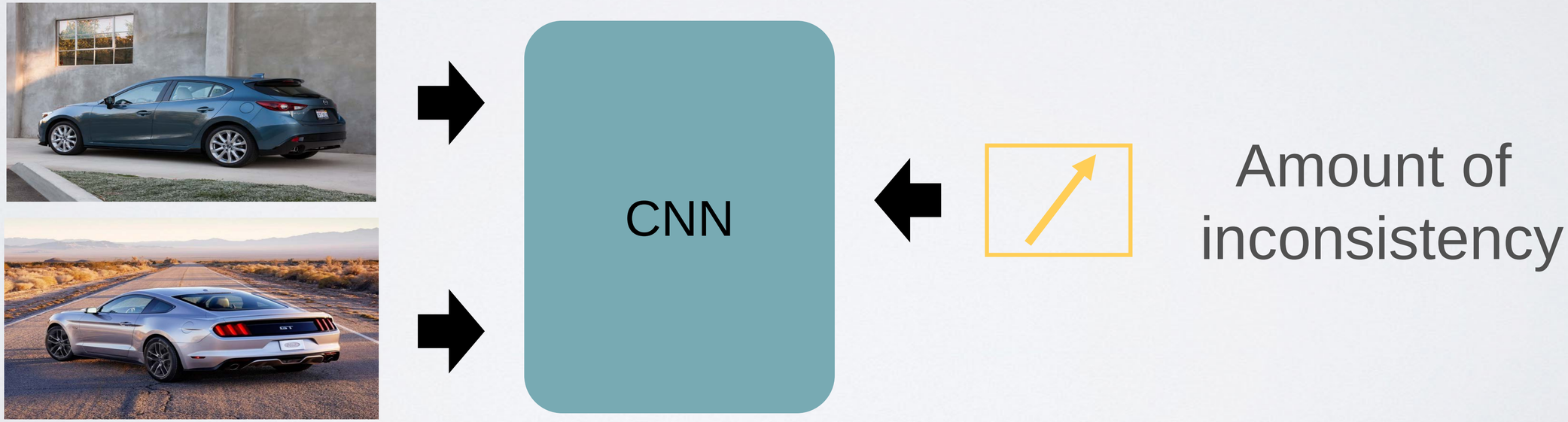
Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$

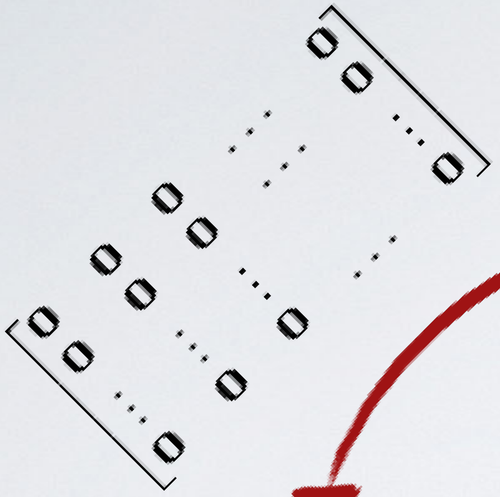


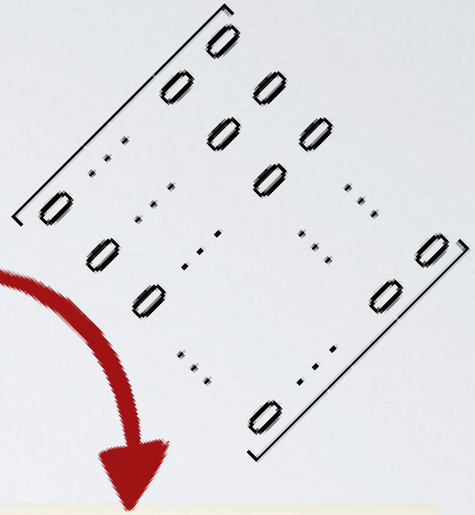
Cycle-consistency as Supervision

- Composite flows along a cycle should be zero
- 2-cycle consistency: $F_{i,j} \circ F_{j,i} = 0$
- 3-cycle consistency: $F_{i,k} \circ F_{k,j} \circ F_{j,i} = 0$



Could be consistent but wrong…

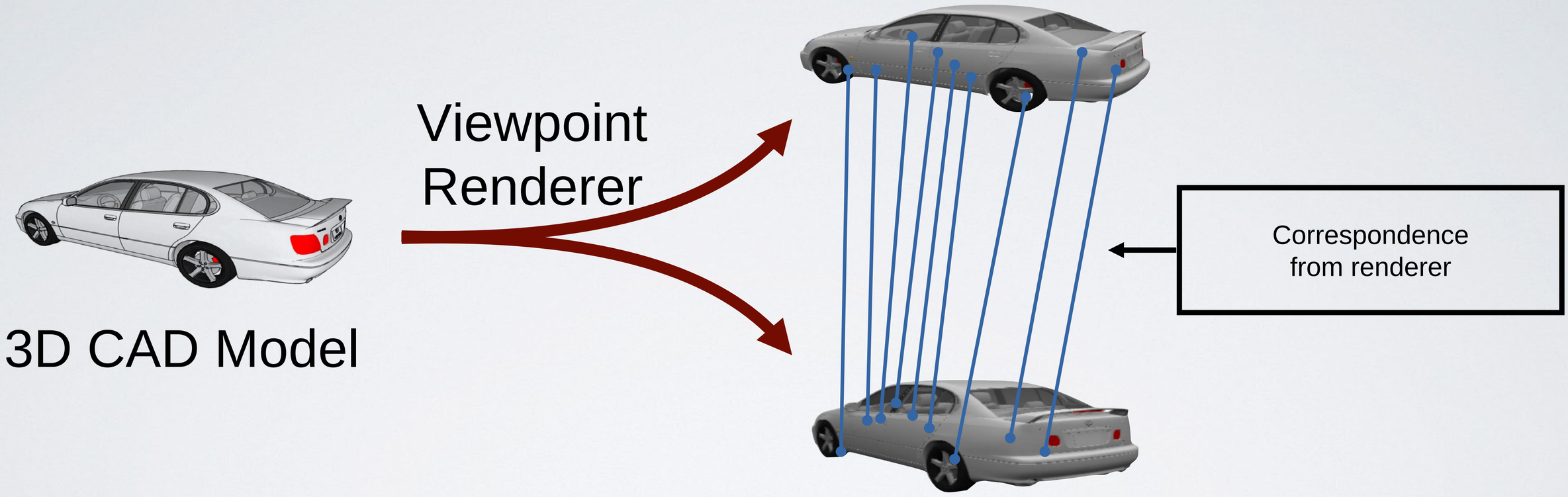




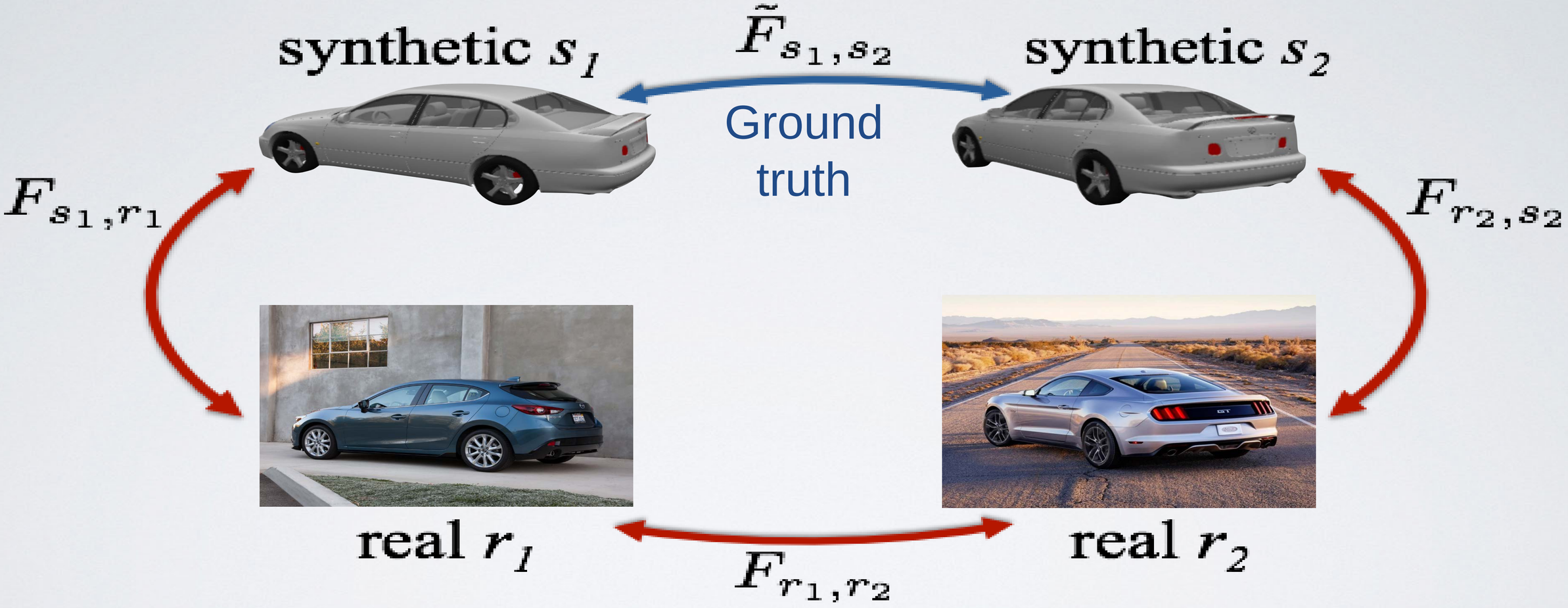
Need an anchor edge!

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Synthetic Correspondence as the Anchor

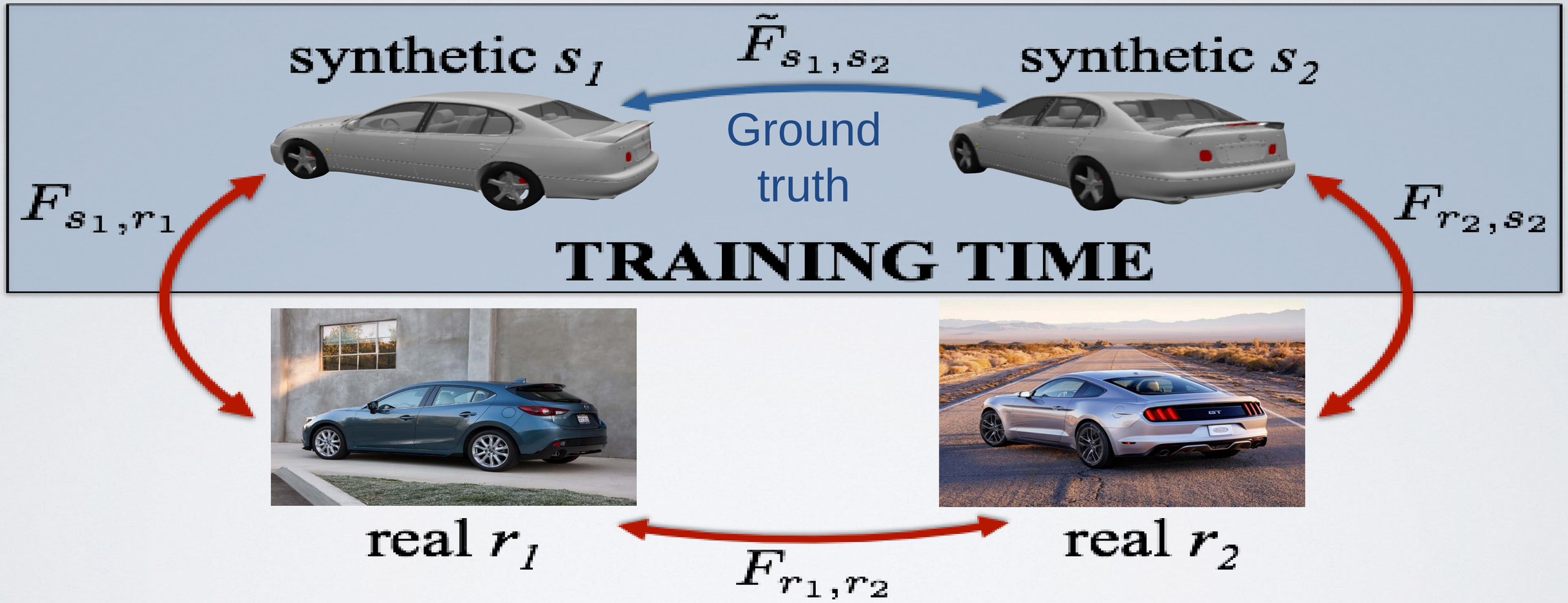


3D-guided Cycle Consistency



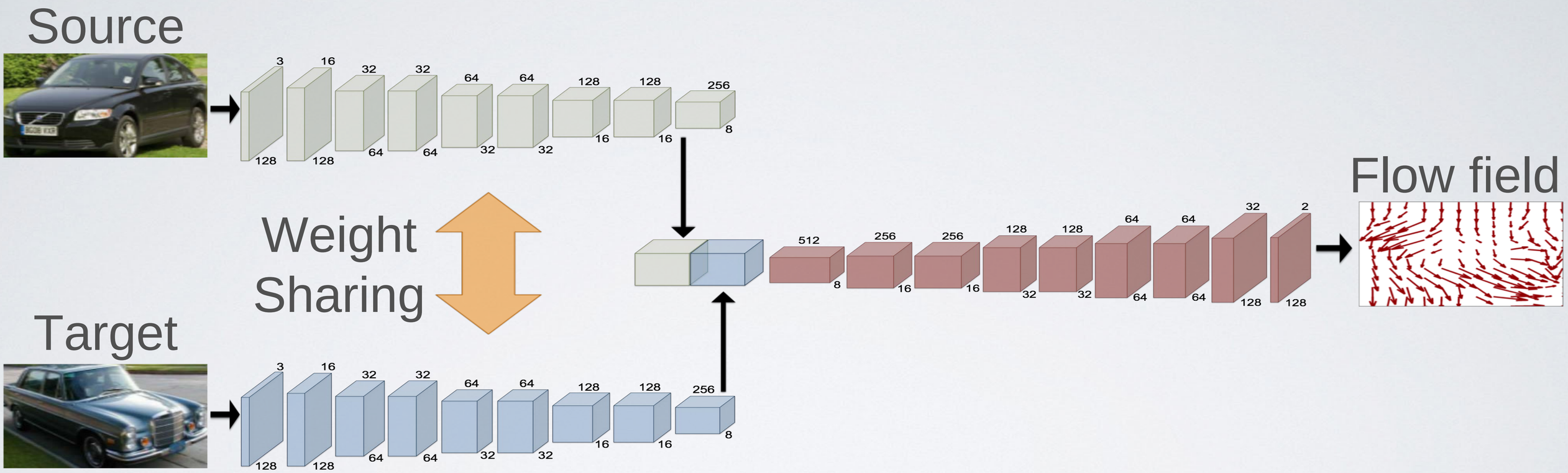
$$\tilde{F}_{s_1,s_2} = F_{s_1,r_1} \circ F_{r_1,r_2} \circ F_{r_2,s_2}$$

3D-guided Cycle Consistency



$$\min \sum_{<s_1,s_2,r_1,r_2>} \mathcal{L}\left(\tilde{F}_{s_1,s_2} - F_{s_1,r_1} \circ F_{r_1,r_2} \circ F_{r_2,s_2}\right)$$

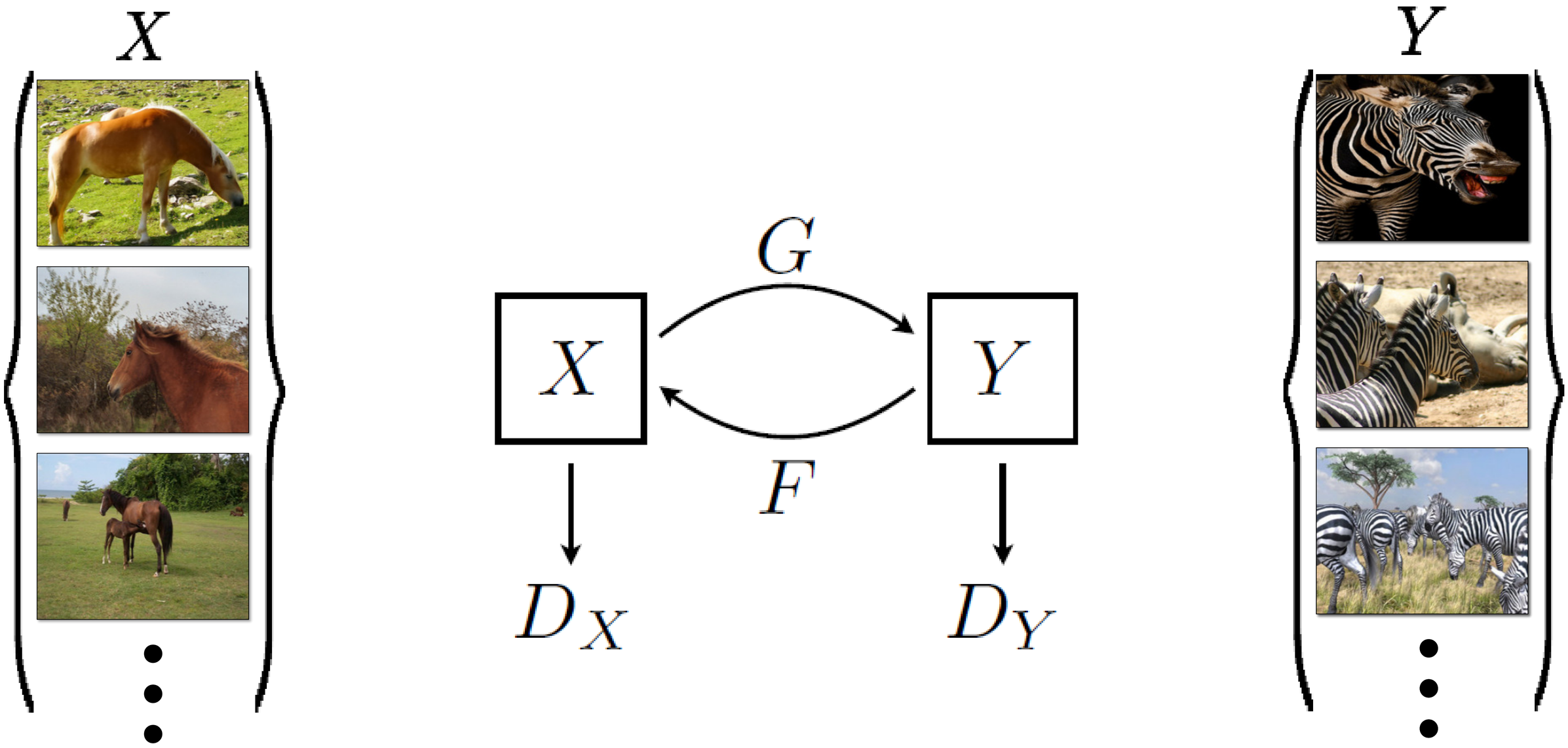
Network Architecture



OUR RESULT

SIFTFLOW

CycleGAN, or “there and back aGAN”



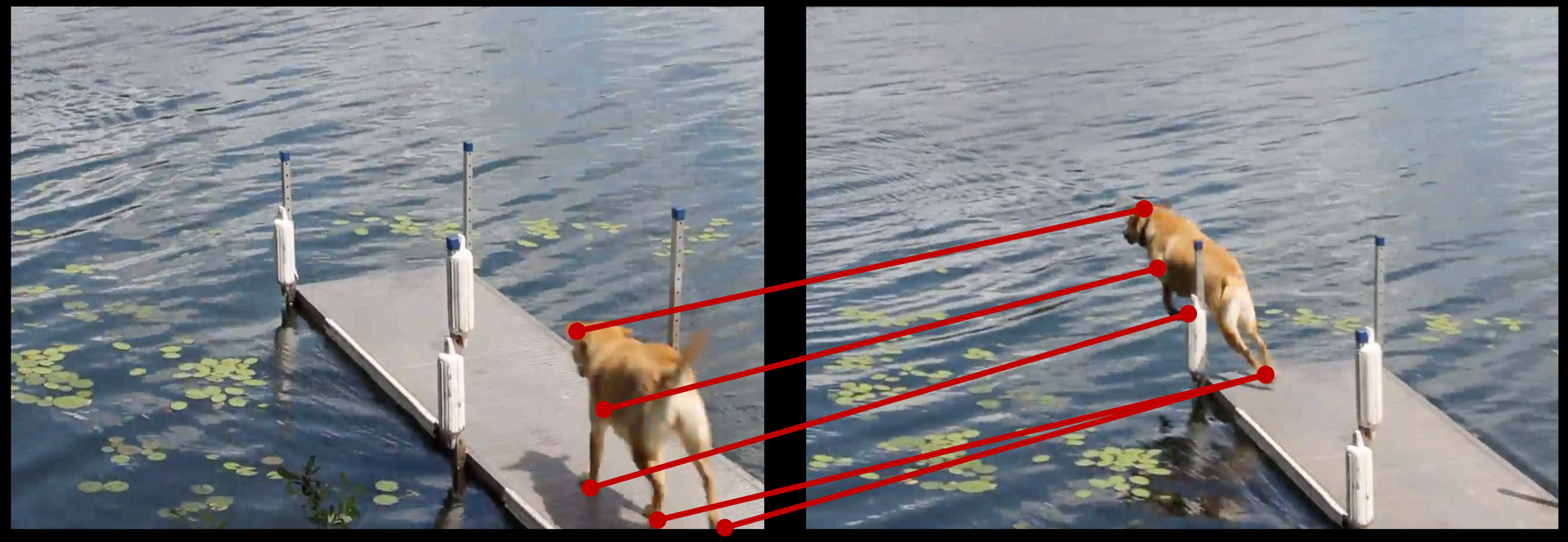
[Zhu*, Park*, Isola, Efros. ICCV 2017]

CVPR 2019

Learning Correspondence from the Cycle-consistency of Time

Xiaolong Wang Allan Jabri Alexei Efros
CMU UC Berkeley

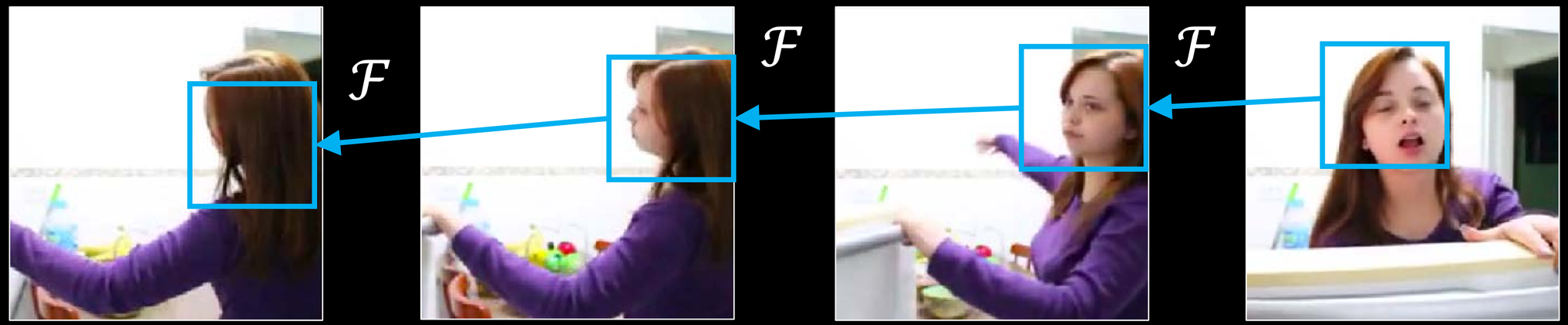
Goal: Learn Correspondence without Human Supervision



The visual world exhibits continuity

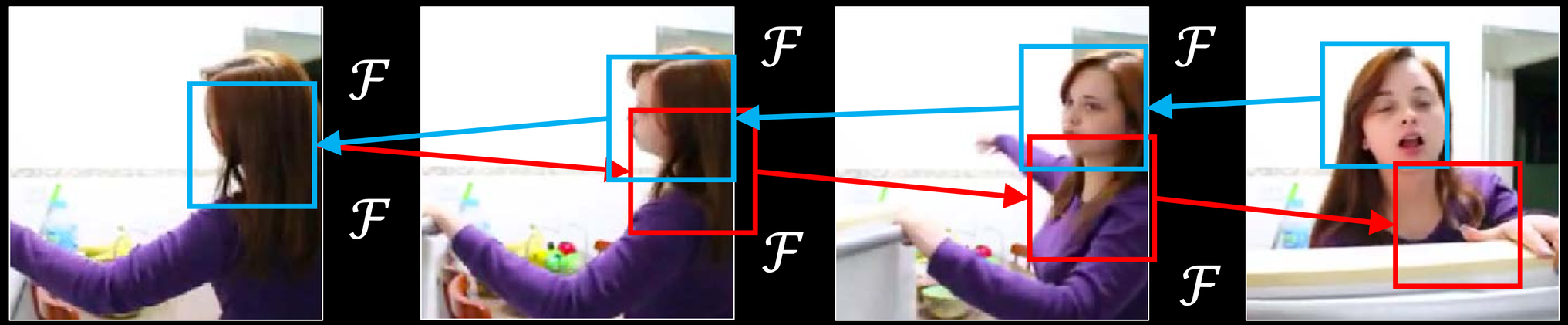
Learning to Track

$\mathcal{F}$: a deep tracker



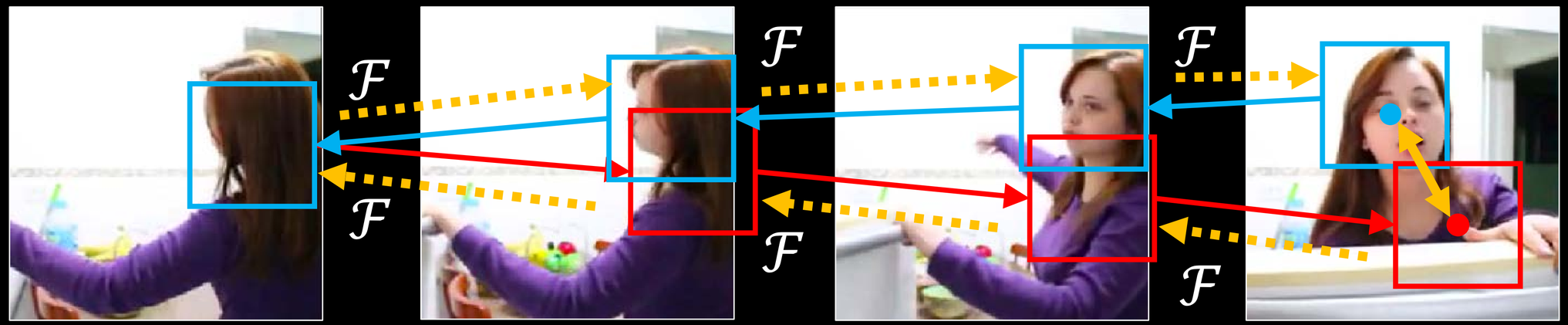
Supervision: Cycle-Consistency in Time

Track backwards



Track forwards, back to the future

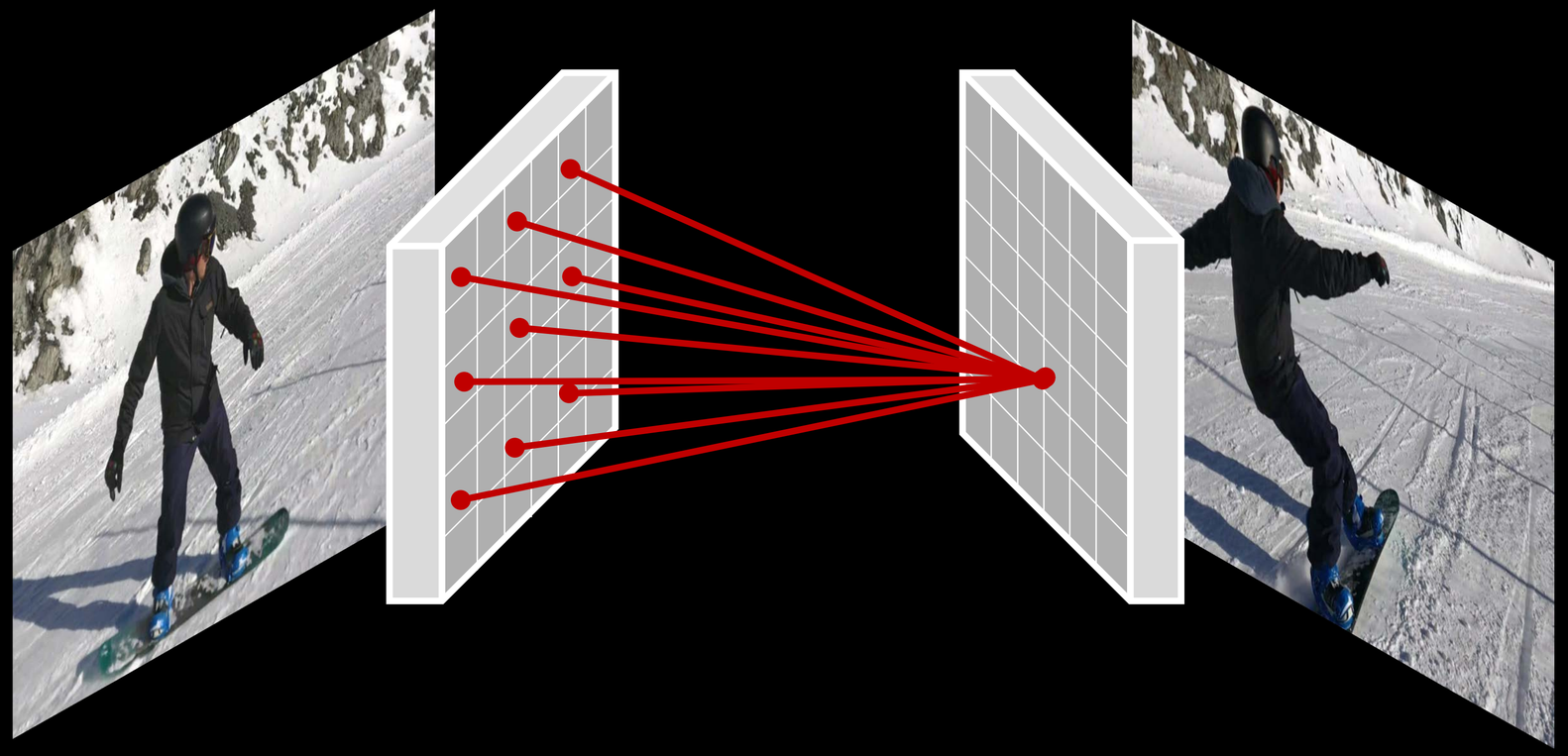
Supervision: Cycle-Consistency in Time



Backpropagation through time along the cycle

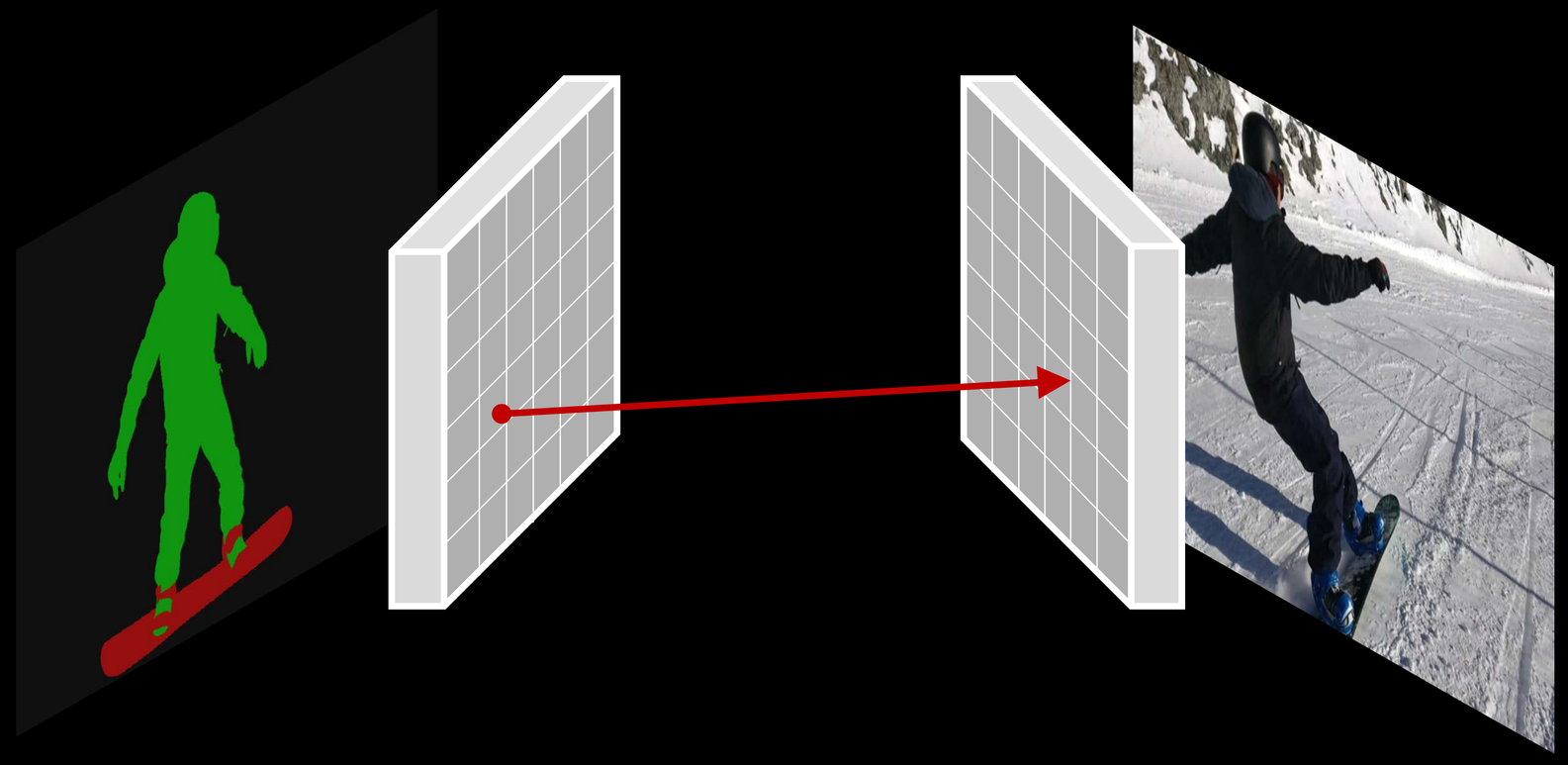
Visualization of Training

Test Time: Nearest Neighbors in Feature Space ϕ



$t-1$ t

Test Time: Nearest Neighbors in Feature Space ϕ



$t-1$ t

Instance Mask Tracking

DAVIS Dataset

DAVIS Dataset: Pont-Tuset et al. *The 2017 DAVIS Challenge on Video Object Segmentation*. 2017.

Pose Keypoint Tracking

JHMDB Dataset

Comparison

Our Correspondence

Optical Flow

Pose Keypoint Tracking

JHMDB Dataset

Method	PCK @.1
Optical Flow	45%
Vondrick et al.	45%
Ours	58%

Vondrick et al. Tracking Emerges by Colorizing Videos. ECCV 2018.

Texture Tracking

DAVIS Dataset

DAVIS Dataset: Pont-Tuset et al. *The 2017 DAVIS Challenge on Video Object Segmentation*. 2017.

Semantic Masks Tracking

Video Instance Parsing Dataset

Zhou et al. *Adaptive Temporal Encoding Network for Video Instance-level Human Parsing*. ACM MM 2018.