# “The Revolution Will Not Be Supervised!”
## (almost) 10 Years Later: A Personal Journey

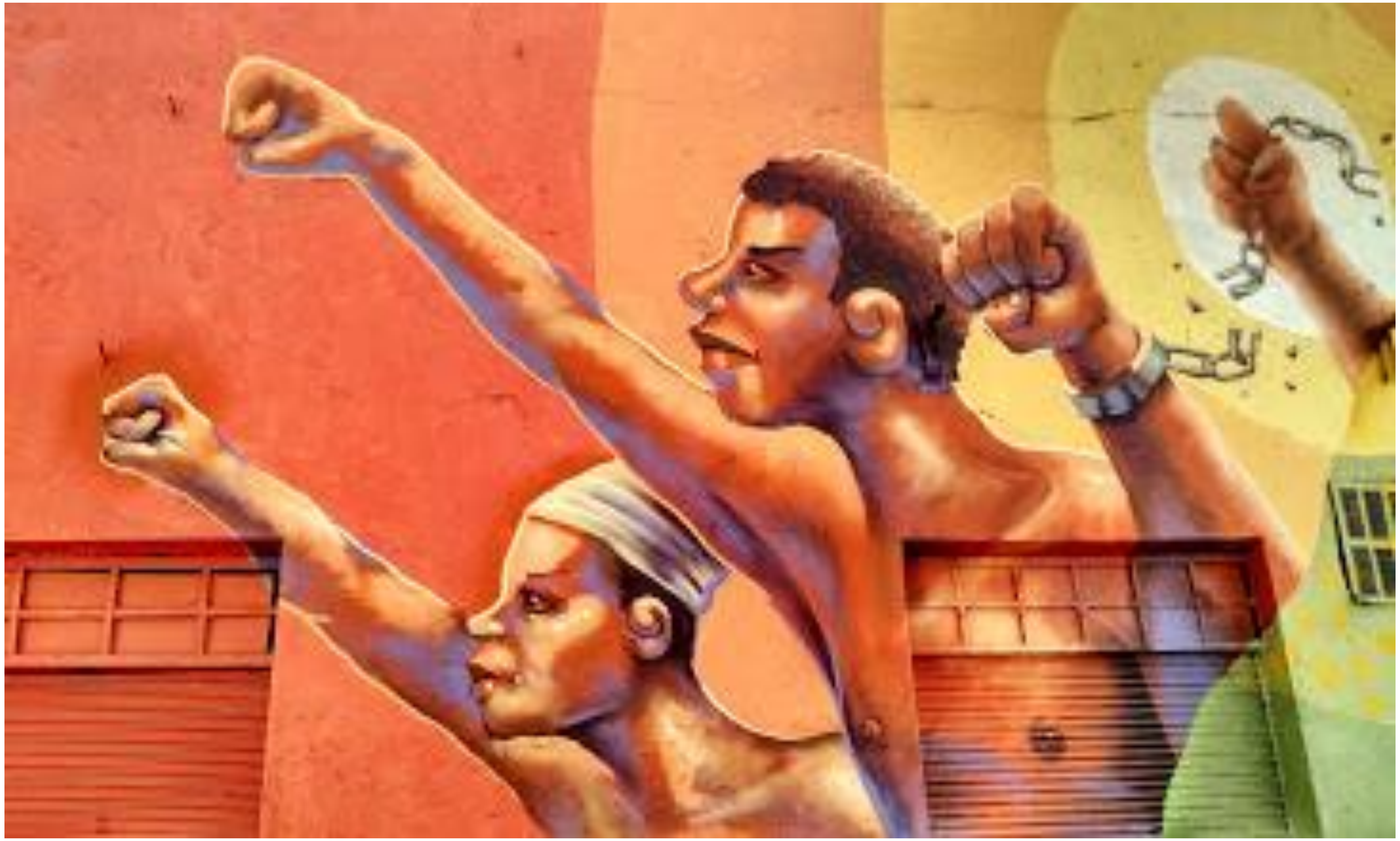


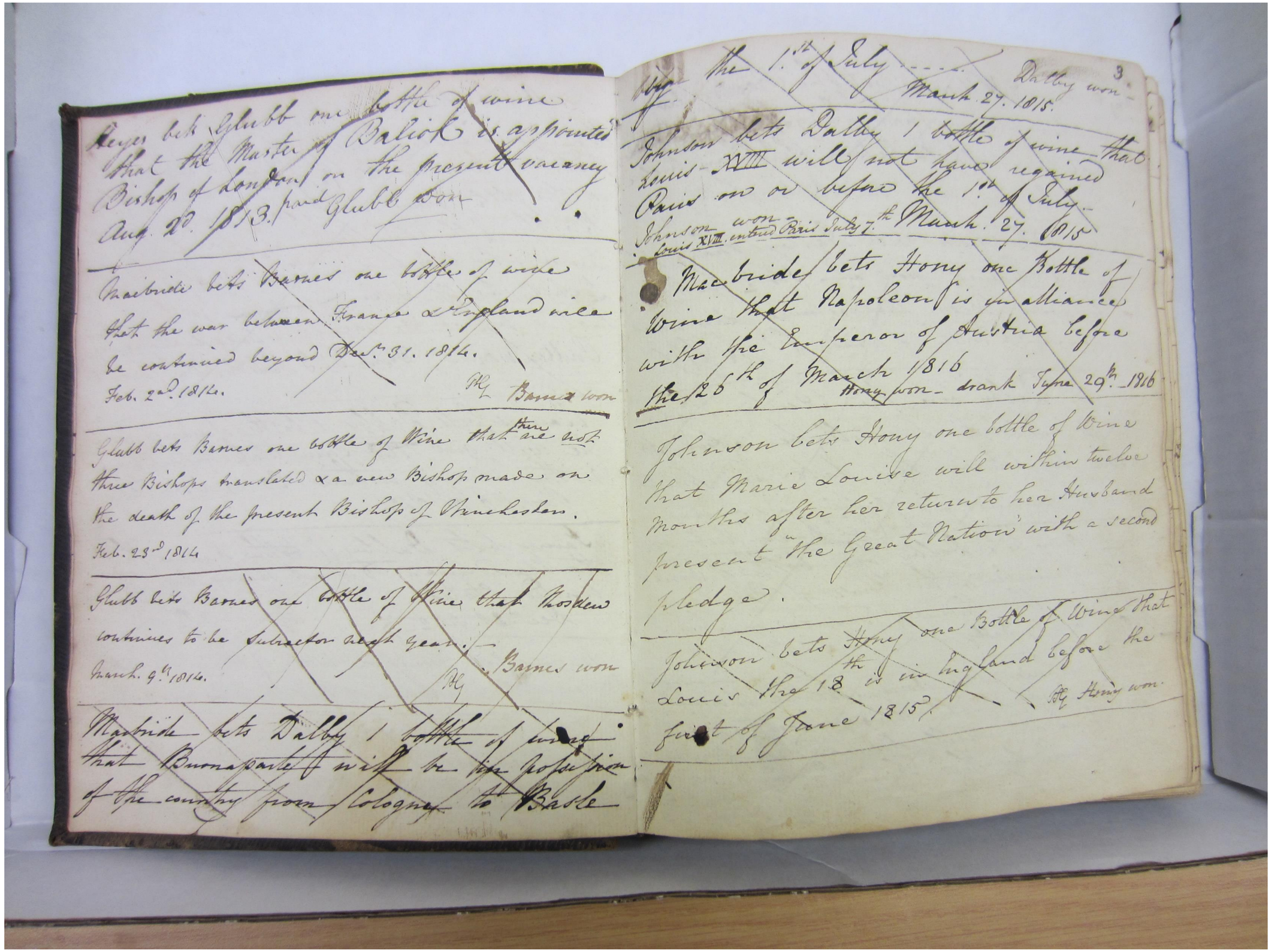
Photo from Santiago, ICCV 2015

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Alexei A. Efros
UC Berkeley

It all started with a bet…

# Long Tradition of Scientific Bets



Keyes bets Glubb one bottle of wine
that the Master of Baliol is appointed
Bishop of London on the present vacancy
Aug. 2d. 1813. paid Glubb won

Macbride bets Barnes one bottle of wine
that the war between France & England will
be continued beyond Decr. 31. 1814.
Feb. 2nd. 1814. HG Barnes won

Glubb bets Barnes one bottle of Wine that there are not
three Bishops translated & a new Bishop made on
the death of the present Bishop of Winchester.
Feb. 23rd 1814

Glubb bets Barnes one bottle of Wine that Rosdew
continues to be Subrector next year.
March. 9th 1814. HG Barnes won

Macbride bets Dalby 1 bottle of wine
that Buonaparte will be in possession
of the country from Cologne to Basle

by the 1st of July ....
March 27. 1815. Dalby won 3

Johnson bets Dalby 1 bottle of wine that
Louis XVIII will not have regained
Paris on or before the 1st of July.
Johnson won Louis XVIII. entered Paris July 7th March. 27. 1815

Macbride bets Hony one Bottle of
Wine that Napoleon is in alliance
with the Emperor of Austria before
the 26th of March 1816
Hony won drank June 29th 1816

Johnson bets Hony one bottle of Wine
that Marie Louise will within twelve
Months after her return to her Husband
present "the Great Nation" with a second
pledge.

Johnson bets Hony one Bottle of Wine that
Louis the 18th is in England before the
first of June 1815. HG Hony won.

Betting Book
Exeter College
Oxford
1815

# The Gelato Bet (2014)

- R-CNN just came out
  - first big success of “Pre-train + Fine-tune” paradigm
- It was surprising (to me), that ImageNet pretraining helped in PASCAL detection
  - Label sets were so different!
- Was it the labels, or just the extra visual data?

# The Gelato Bet

Sept 23, 2014

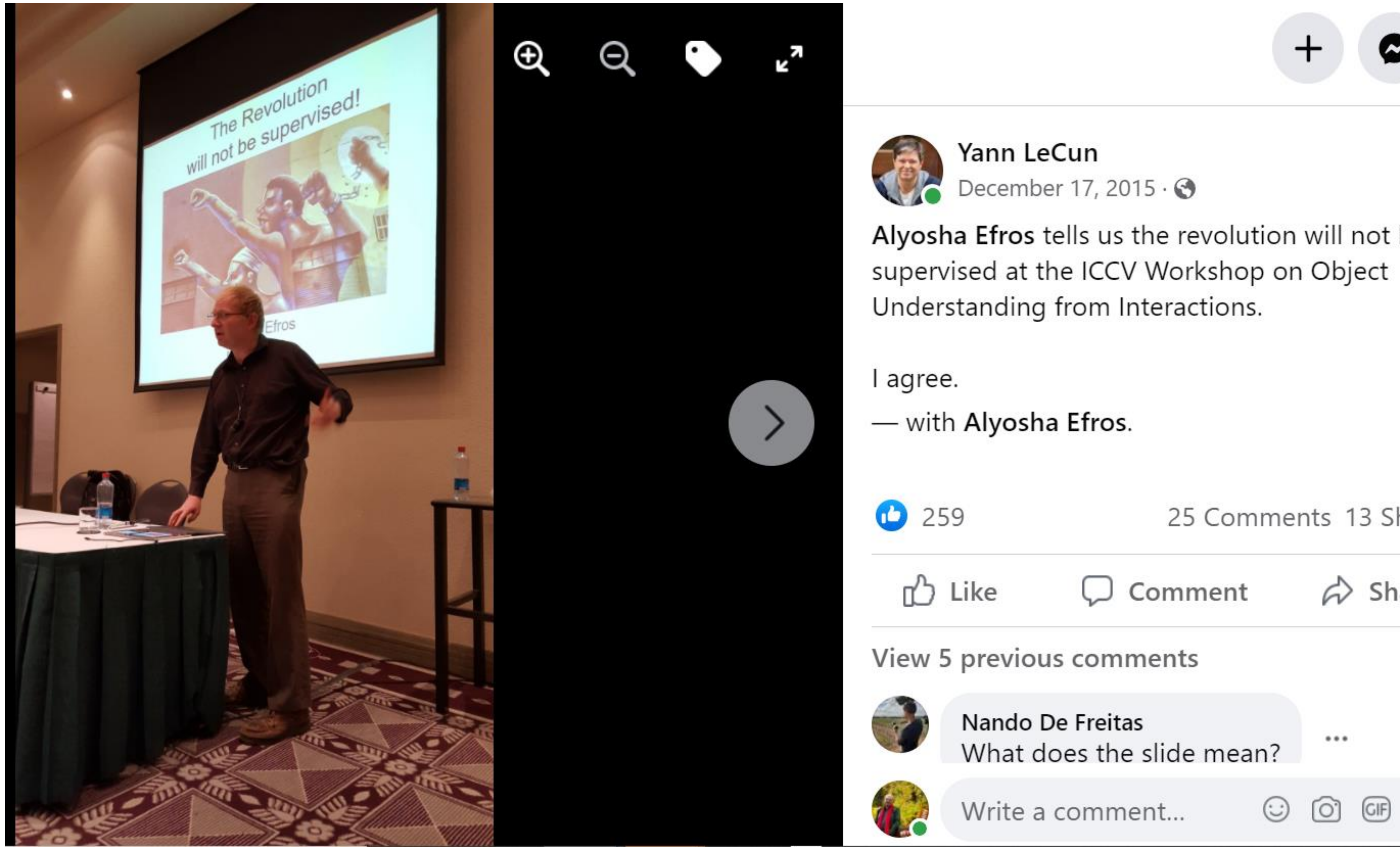
*"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, **without the use of any extra, human annotations** (e.g. ImageNet labels) as pre-training, Mr. Malik promises to buy Mr. Efros <u>one (1) gelato (two scoops: one chocolate, one vanilla).</u>"*

# One year later…

- Of course, I lost the bet:

# One year later…

- …but:
  - First 4 self-supervised papers presented at **ICCV 2015**
    - Doersch et al, Agrawal et al, Wang et al, and Jayaraman et al.
  - Yann LeCun liked my talk and posted on FB
  - The rest is history…



# 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?

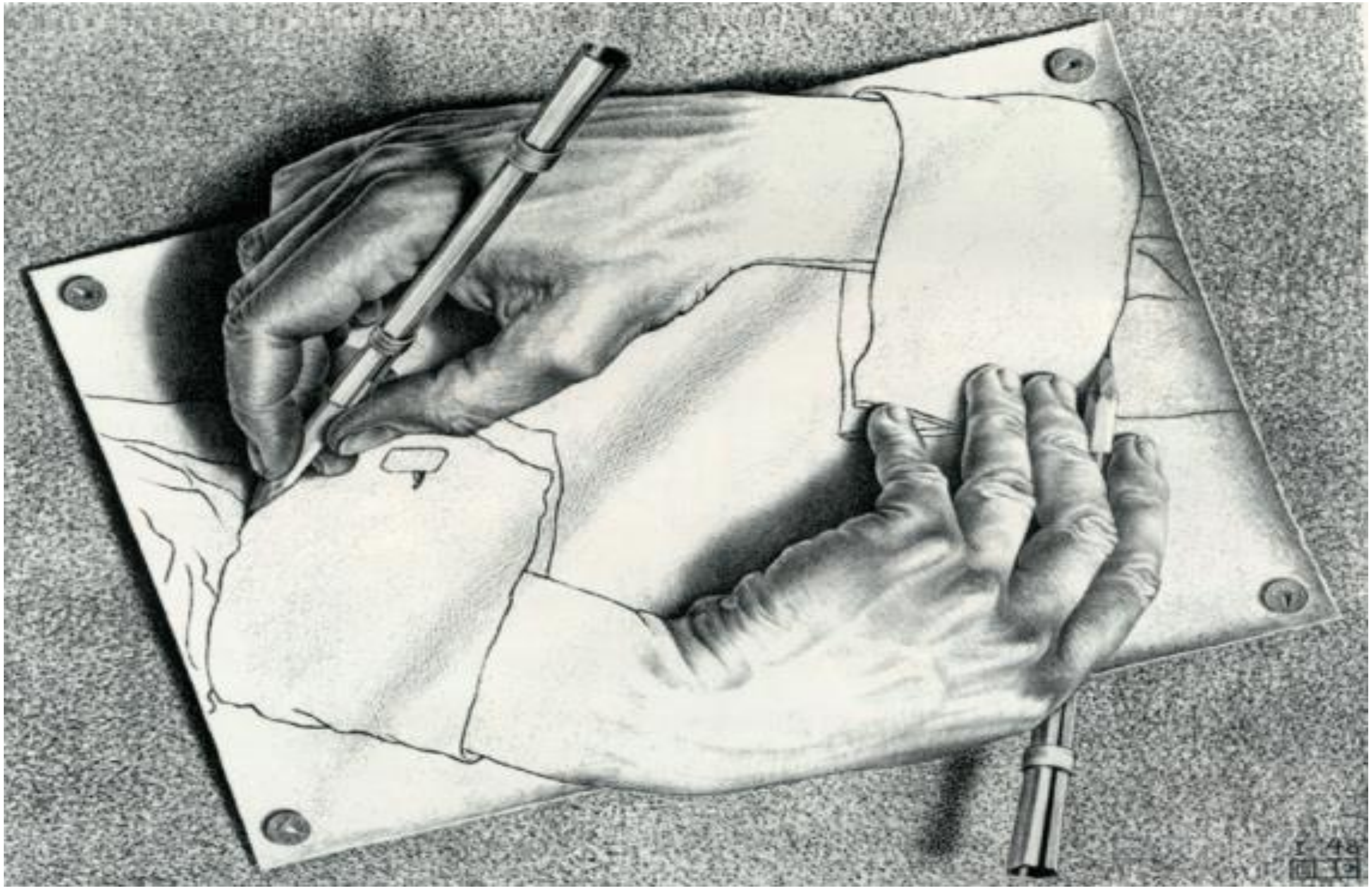
- “To see what is where by looking”
  - Aristotle, Marr, etc.
- .
- “To predict the world”
  - Jakob Uexküll, Jan Koenderink, Moshe Bar, etc.
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

# The world as supervision

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

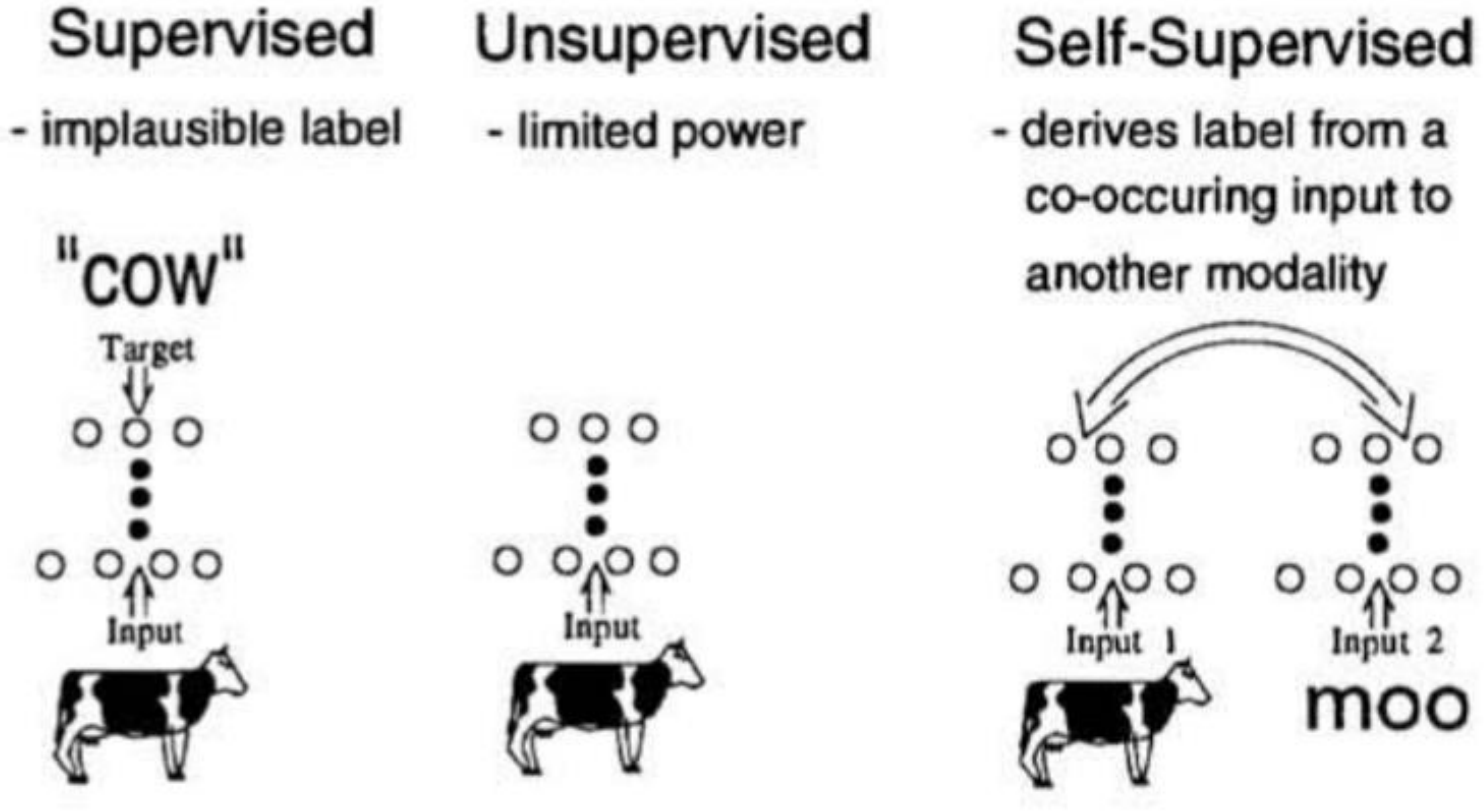
- What’s gonna happen next?
- What’s to my left?
- What can I touch?
- What will make a sound?
- Etc.

# Self-Supervision



Drawing Hands, M.C. Escher, 1948

# Self-Supervision in Multisensory Learning



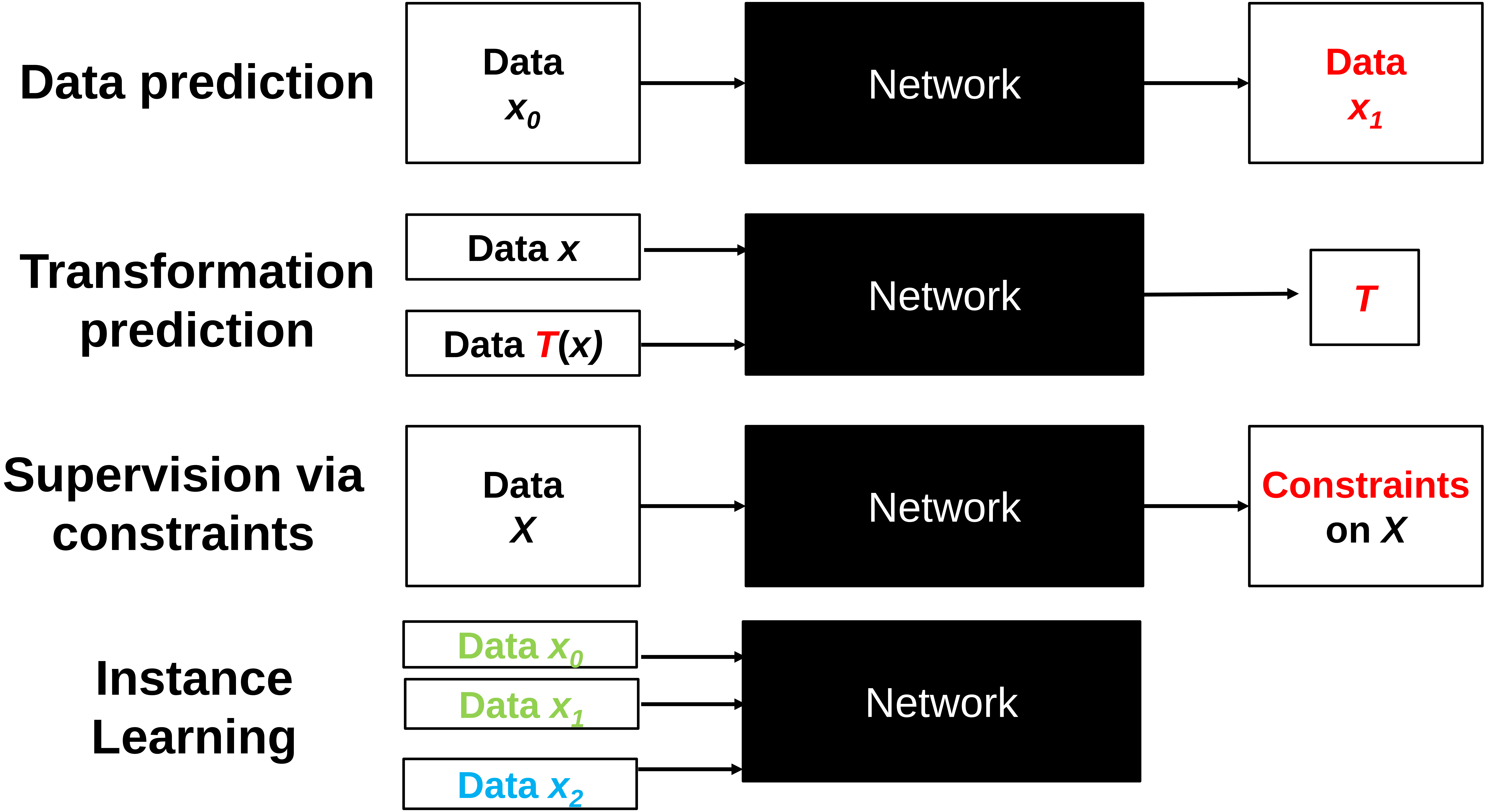

**Virginia De Sa**, "Learning classification from unlabelled data", NIPS 1994

# Context as Supervision

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

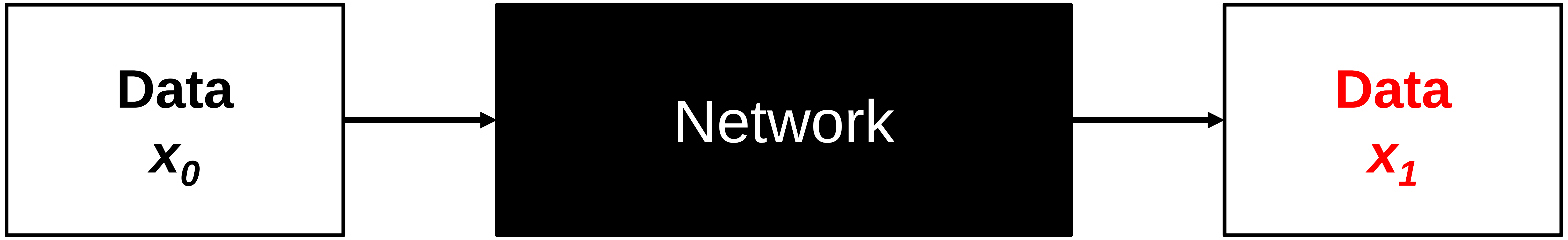

house, where the professor lived without his wife and child; or so he said jokingly sometimes: "Here's where I live. My house." His daughter often added, without resentment, for the visitor's information, "It started out to be for me, but it's really his." And she might reach in to bring forth an inch-high 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

# (Partial) Taxonomy of Self-Supervision

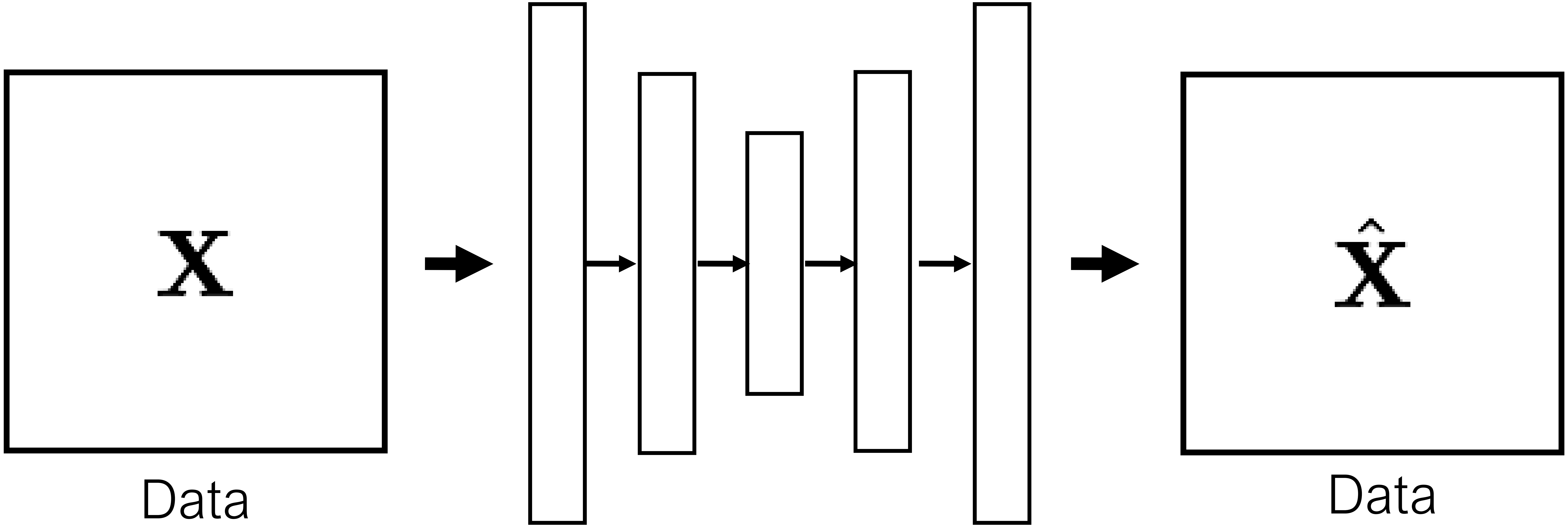


# (Partial) Taxonomy of Self-Supervision

**Data prediction**

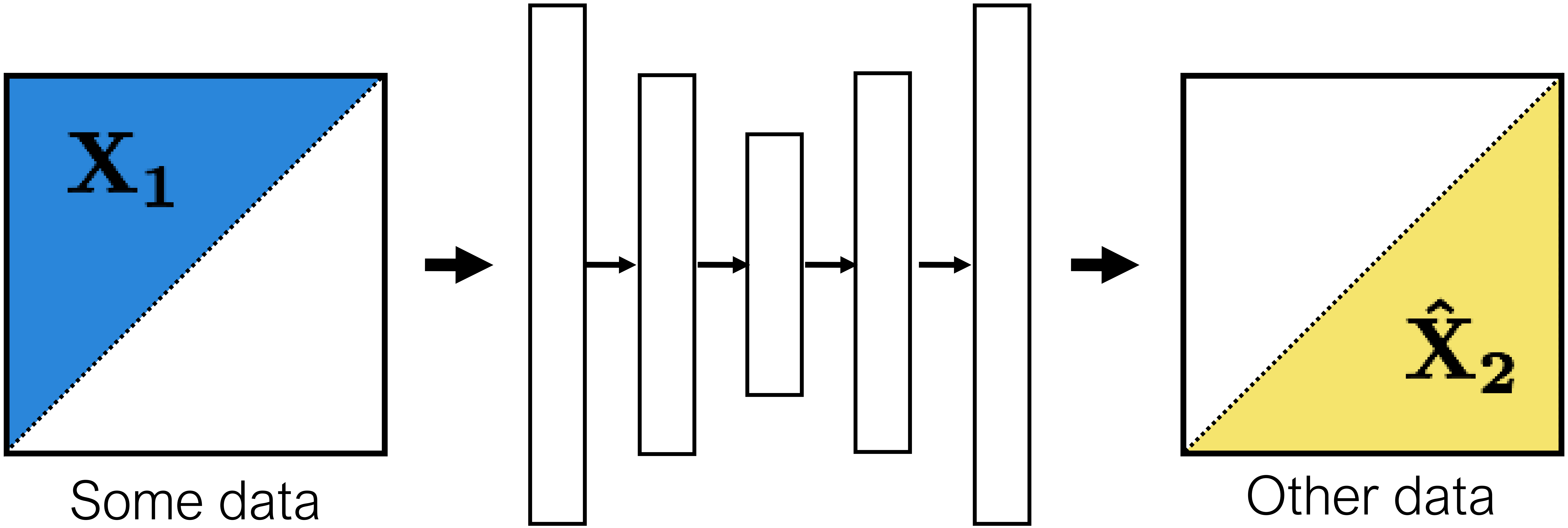


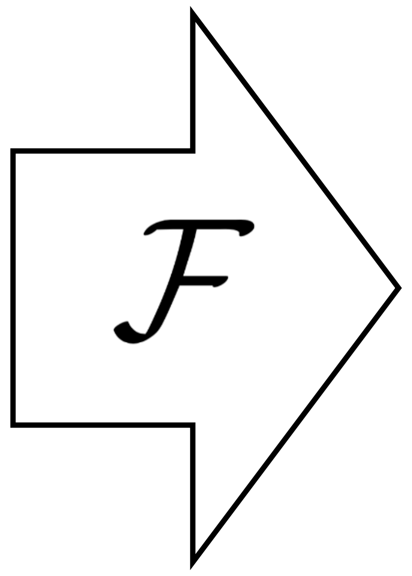
# Self-supervision as data compression




Autoencoders [Hinton & Salakhutdinov, Science 2009]

# Self-supervision as data prediction



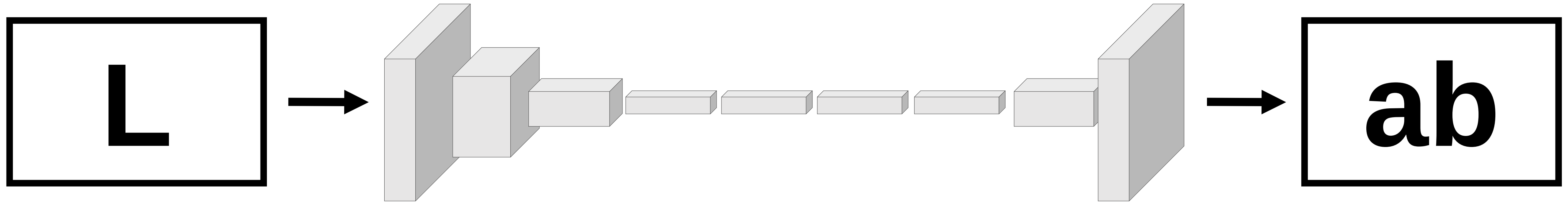



Grayscale image: L channel

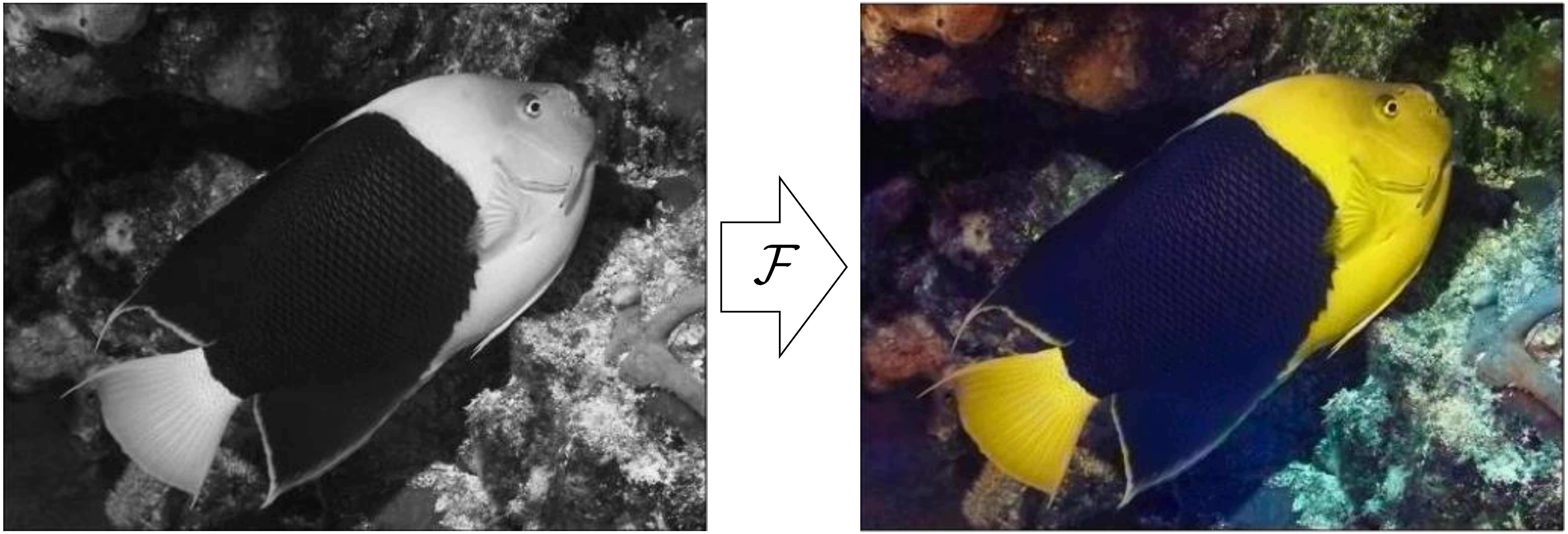
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: ab channels

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

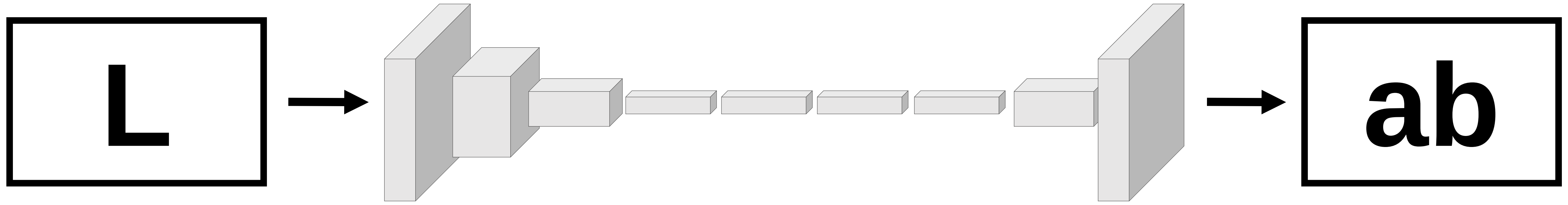



[Zhang, Isola, Efros, ECCV 2016]




Grayscale image: L channel
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

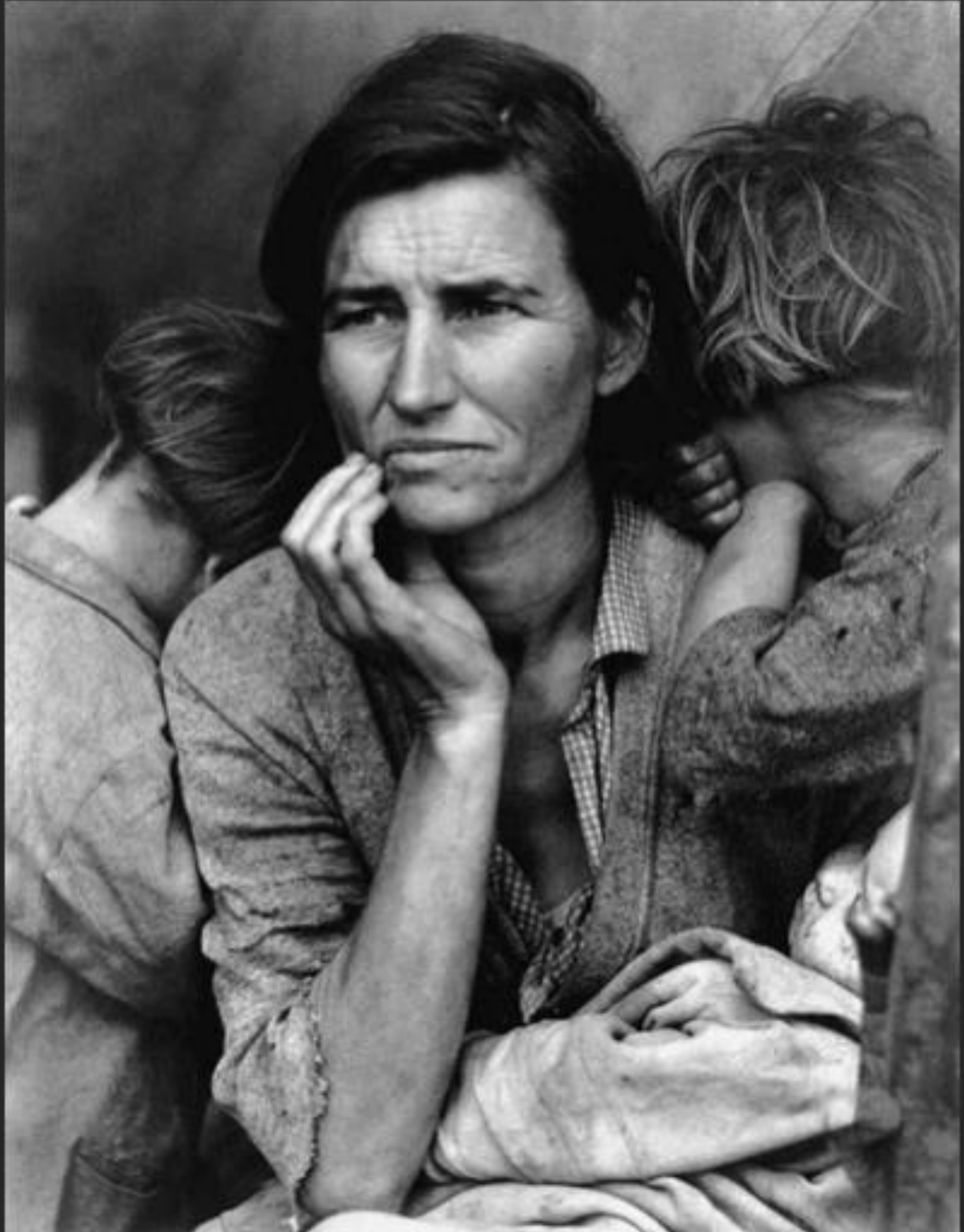
Color information: ab channels
$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



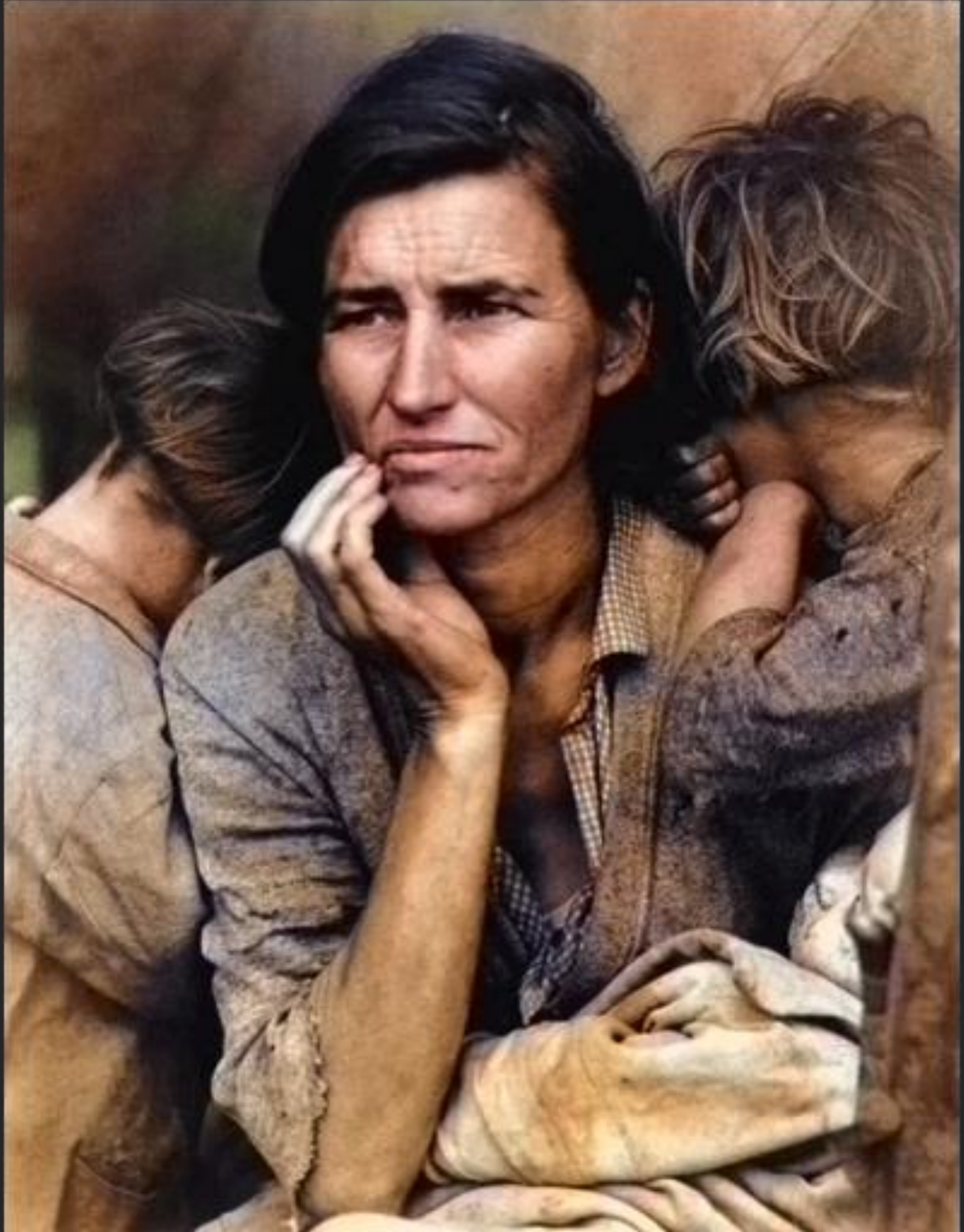

[Zhang, Isola, Efros, ECCV 2016]


Ansel Adams, Yosemite Valley Bridge


Our result



*Migrant Mother*
Dorothea Lange
1936



Our result

$\mathcal{F}$

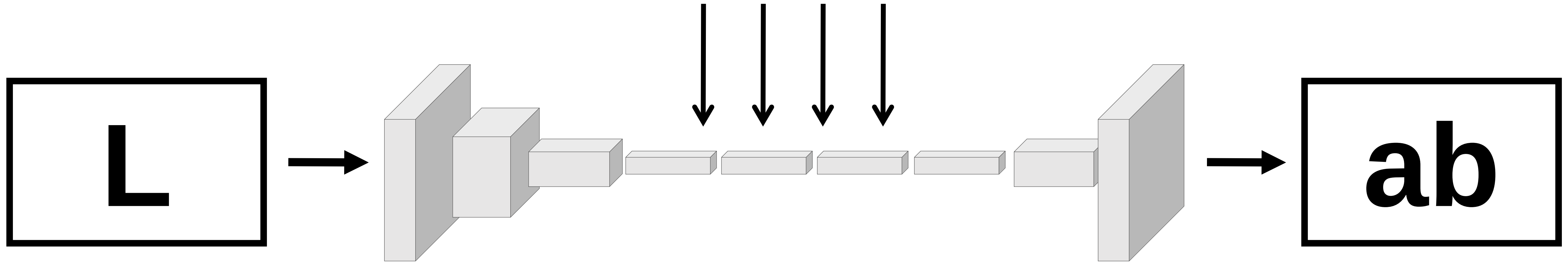
Grayscale image: L chan

Semantics? Higher-level abstraction?

information: ab channels

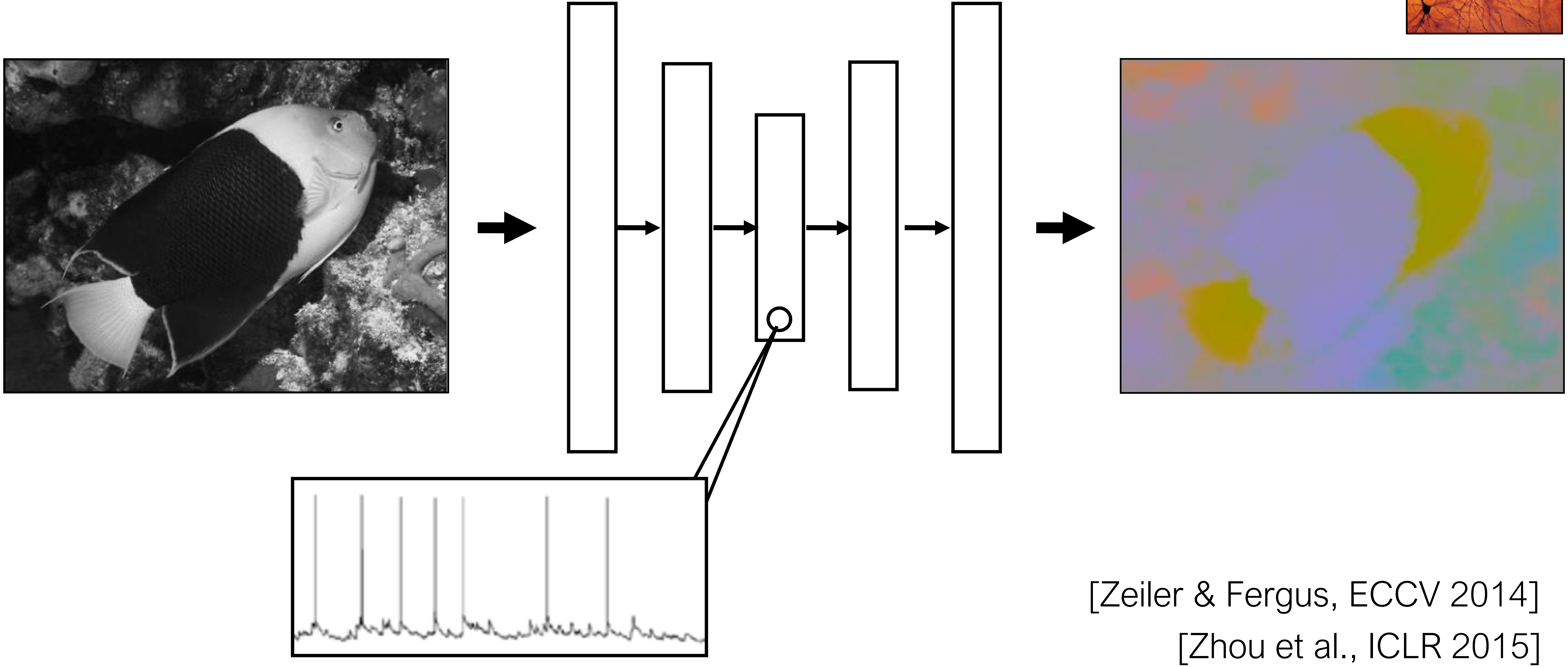
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

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




[Zhang, Isola, Efros, ECCV 2016]

# Deep Net "Electrophysiology"



[Zeiler & Fergus, ECCV 2014]

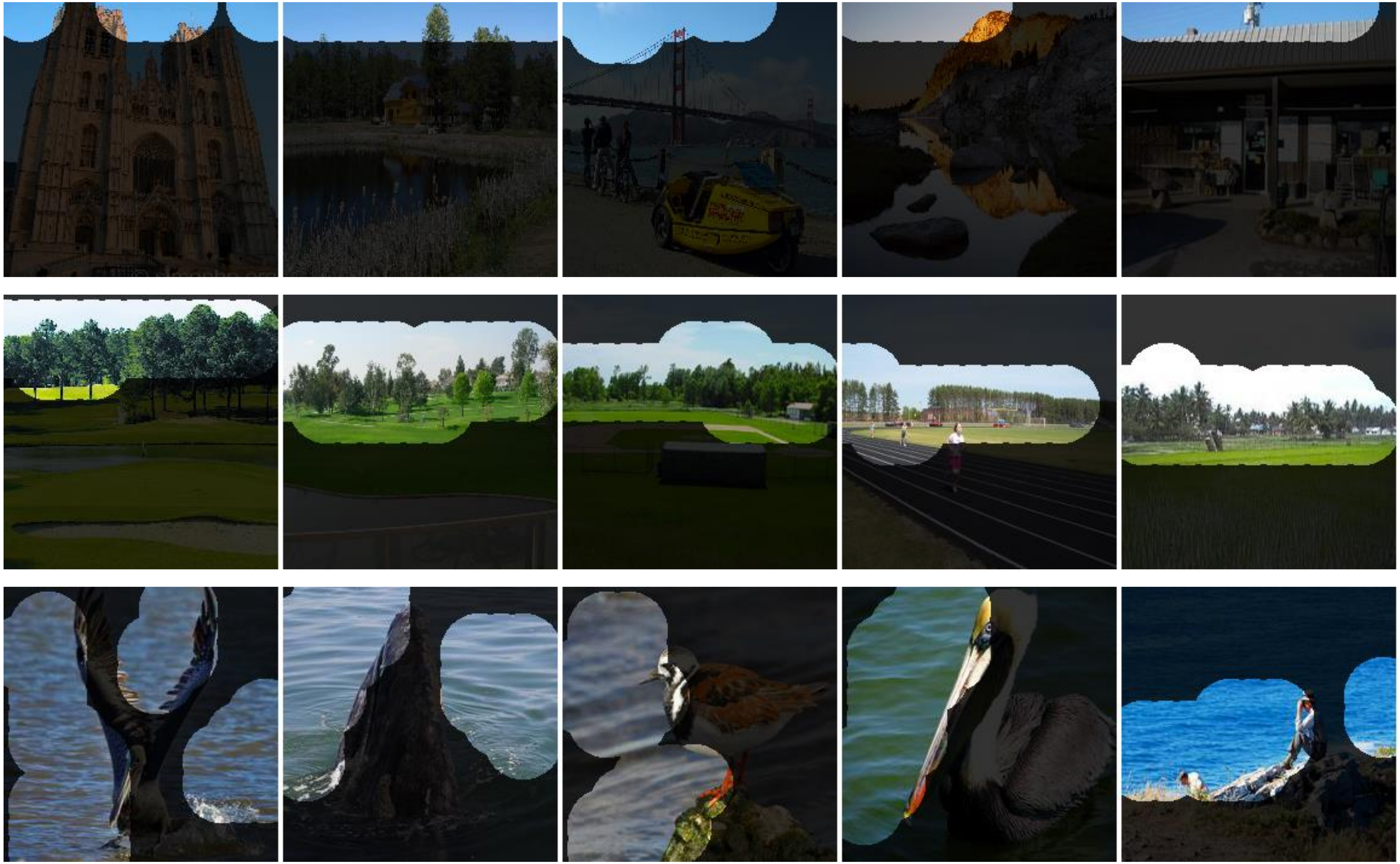
[Zhou et al., ICLR 2015]

# Hidden Unit (conv5) Activations

sky

trees

water

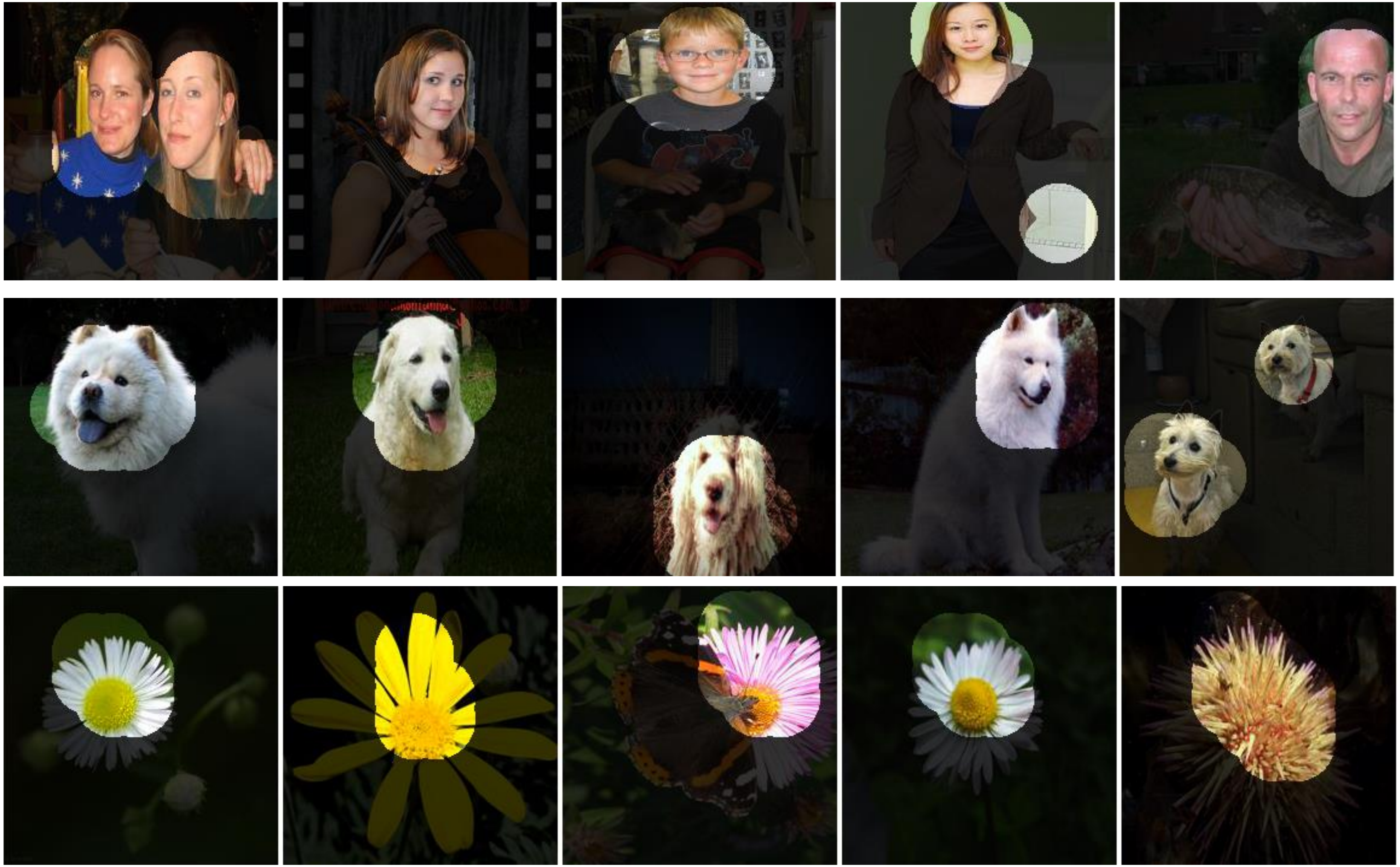


# Hidden Unit (conv5) Activations

faces

dog faces

flowers



# ImageNet Linear "Probing" – big mistake!

12 Zhang, Isola, Efros

| Dataset and Task Generalization on PASCAL [37] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Class. (%mAP) | | | Det. (%mAP) | | Seg. (%mIU) | | |
| **fine-tune layers** | **[Ref]** | **fc8** | **fc6-8** | **all** | **[Ref]** | **all** | **[Ref]** | **all** |
| ImageNet [38] | - | 76.8 | 78.9 | 79.9 | [36] | 56.8 | [42] | 48.0 |
| Gaussian | [10] | – | – | 53.3 | [10] | 43.4 | [10] | 19.8 |
| Autoencoder | [16] | 24.8 | 16.0 | 53.8 | [10] | 41.9 | [10] | 25.2 |
| k-means [36] | [16] | 32.0 | 39.2 | 56.6 | [36] | 45.6 | [16] | 32.6 |
| Agrawal et al. [8] | [16] | 31.2 | 31.0 | 54.2 | [36] | 43.9 | – | – |
| Wang & Gupta [15] | – | 28.1 | 52.2 | 58.7 | [36] | 47.4 | – | – |
| *Doersch et al. [14] | [16] | 44.7 | 55.1 | **65.3** | [36] | **51.1** | – | – |
| *Pathak et al. [10] | [10] | – | – | 56.5 | [10] | 44.5 | [10] | 29.7 |
| *Donahue et al. [16] | – | 38.2 | 50.2 | 58.6 | [16] | 46.2 | [16] | 34.9 |
| Ours (gray) | – | **52.4** | **61.5** | **65.9** | – | 46.1 | – | 35.0 |
| Ours (color) | – | **52.4** | **61.5** | **65.6** | – | 46.9 | – | **35.6** |

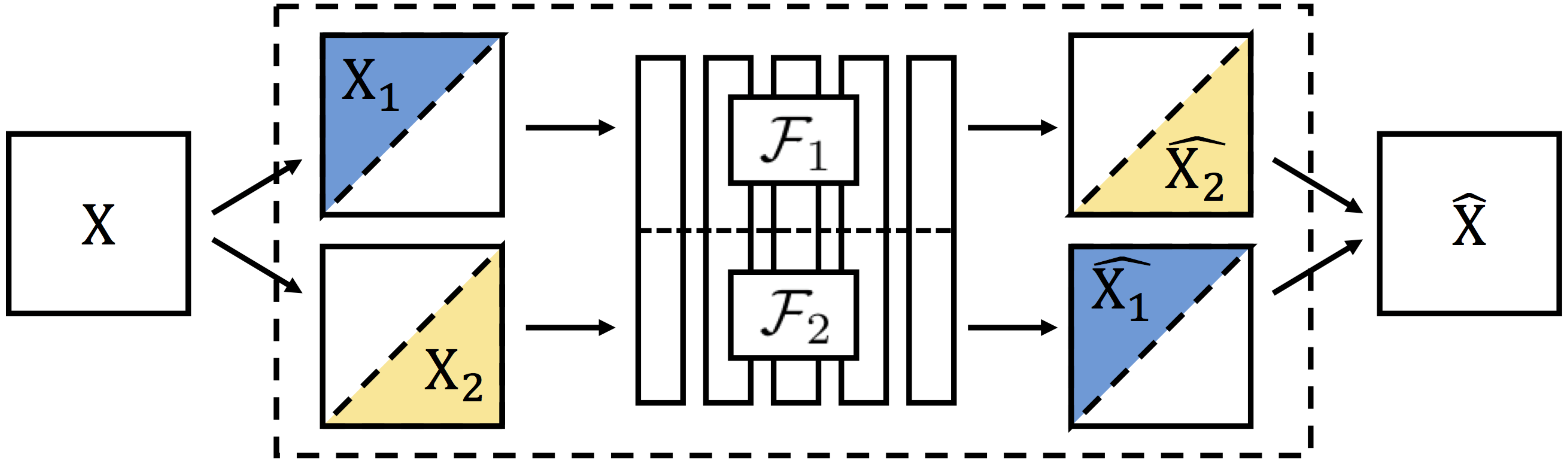
ion **Table 2.** PASCAL Tests

The "probe" has 2000 * 1000 = 2,000,000 parameters!

[Zhang, Isola, Efros, ECCV 2016]

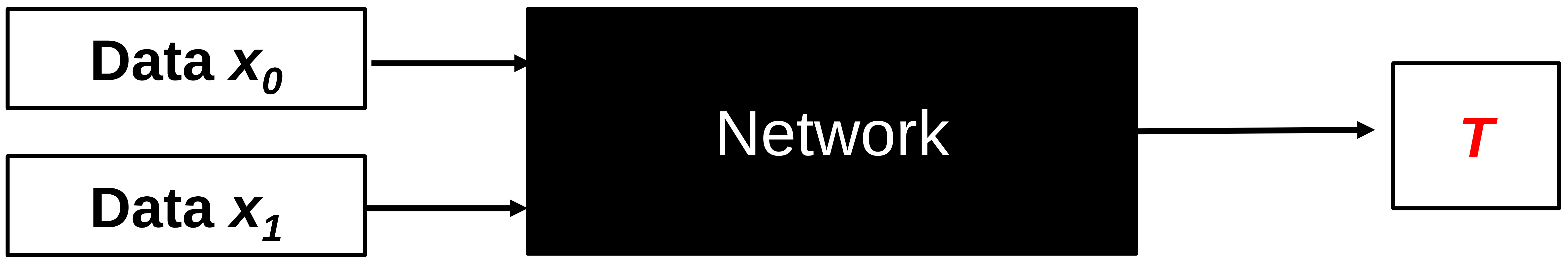
# Self-supervision as data prediction

Context Encoder
Pathak et al. CVPR 2016

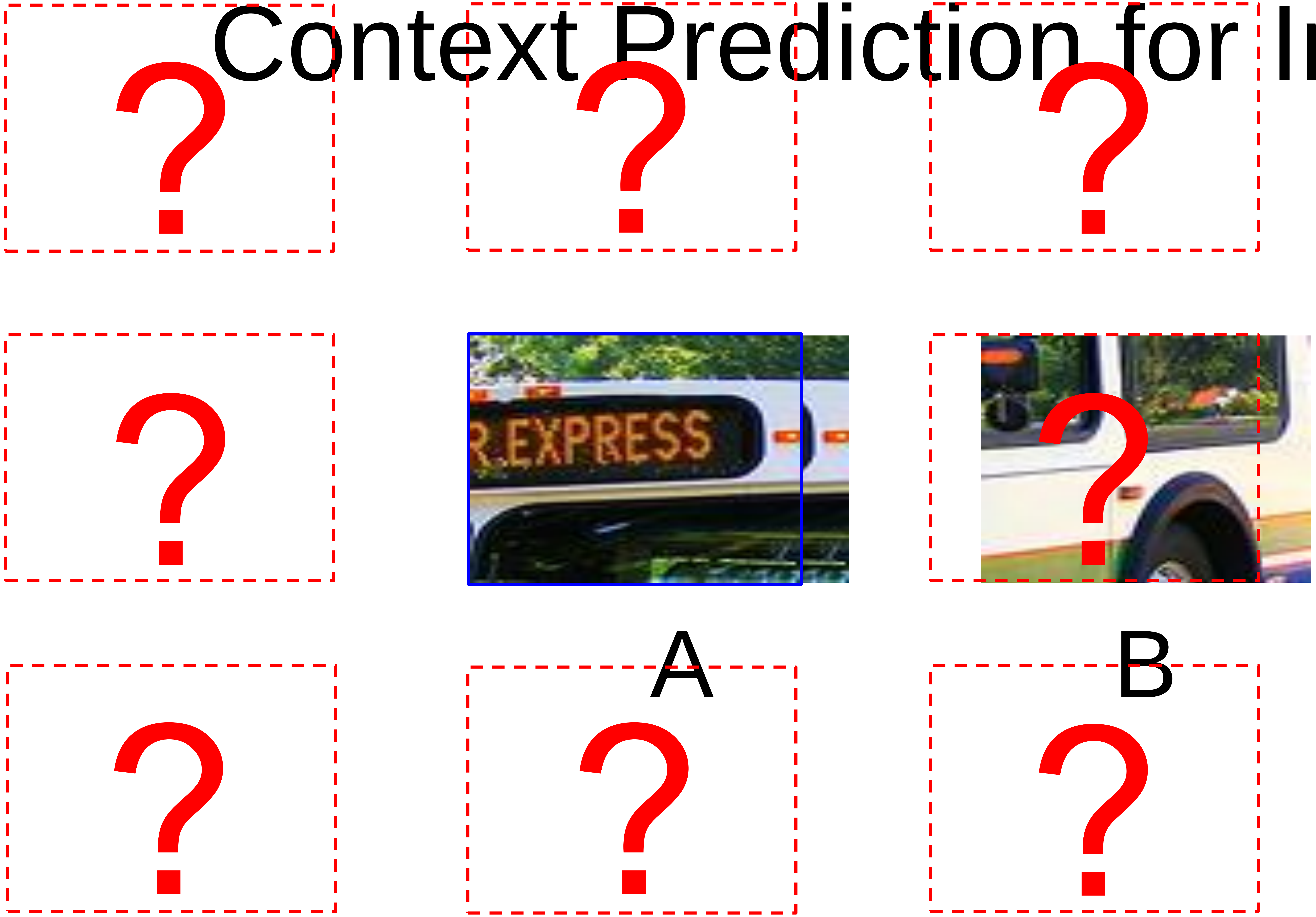



Split-brain Autoencoder, Zhang et al, CVPR 2017

# Self-supervision as transformation predication

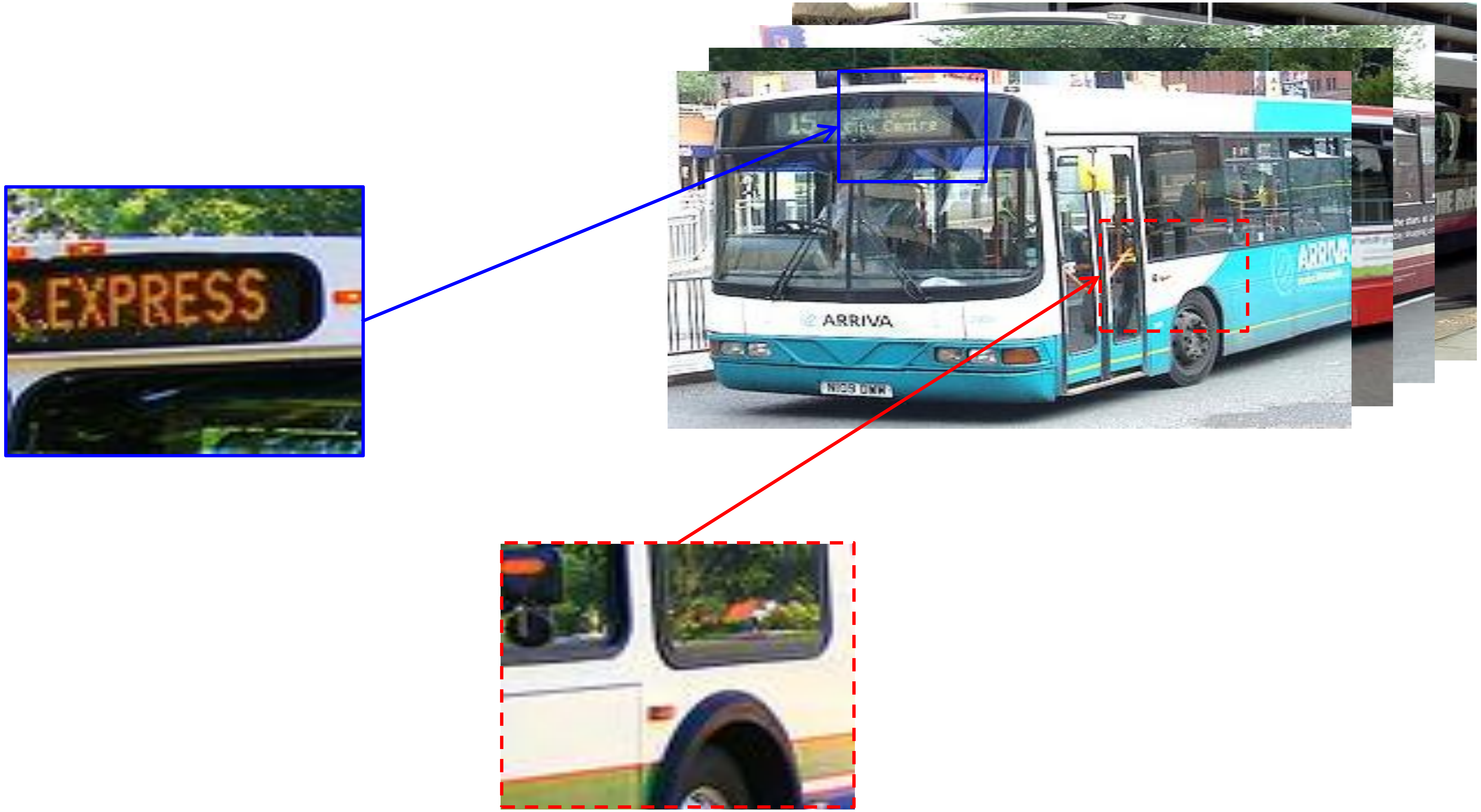


# Context Prediction for Images

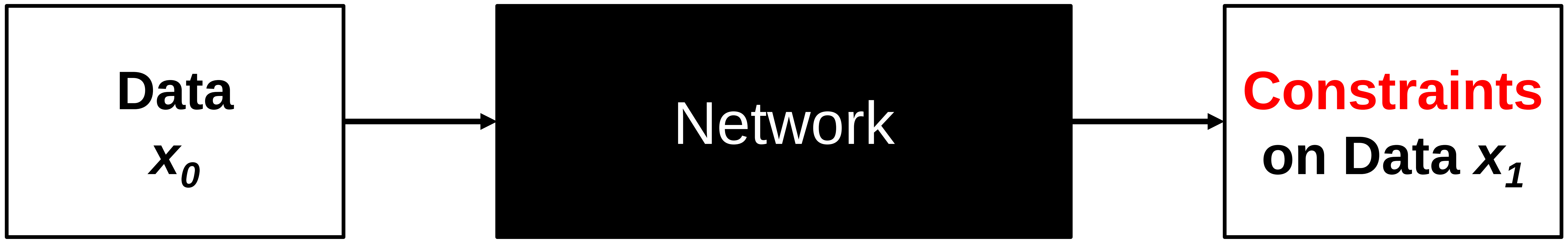



**Deorsch, Gupta, Efros ICCV 2015**

# Semantics from a non-semantic task



# Self-Supervision via Constraints




$F(x) = y$

- direct supervision

$F(x) \in Y$

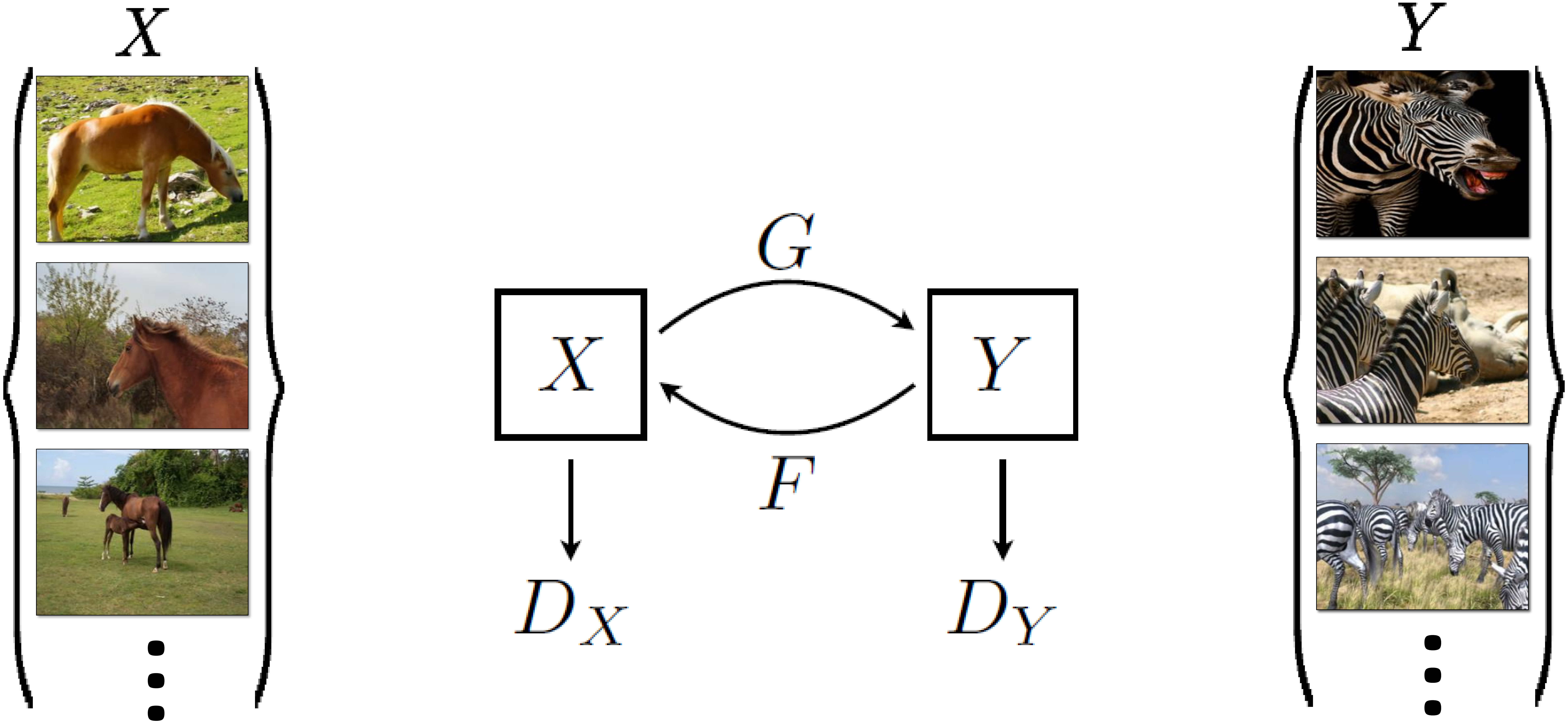
- GANs

$G(F(x)) = x$

- cycle-consistency

- …

# **CycleGAN**, or “there and back aGAN”

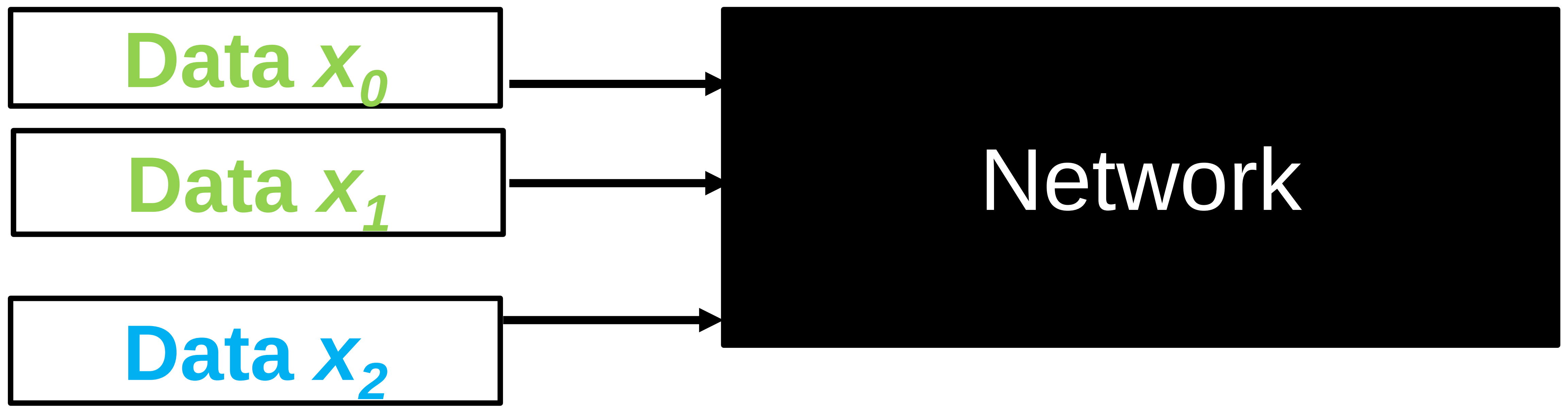



[Zhu*, Park*, Isola, Efros. ICCV 2017]

# Video

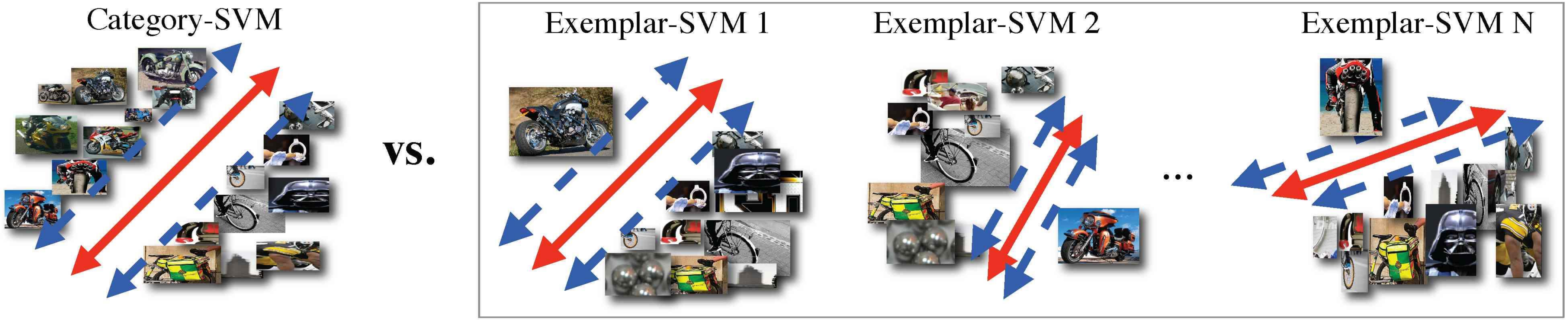
# Instance Learning

**Instance Learning**

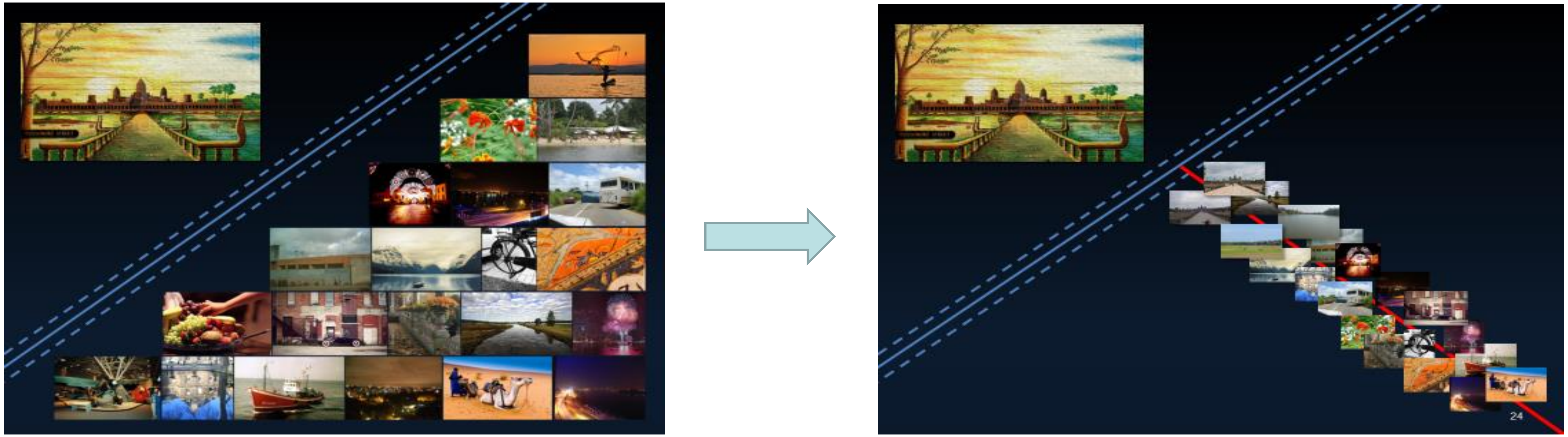


# **Exemplar-SVM:** *defining yourself by what you are <u>not</u>*

[Malisiewicz, Gupta, Efros, ICCV'11]




One-against-all learning for image retrieval [Srivastava et al, SIGGRAPH'11]



# Exemplar-CNN [Dosovitskiy et al, NIPS'14]

- single parametric representation (CNN)
- Data augmentation

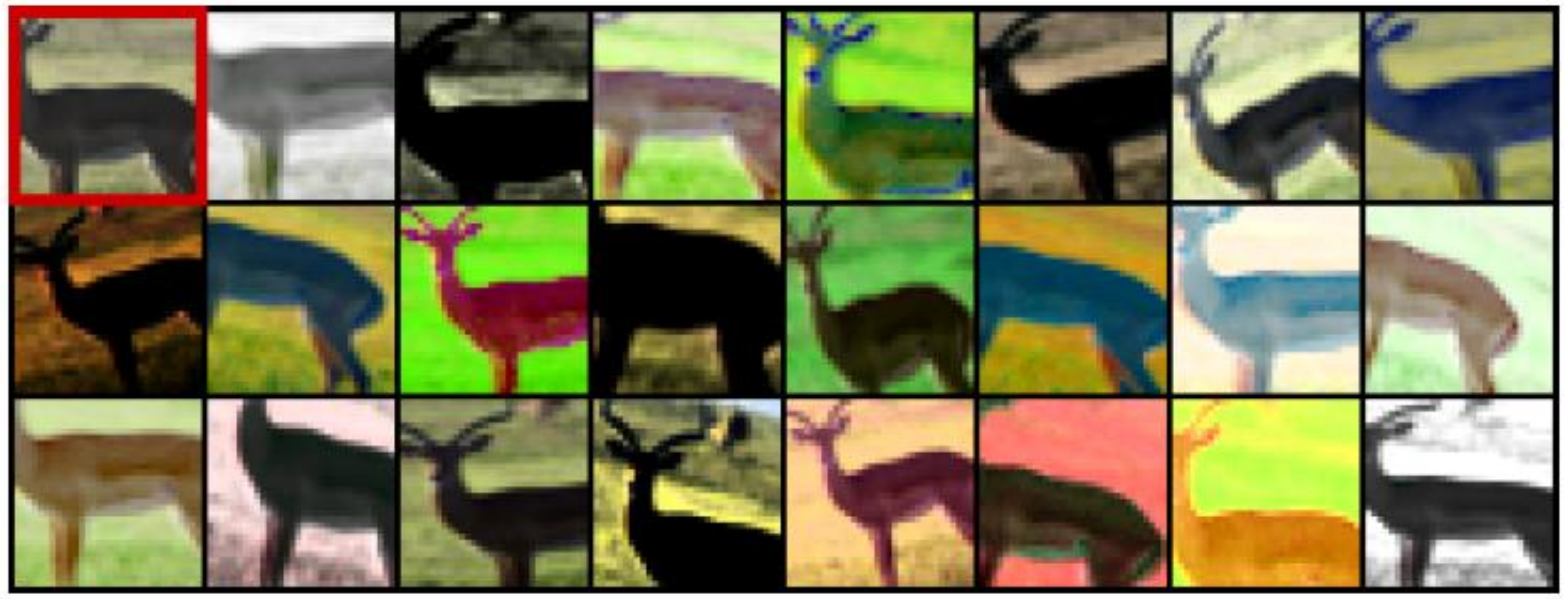
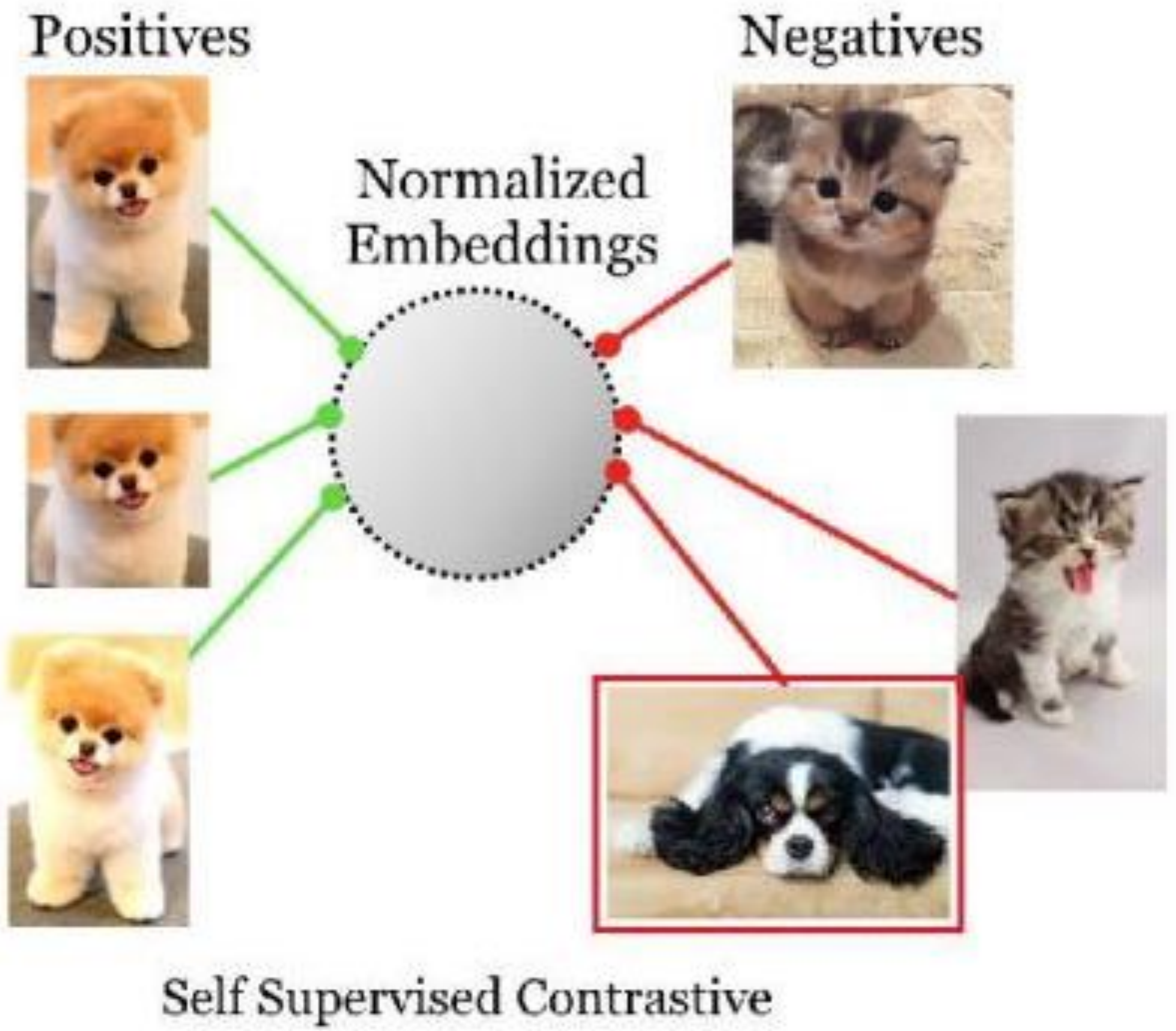


Fig. 2. Several random transformations applied to one of the patches extracted from the STL unlabeled dataset. The original ('seed') patch is in the top left corner.

# Modern Day: representations via Similarity Learning

- Metric Learning
- Siamese Nets
- Contrastive Learning
- etc

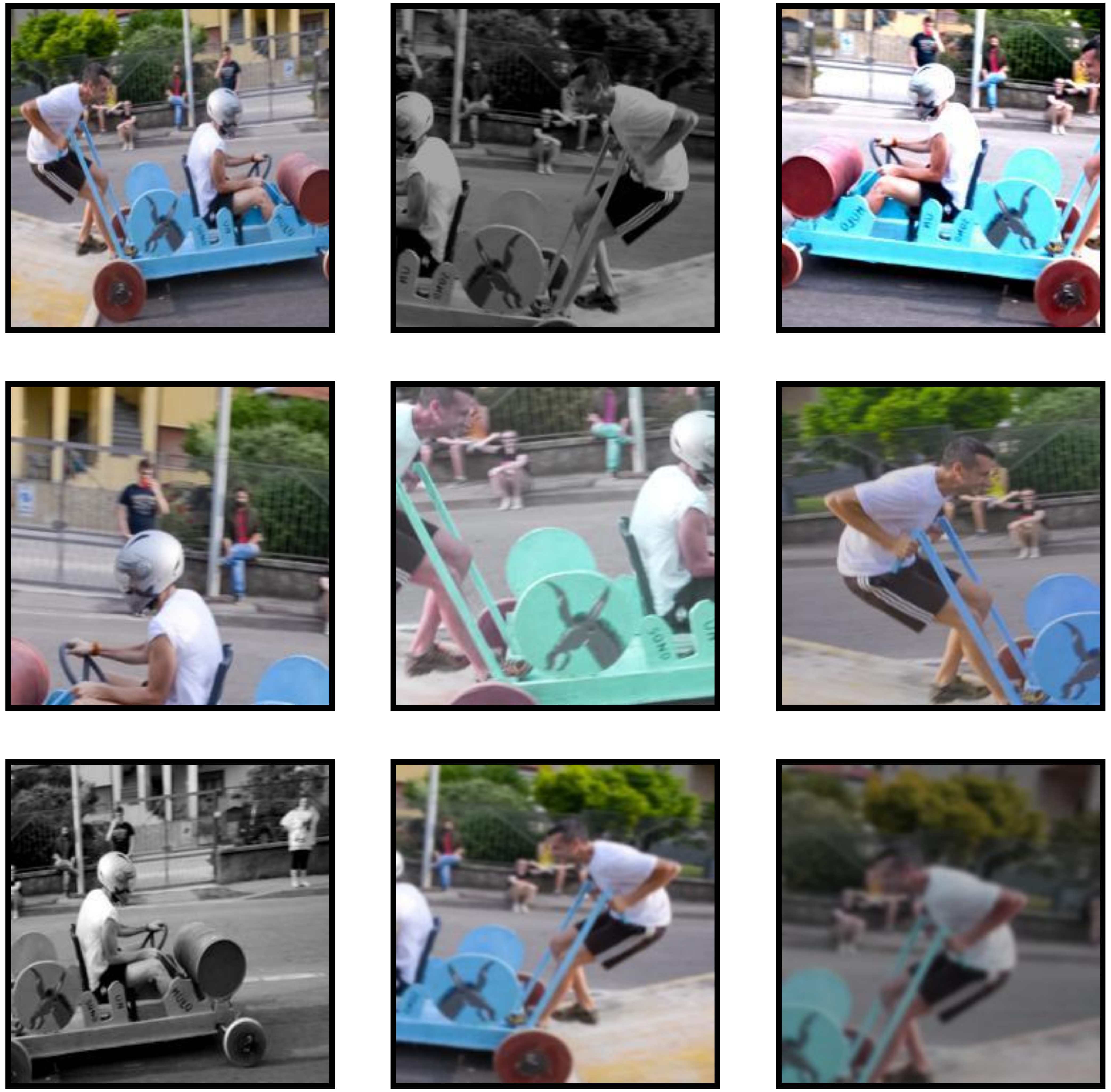



Becker et al (1992)
de Sa (1993)
Bromley et al (1994)
Chopra et al (2005)
Dosovitsky et al (2014)
Bojanowski et al (2017)
Wu et al (2018)
van den Oord et al (2019)
Tian et al (2019)
He et al (2019)
Chen et al (2020)

1. Improvements in representation learning (e.g. Contrastive)
2. **Improved Data Augmentations (e.g. cropping)**

# Data Augmentation

Views

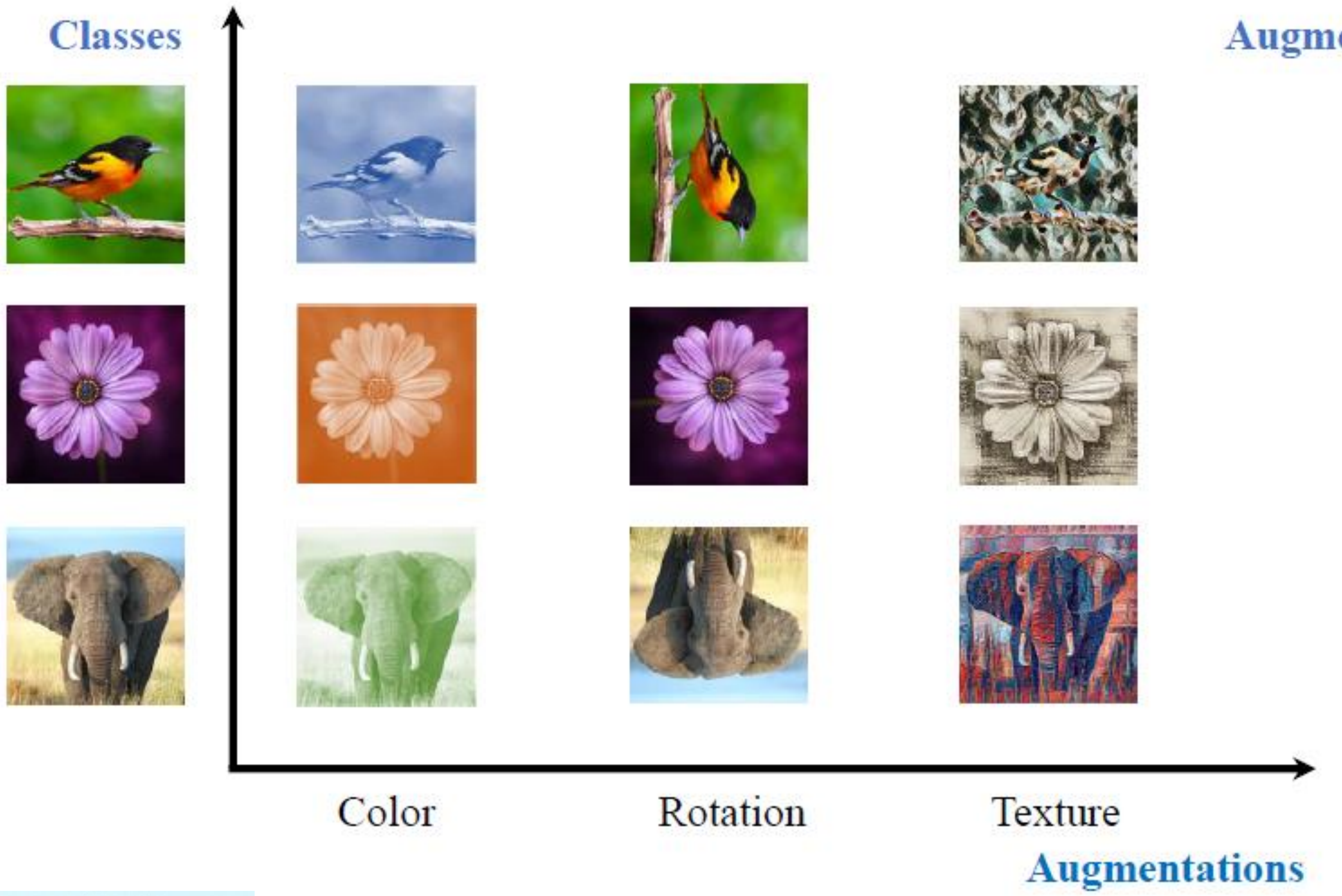


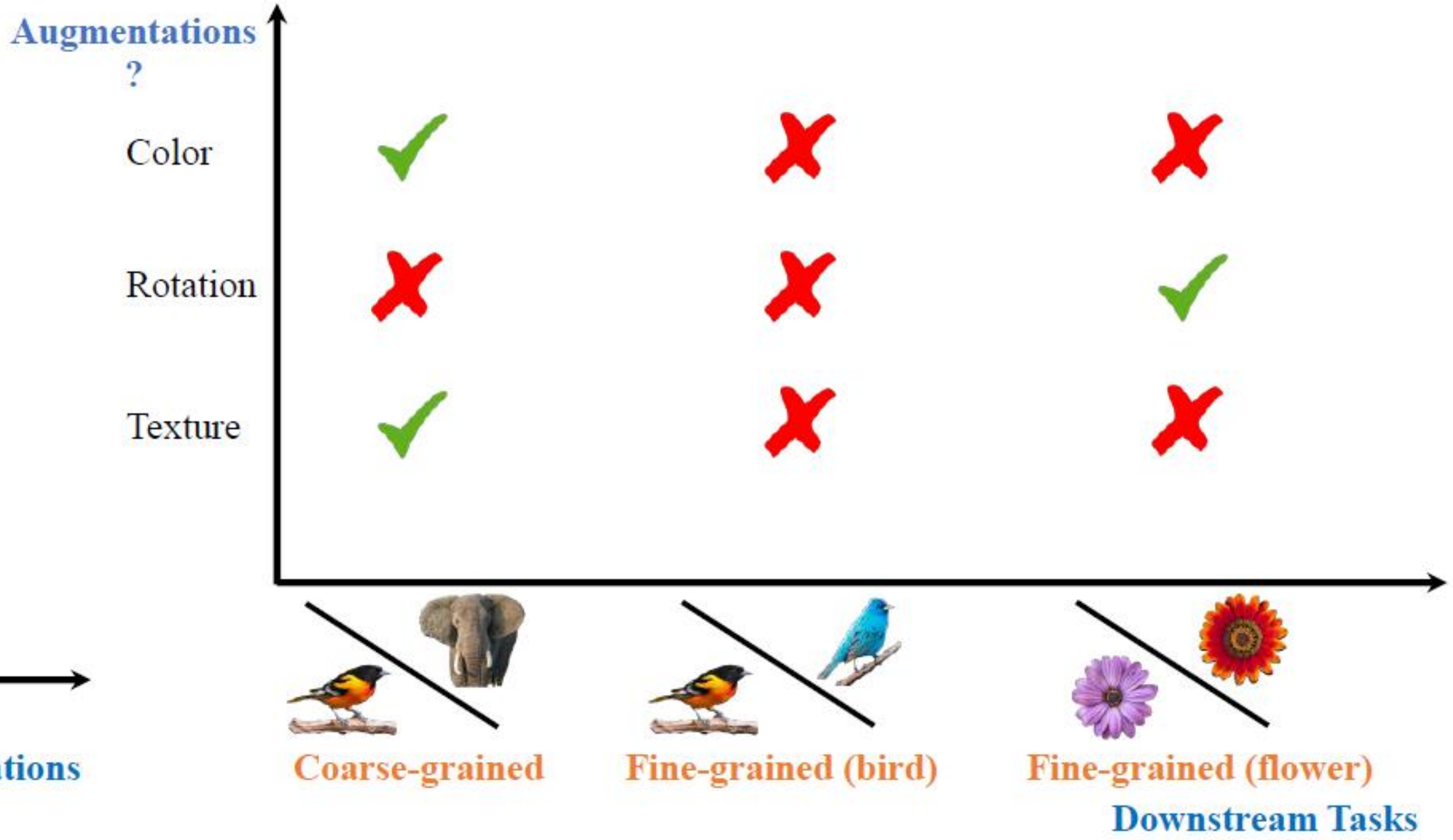
input

color
crop
flip
blur

SimCLR augmentations (Chen et al, 2020)

# The choice of data augmentation is itself supervision

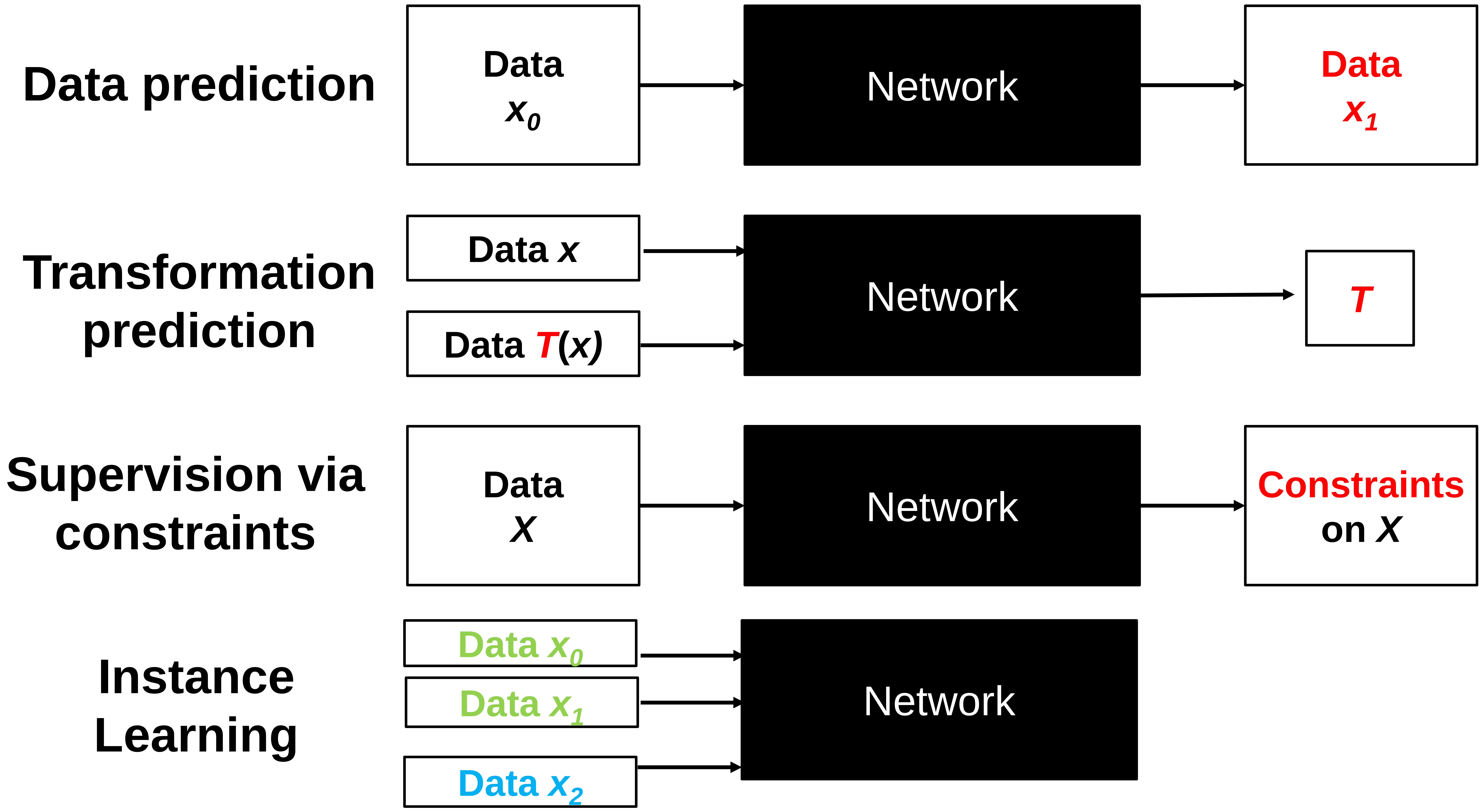



What should not be contrastive in contrastive learning
T Xiao, X Wang, AA Efros, T Darrell - ICLR 2021

# So, where are we now?

# (Partial) Taxonomy of Self-Supervision



# And the current winner is…

**Data prediction**

**Data** $x_0$ → Network → **Data** $x_1$

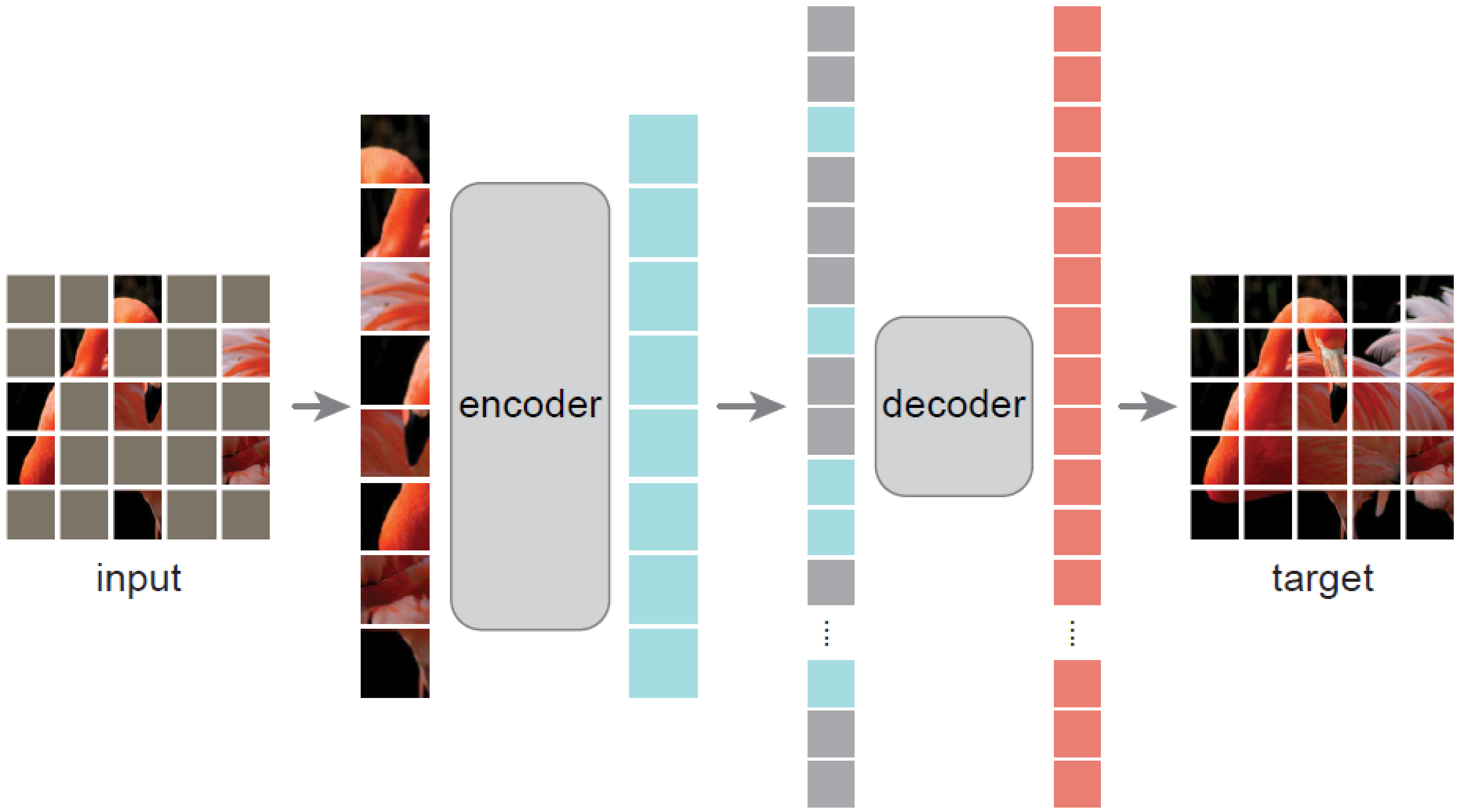
Context Encoder
Pathak et al. CVPR 2016

Masked Autoencoder
He et al. 2021

# And the current winner is…

**Data prediction**

**Data** $x_0$ → Network → **Data** $x_1$




Masked Autoencoder
He et al. 2021

# GPT-envy: “prompting” instead of fine-tuning

Prompt: *“Je suis désolé I’m sorry*

*J’adore la glace”*

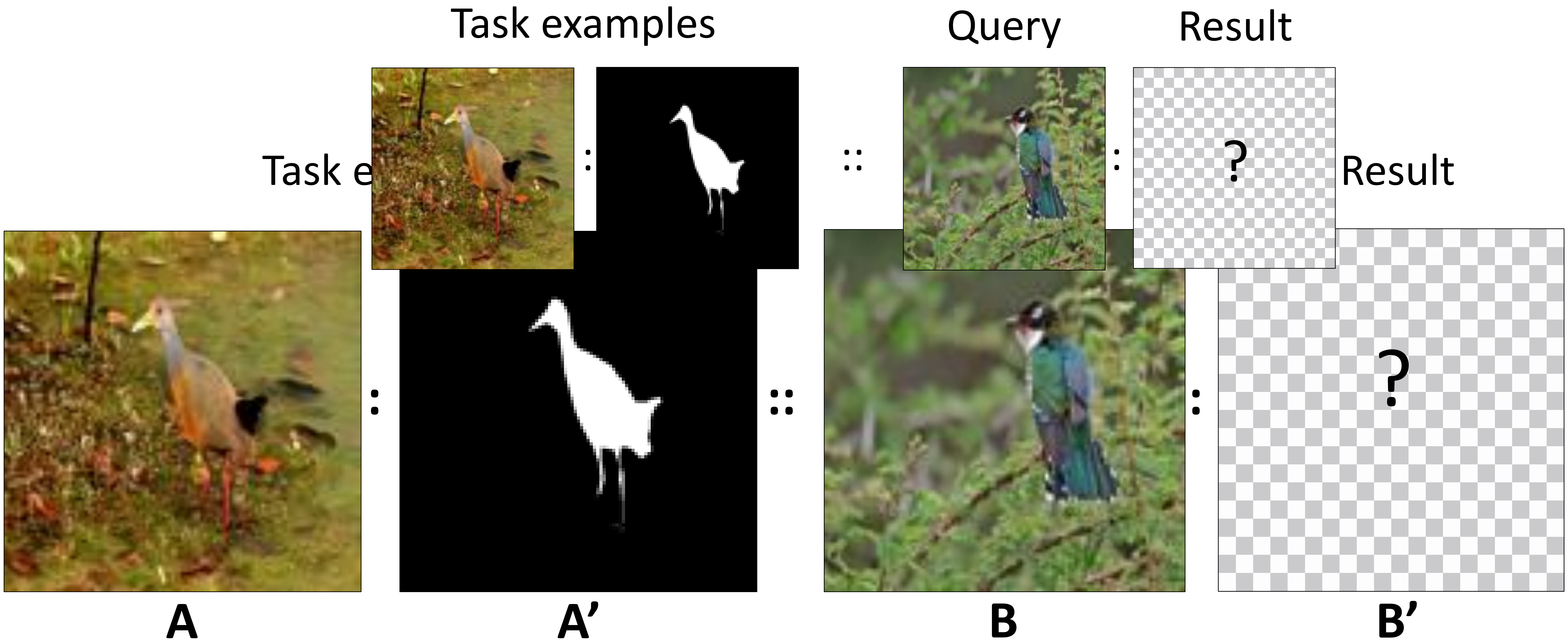
GPT-3

Output: *I love ice cream*

Can we do this for **visual data?**

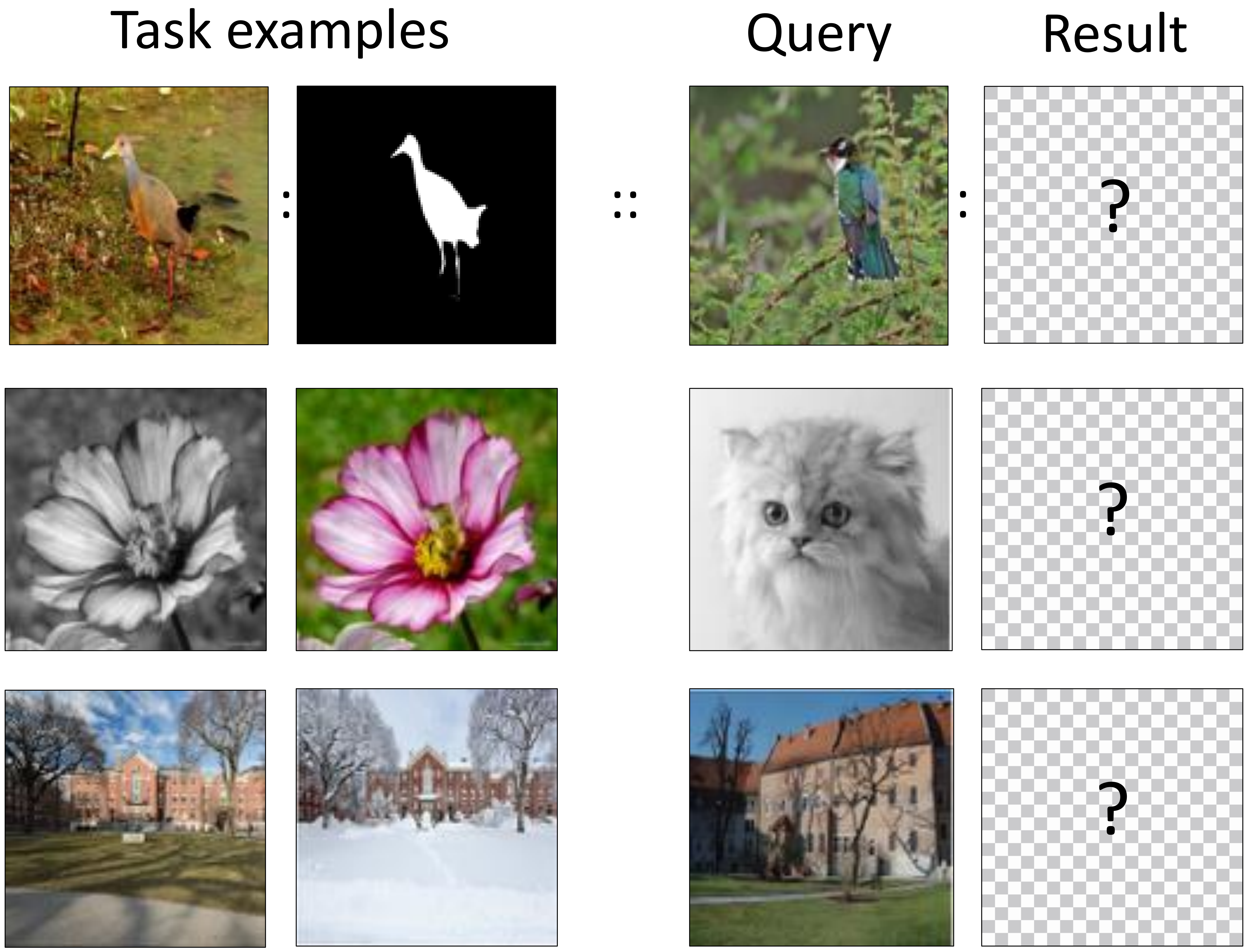
Brown et al. , “Language Models are Few-Shot Learners”, NeurIPS 2020

# Visual Prompting




Hertzmann, Aaron, et al. "Image analogies." SIGGRAPH 2001.

# Visual Prompting



# Visual Prompting via Image Inpainting

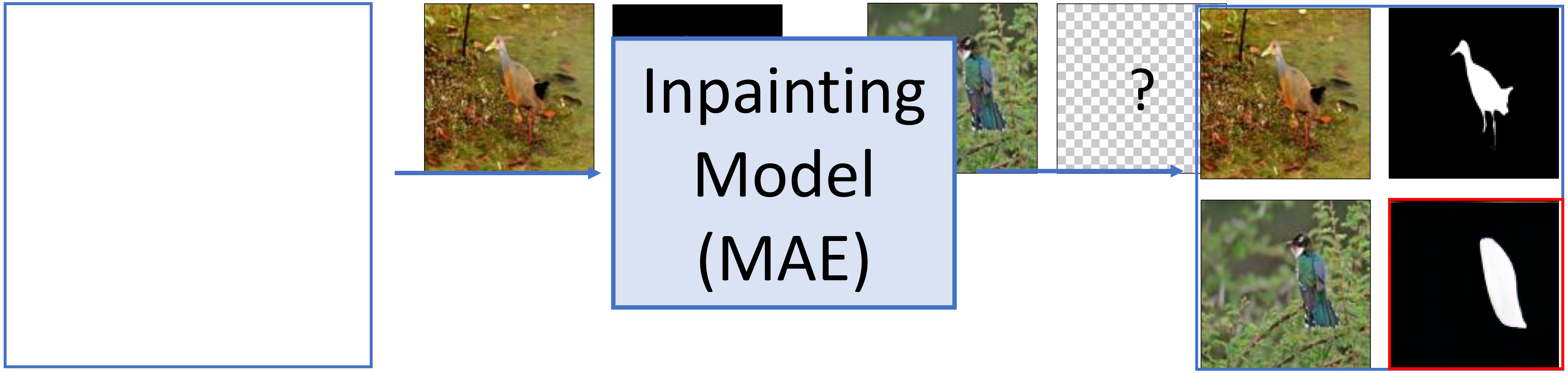
**Amir Bar** **Yossi Gandelsman**

Trevor Darrell, Amir Globerson, Alexei A Efros

NeurIPS 2022


* Equal contribution

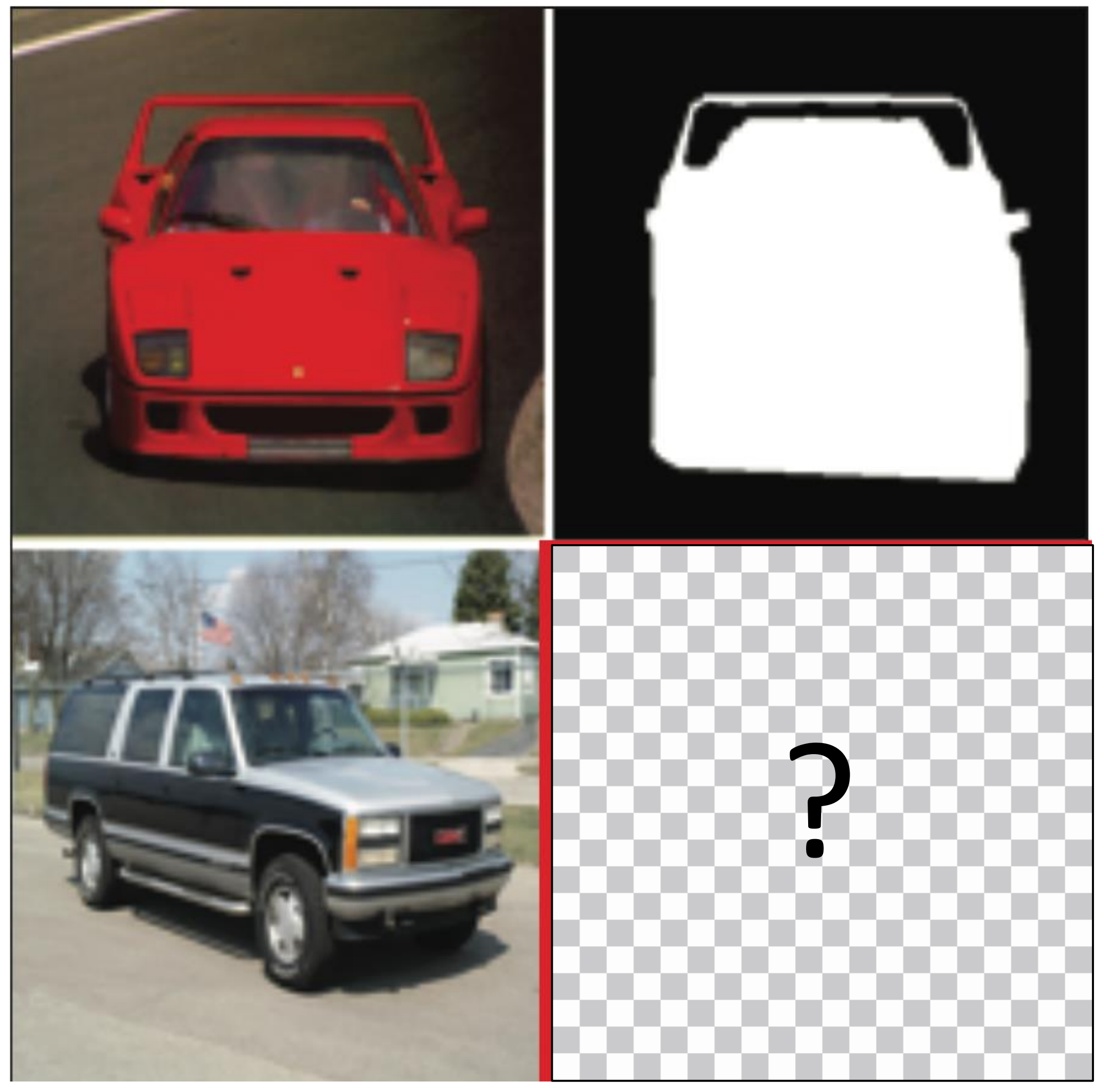
# Deep Inpainting to the rescue!



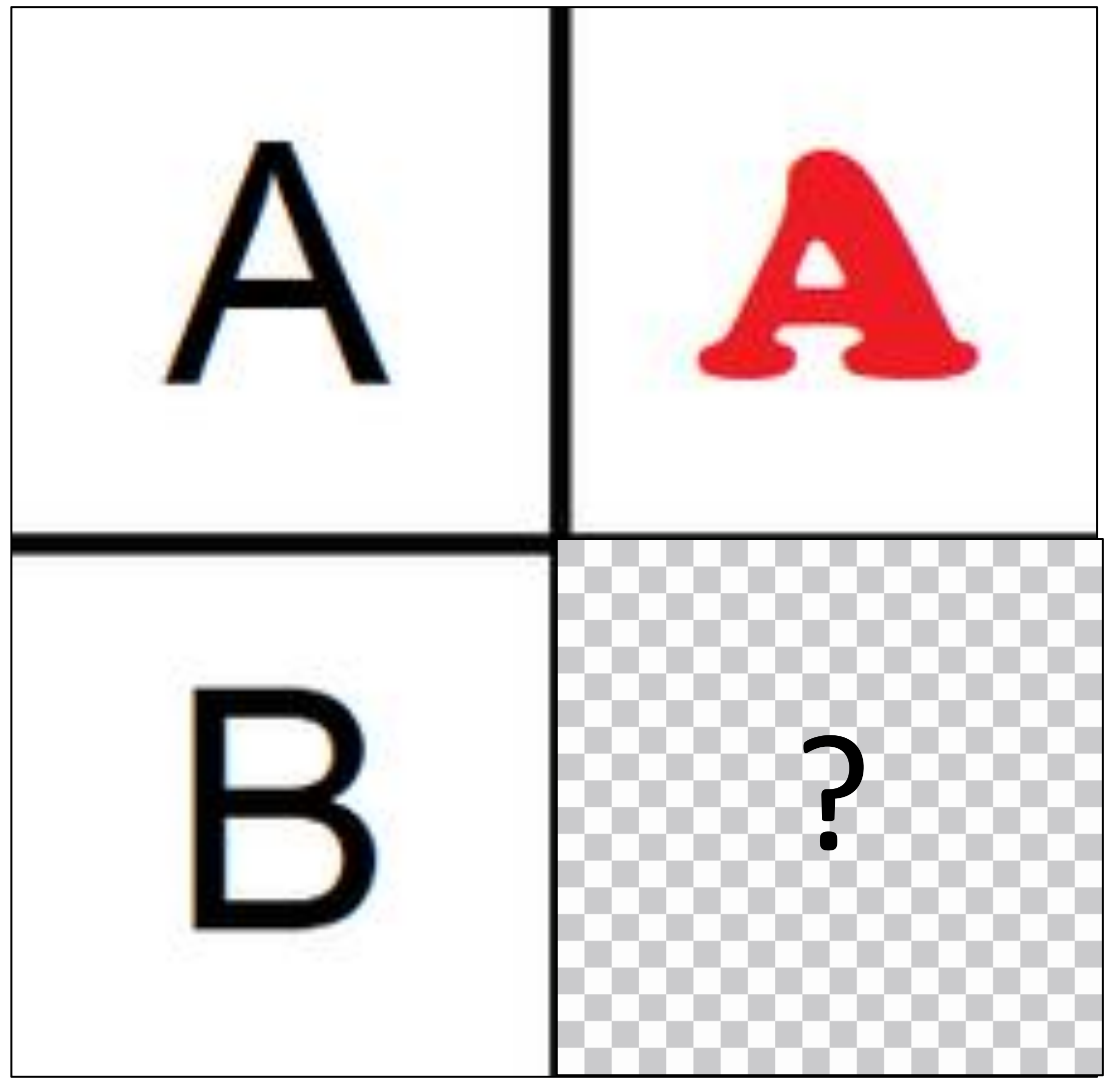

Visual prompt image

# Wide range of tasks

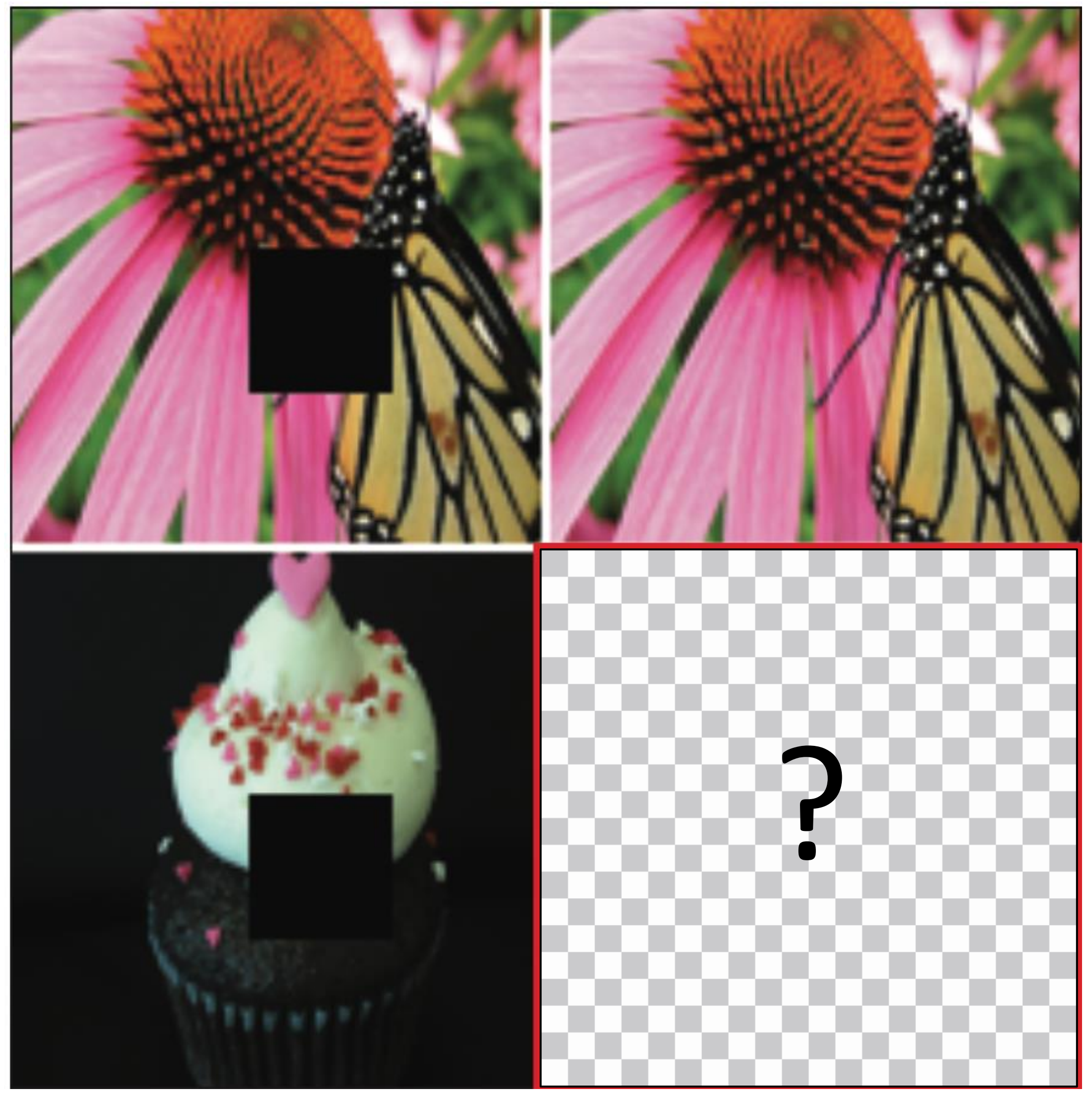

Colorization




Segmentation

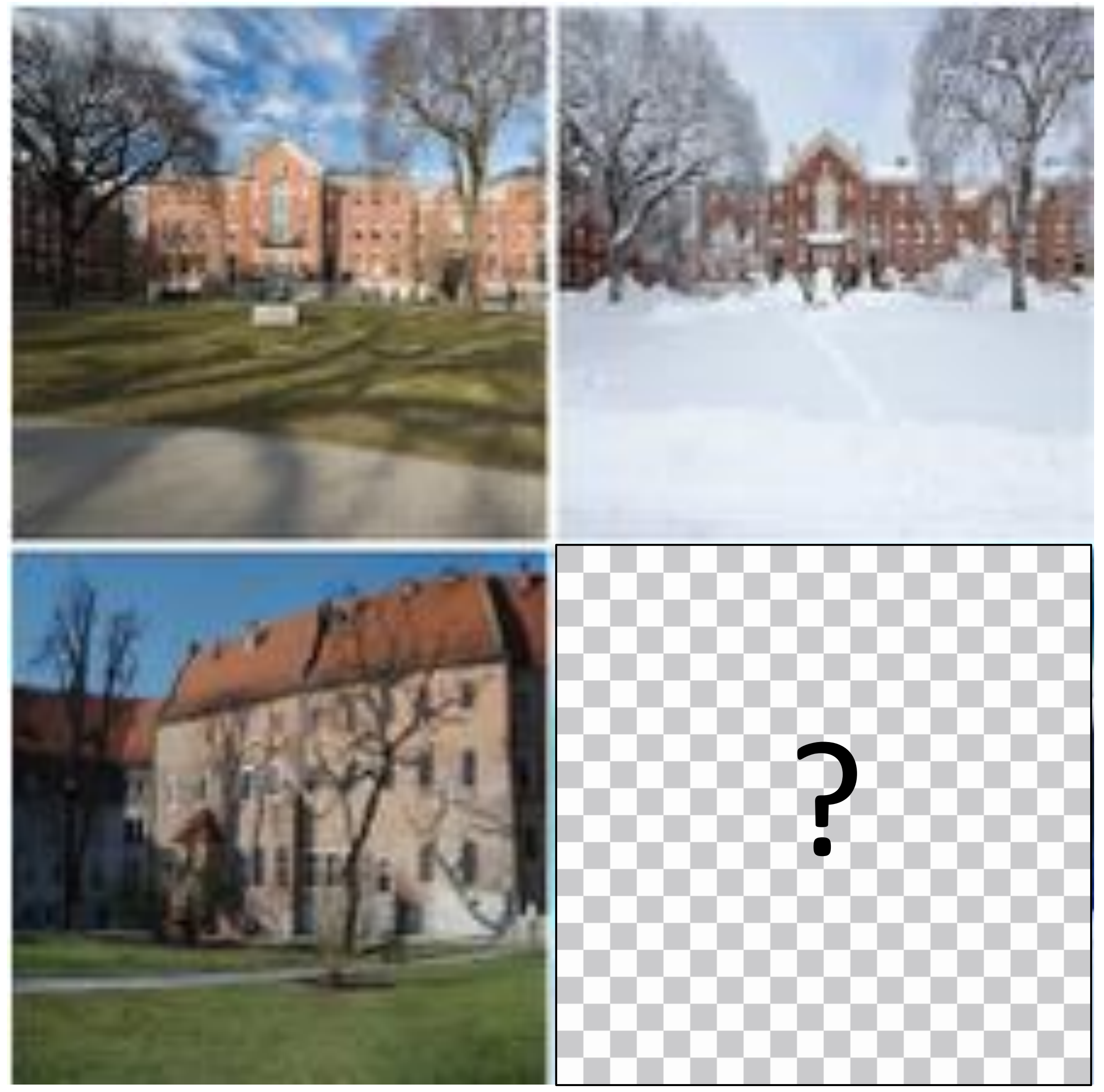



Font Style Transfer




Inpainting

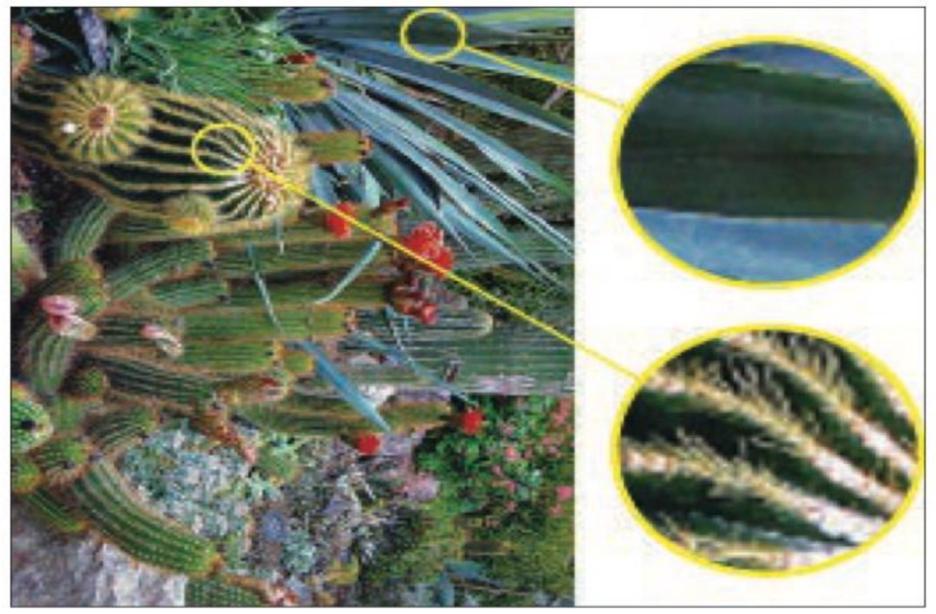

Edge Detection

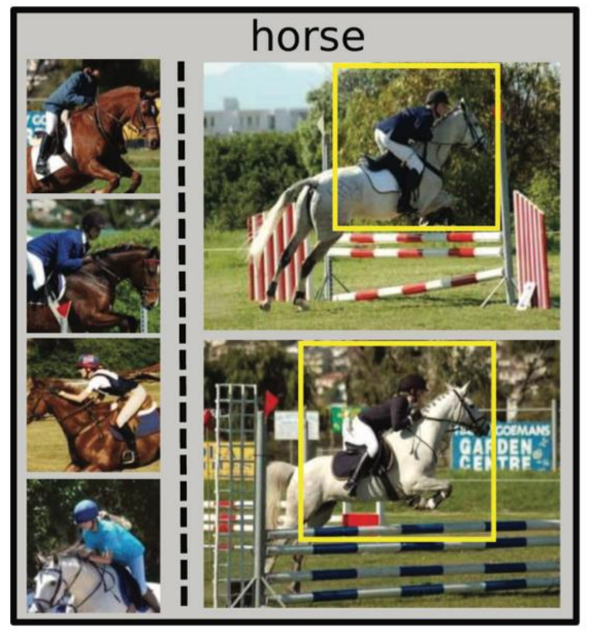


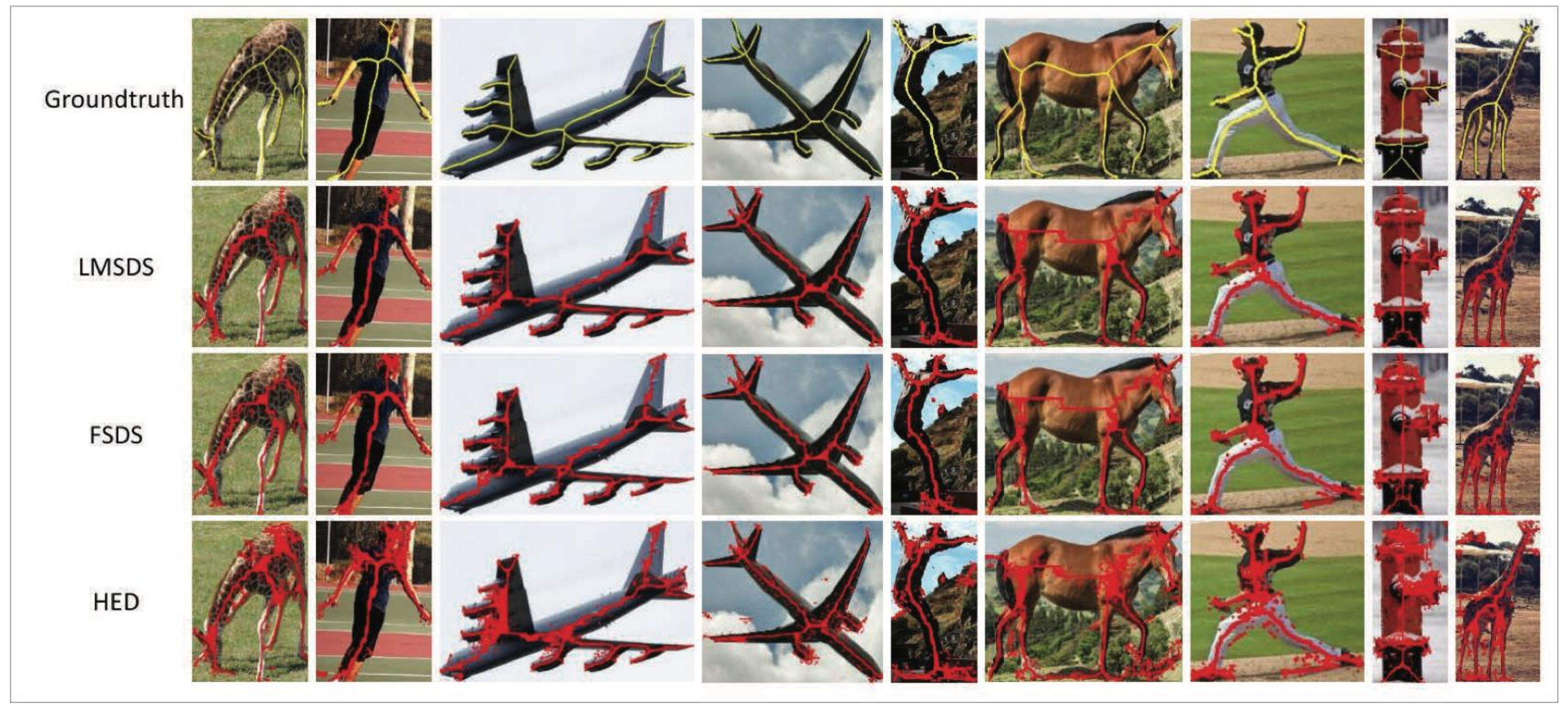

Style Transfer

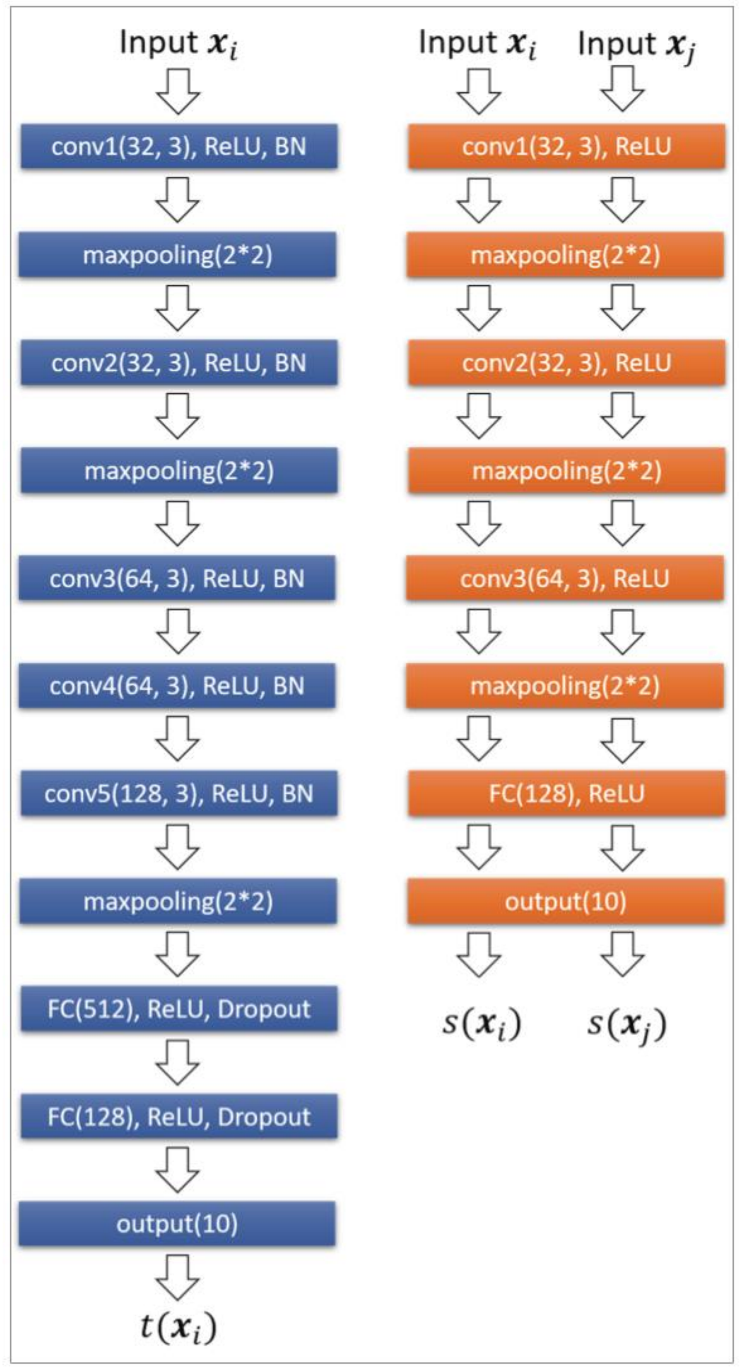
# Training Data

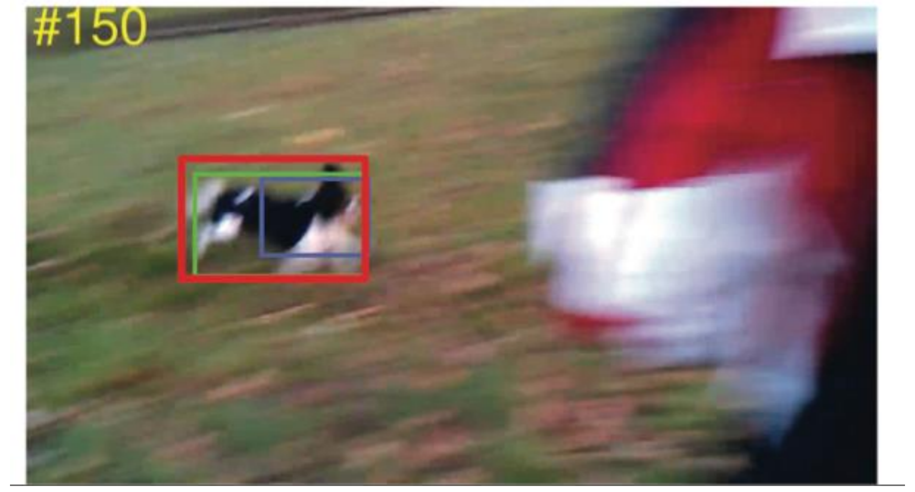
# Computer Vision Figures Dataset (CVFD)

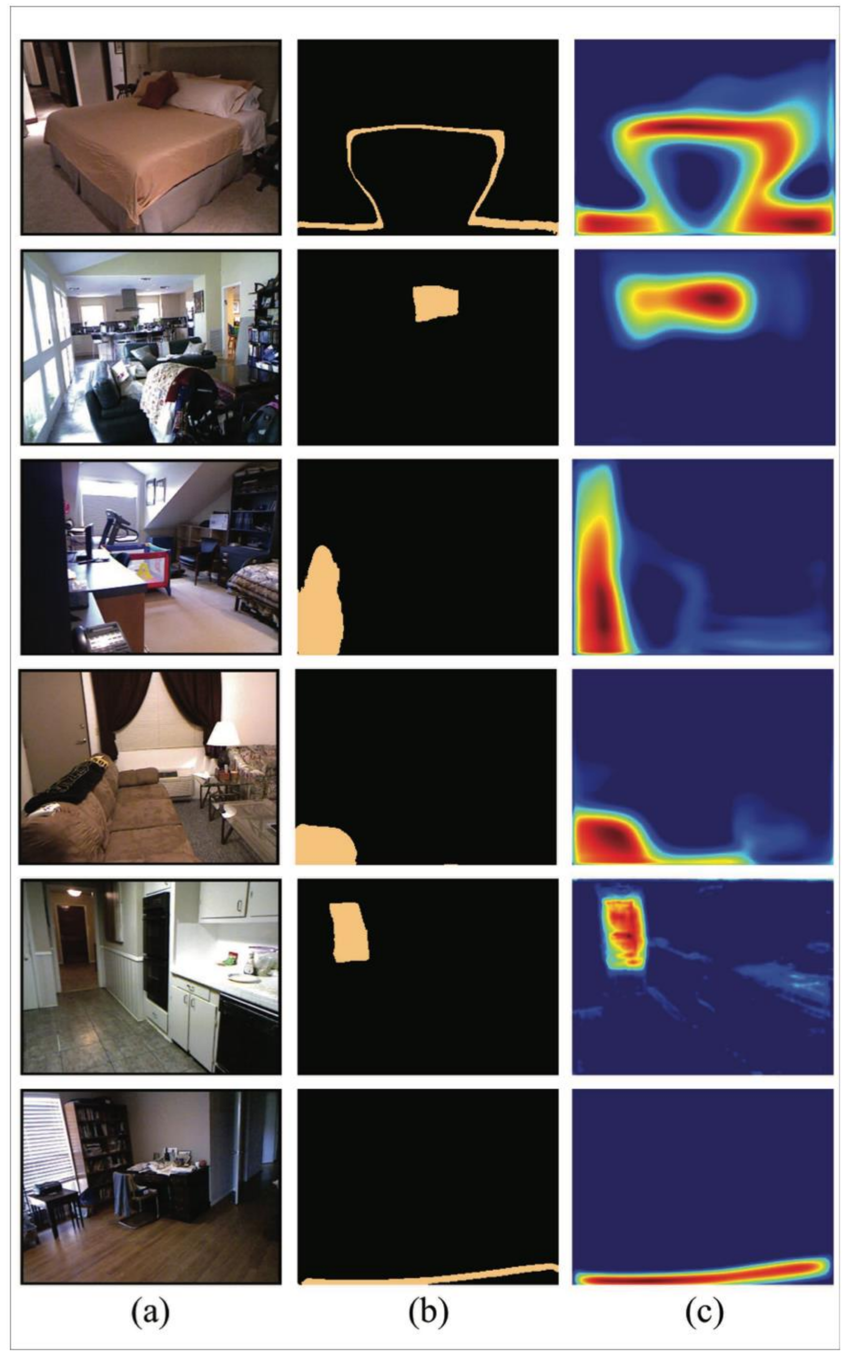




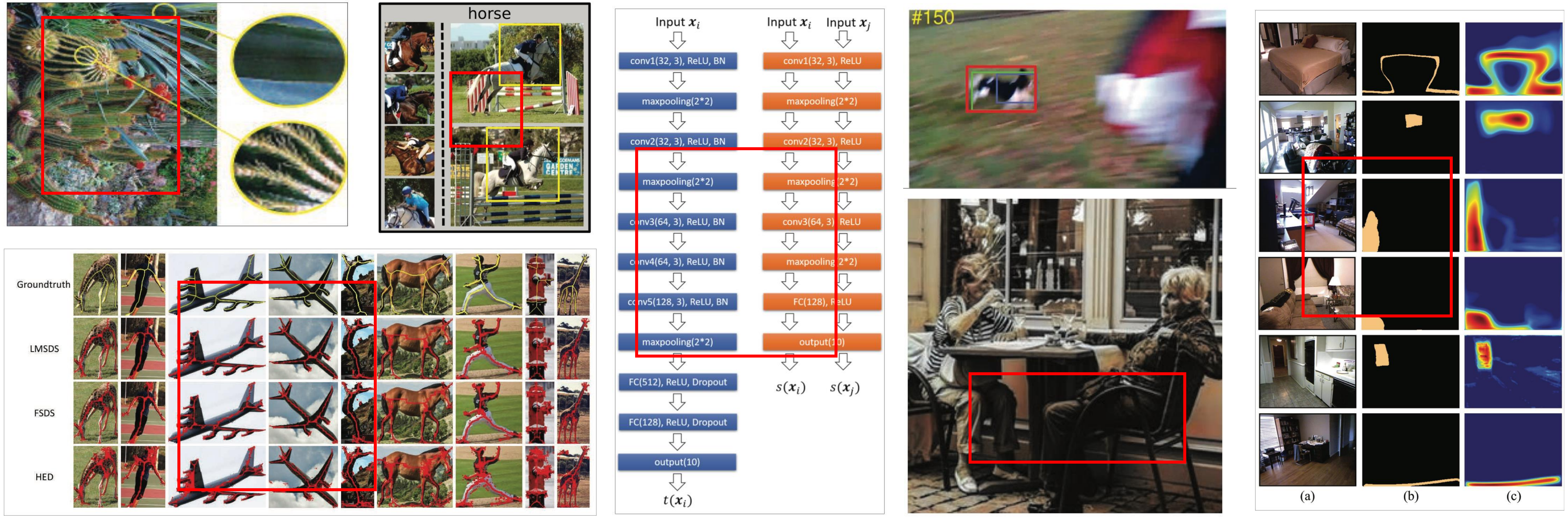




- 88k images from cs.CV arxiv papers (2010 to 2022)
- Many figures have grid-like structure

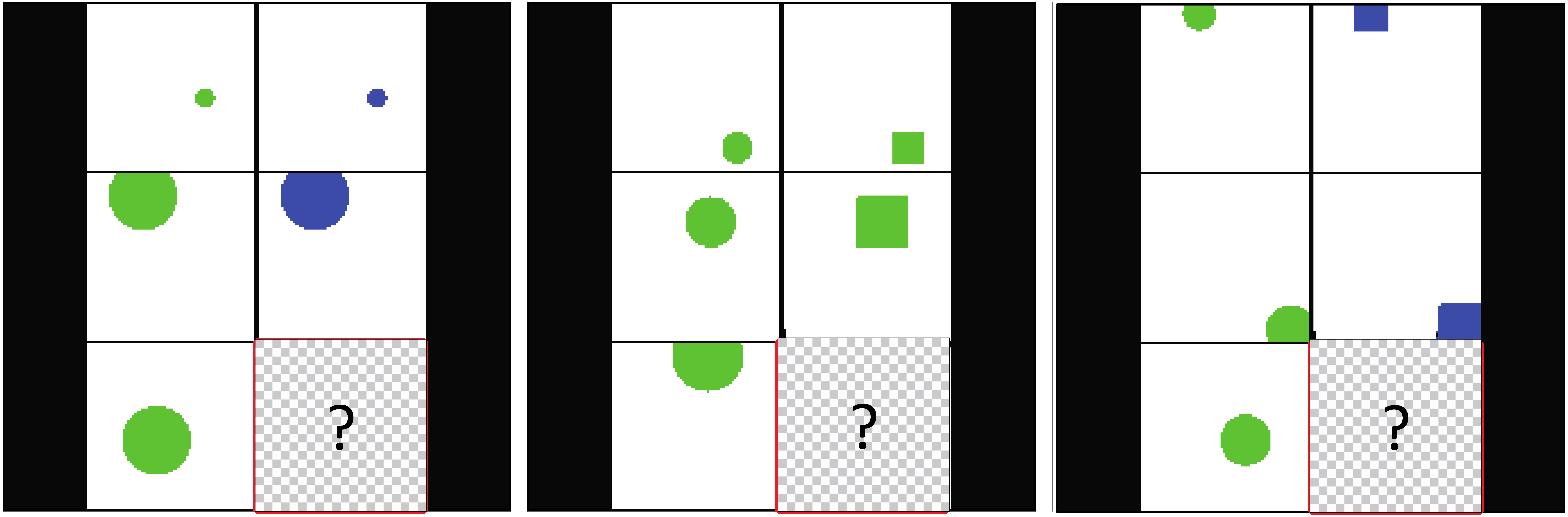
# Training Time




- Random 224x224 crops from figure images
  - No parsing
- We train MAE-VQGAN, a variant of Masked-Autoencoder (MAE)

He et al. , “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022.

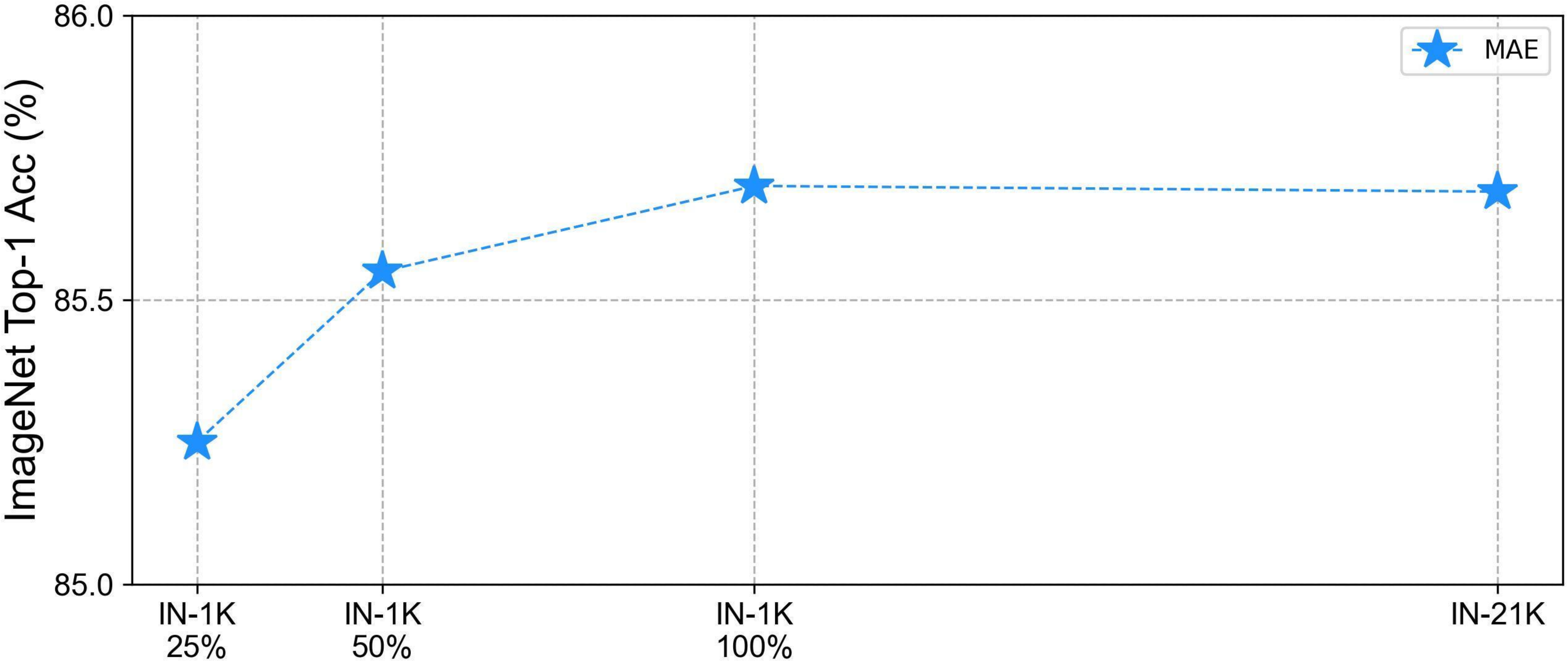
# Synthetic Experiments




Color Change Shape Change Color & Shape Change

# Various tasks

# But MAE doesn't seem to scale ☹



# Sequential Modeling Enables Scalable Learning for Large Vision Models

**Yutong Bai, Xinyang Geng**,
Karttikeya Mangalam, Amir Bar,
Alan Yuille, Trevor Darrell, Jitendra Malik, Alexei A Efros

On arxiv

# MAE -> LLM (but without language!)

**Data:** ImageNet

**Architecture:** Masked Autoencoders

**Loss function:** L2 regression loss

**Task Specification:** Finetune

# MAE -> LLM (but without language!)

**Data:** ~~ImageNet~~ 1.68B of images, 420B tokens, 50 Datasets

**Architecture:** Masked Autoencoders

**Loss function:** L2 regression loss

**Task Specification:** Finetune

# MAE -> LLM (but without language!)

**Data:** ~~ImageNet~~ 1.68B of images, 420B tokens, 50 Datasets

**Architecture:** ~~Masked Autoencoders~~ Autoregressive Model

**Loss function:** L2 regression loss

**Task Specification:** Finetune

# MAE -> LLM (but without language!)

**Data:** ~~ImageNet~~ 1.68B of images, 420B tokens, 50 Datasets

**Architecture:** ~~Masked Autoencoders~~ Autoregressive Model

**Loss function:** ~~L2 regression loss~~

**Task Specification:** Finetune

# MAE -> LLM (but without language!)

**Data:** ~~ImageNet~~ 1.68B of images, 420B tokens, 50 Datasets

**Architecture:** ~~Masked Autoencoders~~ Autoregressive Model

**Loss function:** ~~L2 regression loss~~ Cross Entropy for next token

**Task Specification:** Finetune

# MAE -> LLM (but without language!)

**Data:** ~~ImageNet~~ 1.68 B of images, 420 B tokens, 50 Datasets

**Architecture:** ~~Masked Autoencoders~~ Autoregressive Model

**Loss function:** ~~L2 regression loss~~ Cross Entropy for next token

**Task Specification:** ~~Finetune~~ prompting

across:
images,
videos,
supervised / unsupervised
synthetic /real,
all kinds of tasks
2D / 3D / 4D data etc.

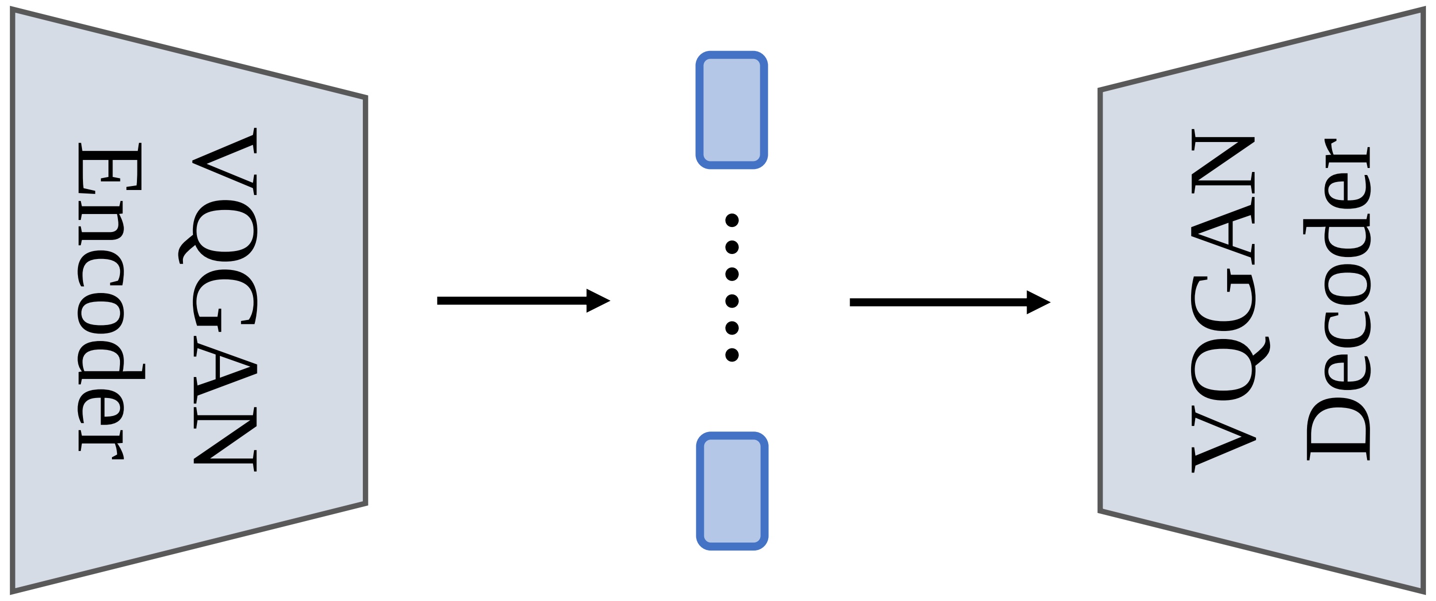
| Dataset | Tokens (Millions) | Annotation Type | Annotation Source |
|---|---|---|---|
| **Unpaired Image Data** | | | |
| LAION 5B [71] (1.5B images subset) | 380690 | - | - |
| **Images with Annotations** | | | |
| ImageNet 1K [25] | 1317.40 | Image Classification | Ground Truth |
| COCO [54] | 363 | Object Detection | MMDetection [16] |
| ADE 20K [100], Cityscapes [22] | 66.88 | Semantic Segmentation | Ground Truth |
| COCO [54], ImageNet 1K [25] | 2078.06 | Semantic Segmentation | Mask2Former [19] |
| COCO [54], lvmhp [51], mpii [4], Unite [49] | 950.79 | Human Pose | MMPose[21] |
| COCO [54], ImageNet 1K [25] | 1623.85 | Depth Map Image | DPT [67] |
| Subset of InstructPix2Pix [34] | 415.46 | Style Transfer | InstructPix2Pix [34] |
| COCO[54], ImageNet 1K[25] | 1623.85 | Surface Normal Image | NLL-AngMF [7] |
| COCO [54], ImageNet 1K [25] | 1623.85 | Edge Detection | DexiNed [79] |
| DID-MDN [98] | 35.06 | Rainy and Clean Image Pairs | Ground Truth |
| SIDD [3] | 245.76 | Denoised Image | Ground Truth |
| LOL[89] | 0.458 | Light Enhanced Image | Ground Truth |
| ImageNet 1K [25] | 1321.07 | Grayscale and Colorized Image Pairs | Ground Truth |
| ImageNet 1K [25] | 1321.07 | Inpainting | Ground Truth |
| Kitti [34] | 9.21 | Stereo | Ground Truth |
| **Videos** | | | |
| UCF101 [78] | 109.11 | - | - |
| DAVIS [65] | 0.36 | - | - |
| HMDB [48] | 55.41 | - | - |
| ActivityNet [13] | 380.63 | - | - |
| Moments in Time [59] | 2979.00 | - | - |
| Multi-moments in Time [60] | 4124.04 | - | - |
| Co3D [69] | 228.75 | - | - |
| Charades v1 [76] | 241.53 | - | - |
| Something-something v2 [37] | 904.57 | - | - |
| YouCook [23] | 3.14 | - | - |
| Kinetics 700 [14] | 7092.04 | - | - |
| MSR-VTT [92] | 57.34 | - | - |
| Youtube VOS [93] | 63.70 | - | - |
| jester [57] | 606.47 | - | - |
| diving48 [52] | 150.73 | - | - |
| MultiSports [53] | 78.44 | - | - |
| CharadesEgo [77] | 193.06 | - | - |
| AVA [61] | 117.96 | - | - |
| Ego4D [38] | 1152.12 | - | - |
| **Videos with Annotations** | | | |
| VIPSeg [58] | 64.47 | Video Panoptic Segmentation | Ground Truth |
| Hand14K [32] | 1.96 | Hand Segmentation | Ground Truth |
| AVA [61] | 122.88 | Video Detection | Ground Truth |
| JHMDB [43] | 19.00 | Optical Flow | Ground Truth |
| JHMDB [43] | 37.92 | Video Human Pose | Ground Truth |
| **Synthetic 3D Views** | | | |
| Objaverse [24] Rendered Multiviews | 217.85 | - | - |

Sentence -> Visual Sentence

# Single images

**<BOS>**  **<EOS>**

**Tokenizer**




tok1 tok2 tok3 •••

# Videos

<BOS>  ... <EOS>

# Image sequences

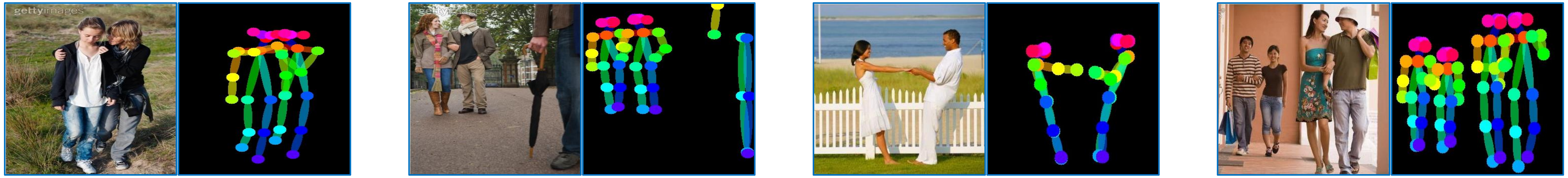
<BOS>  ••• <EOS>

# categories

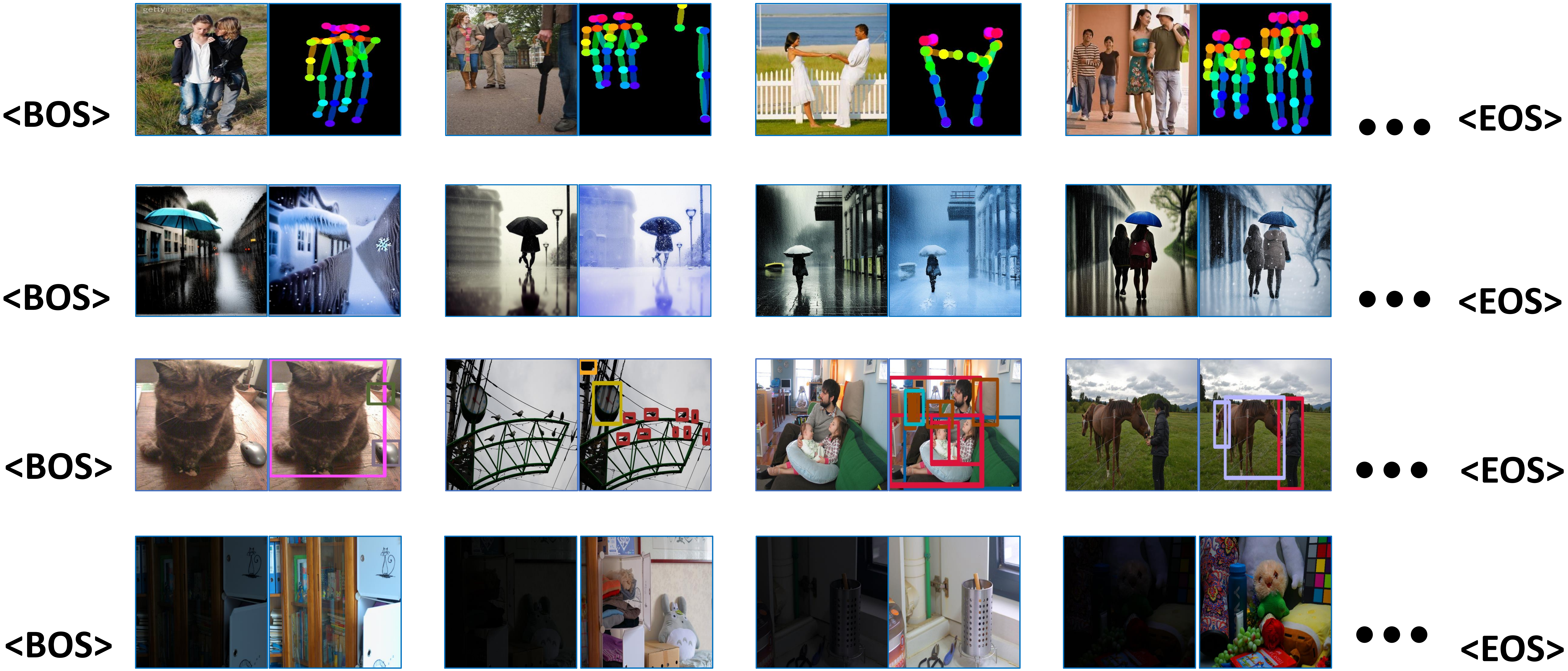
**<BOS>**


**••• <EOS>**

# Images with annotation

<BOS>  ... <EOS>

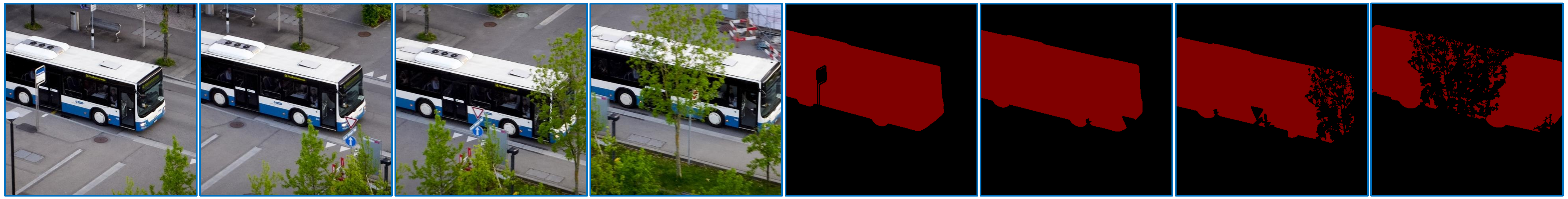
# Images with annotation



# Images with free form annotation

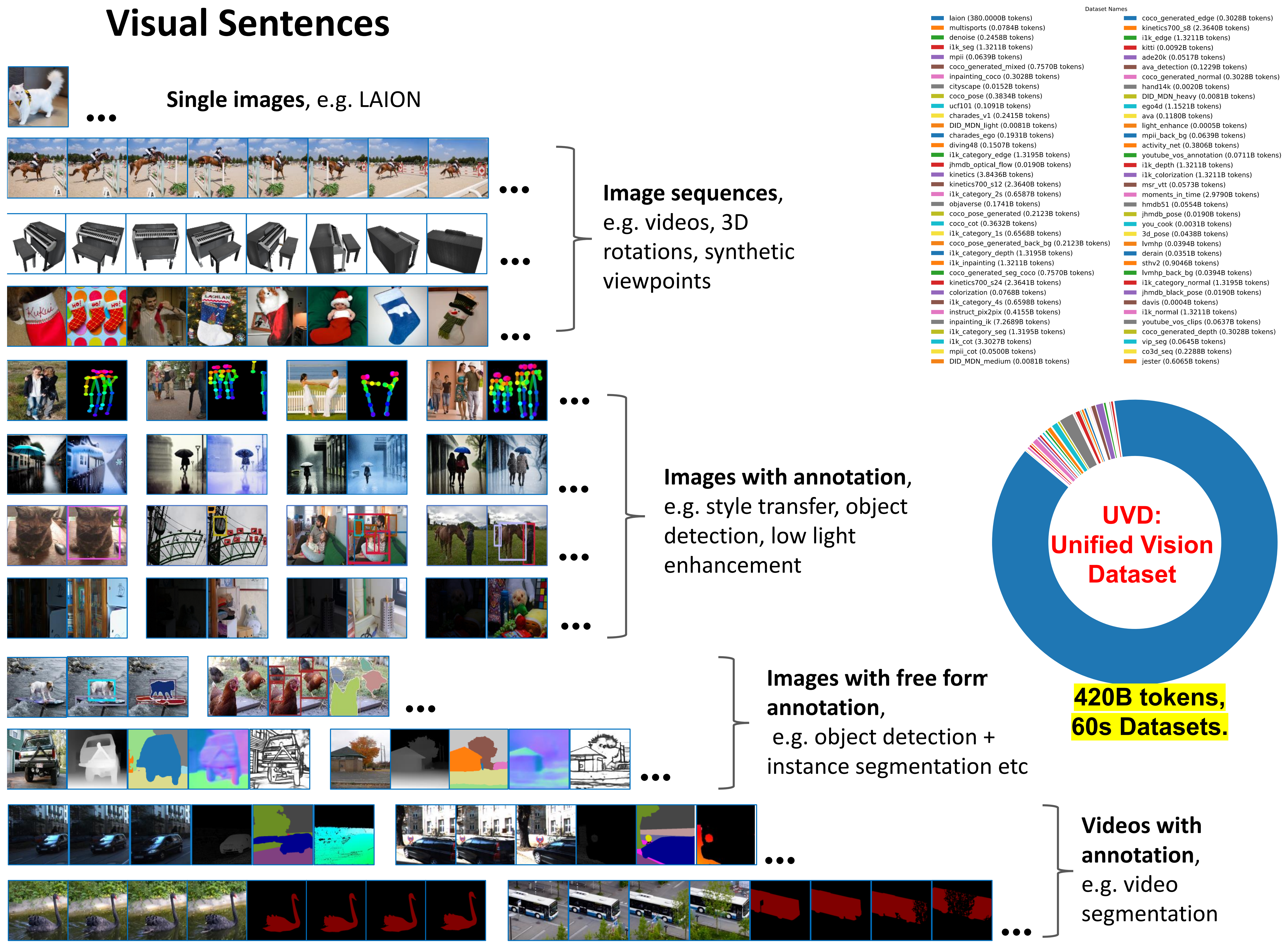
# Videos with annotation

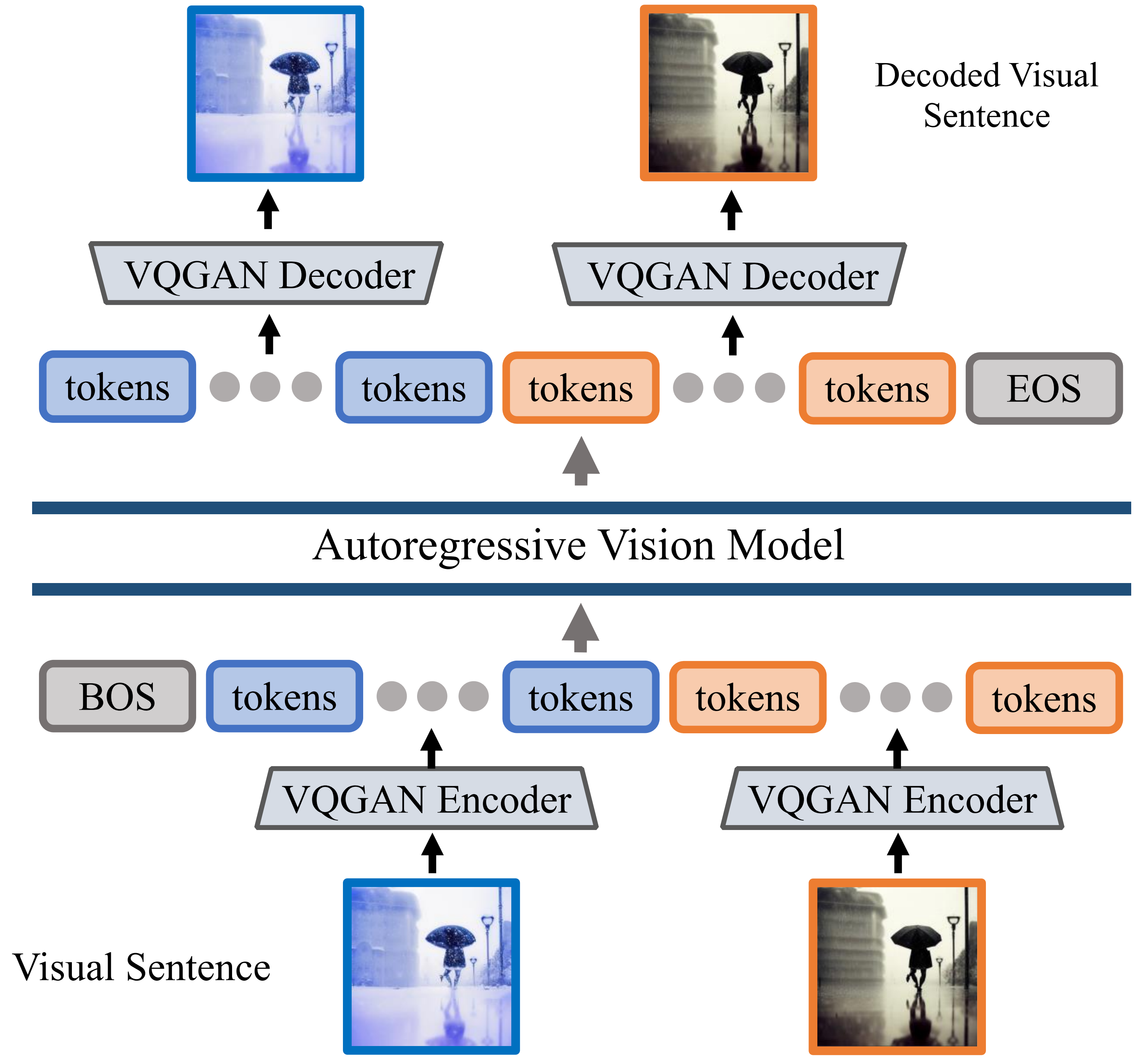
<BOS>



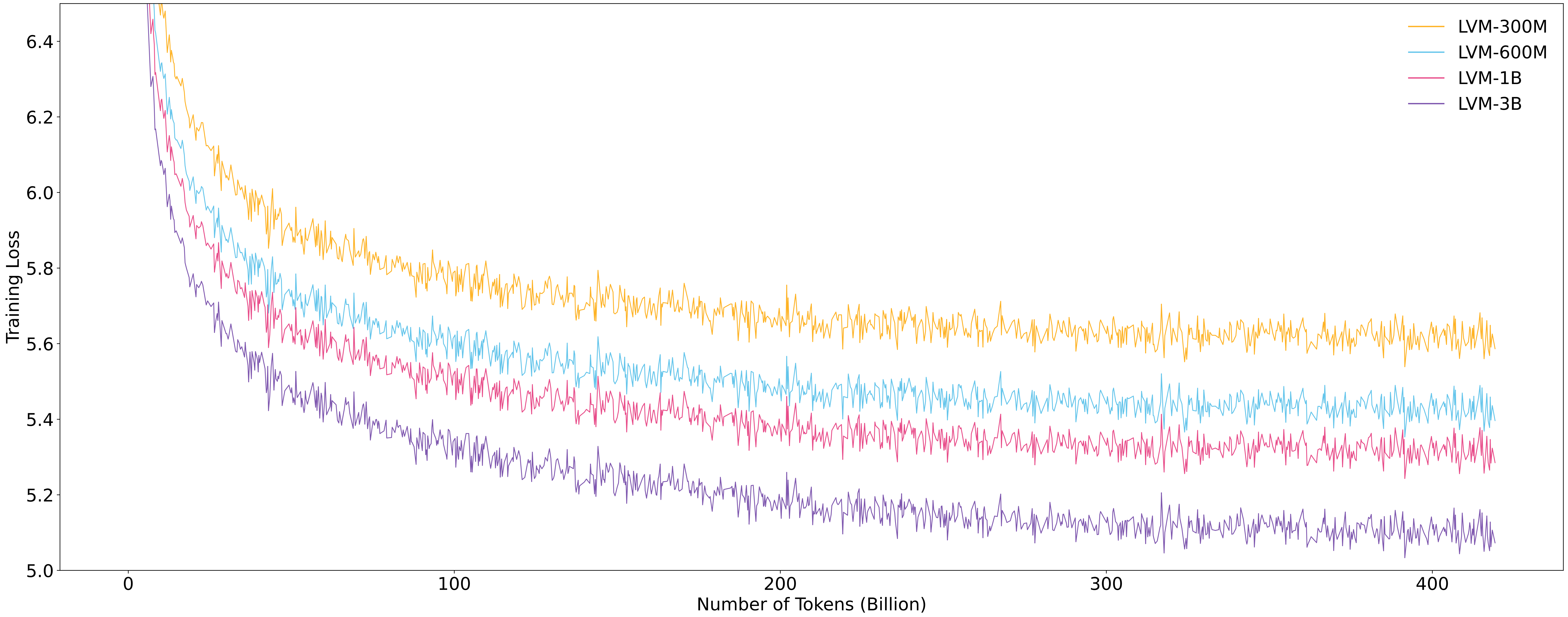
<EOS>

# LVM: Large Vision Model



# Training Loss (1 epoch) ~ Validation Loss



# Sequential Prompting

**Prompts**

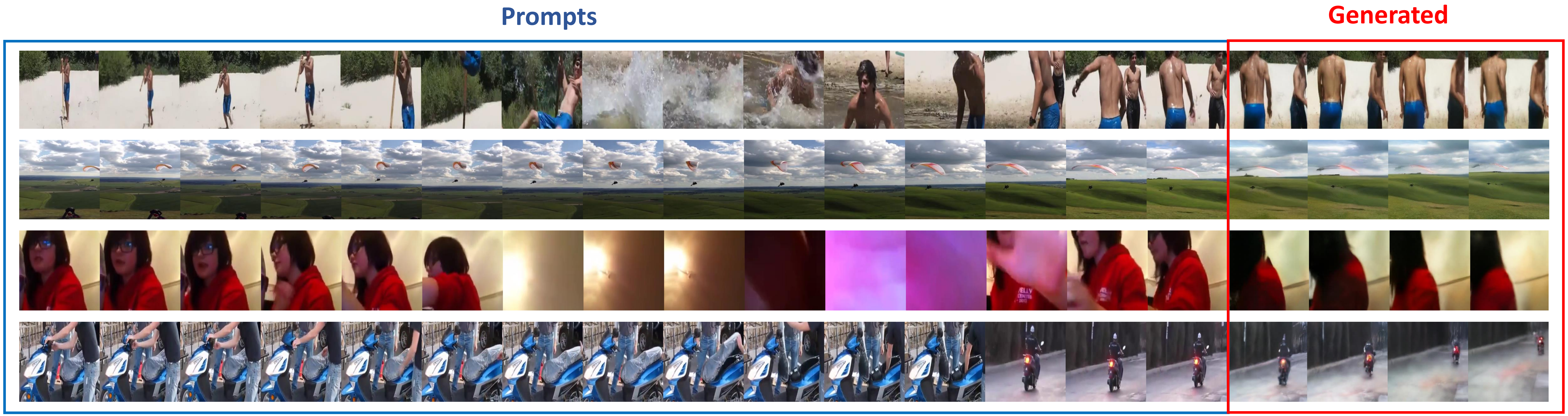
# Sequential Prompting

# Sequential Prompting

# Sequential Prompting

# Longer contexts

# Sequential Prompting



# Sequential Prompting

# Analogy Prompting

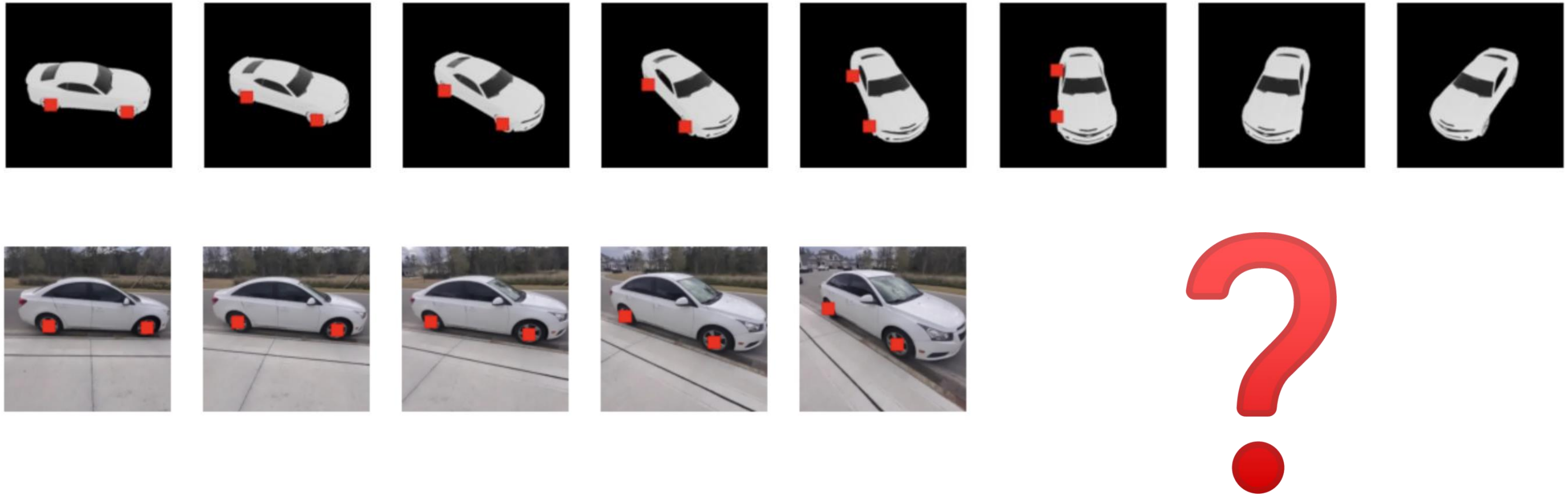
# Analogy Prompting

Prompts
Generated

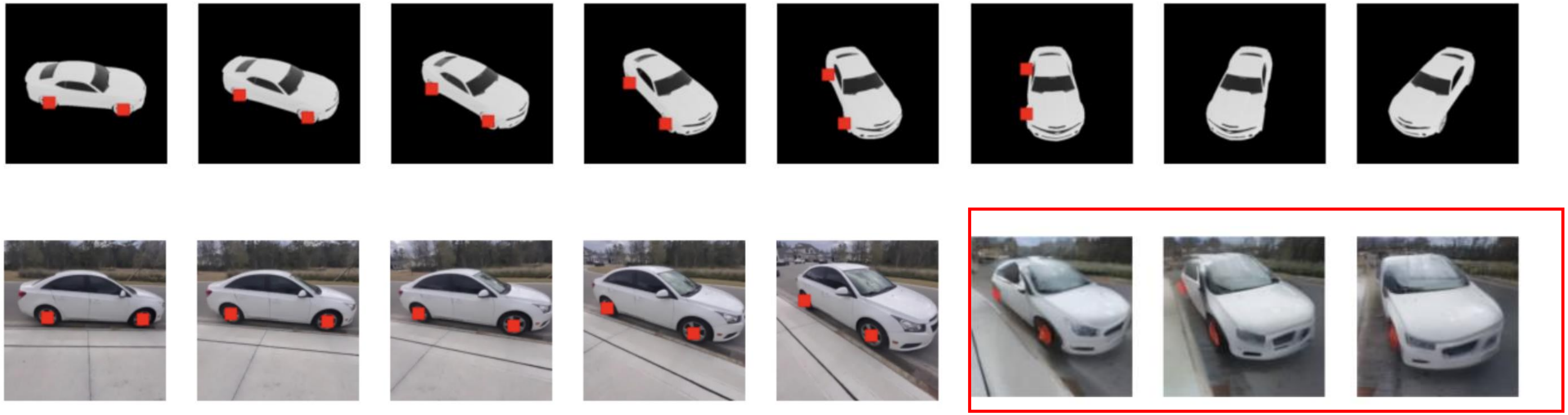
# More complicated

# More complicated

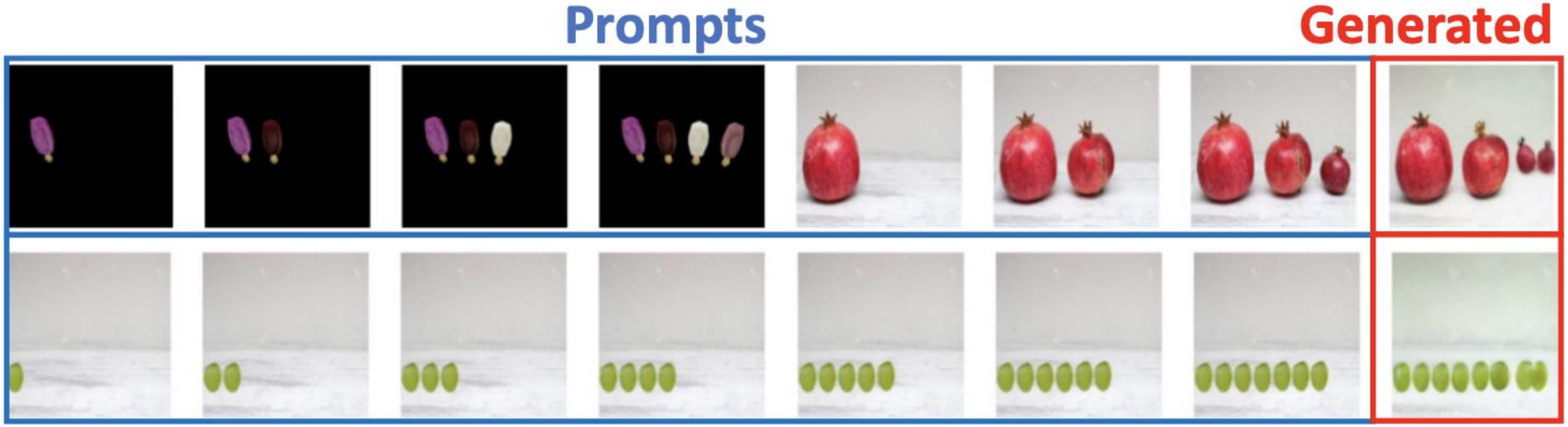
# More complicated



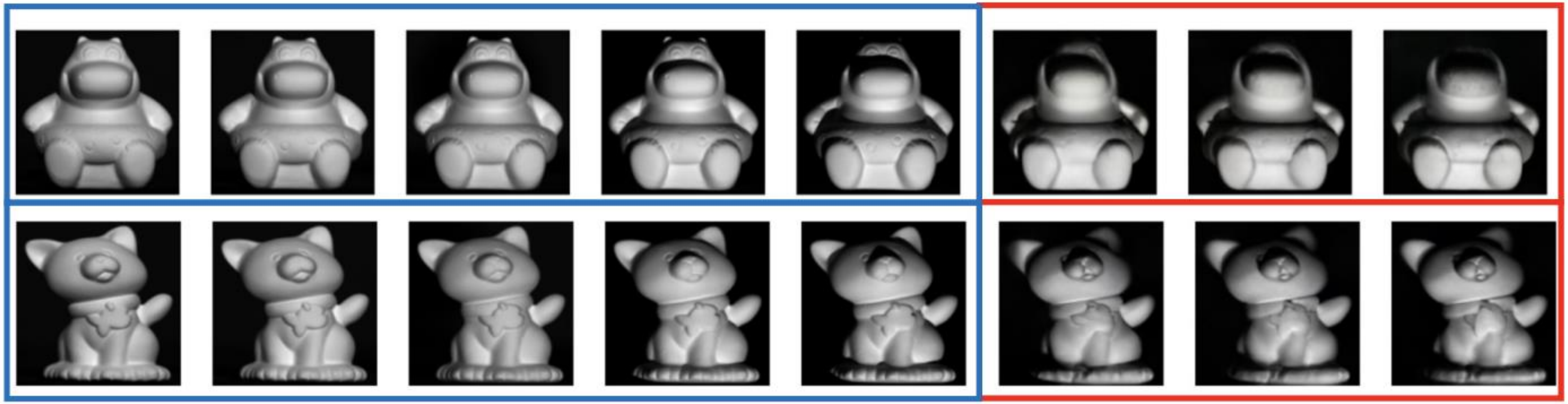
# More complicated



# Unseen tasks



# Unseen tasks



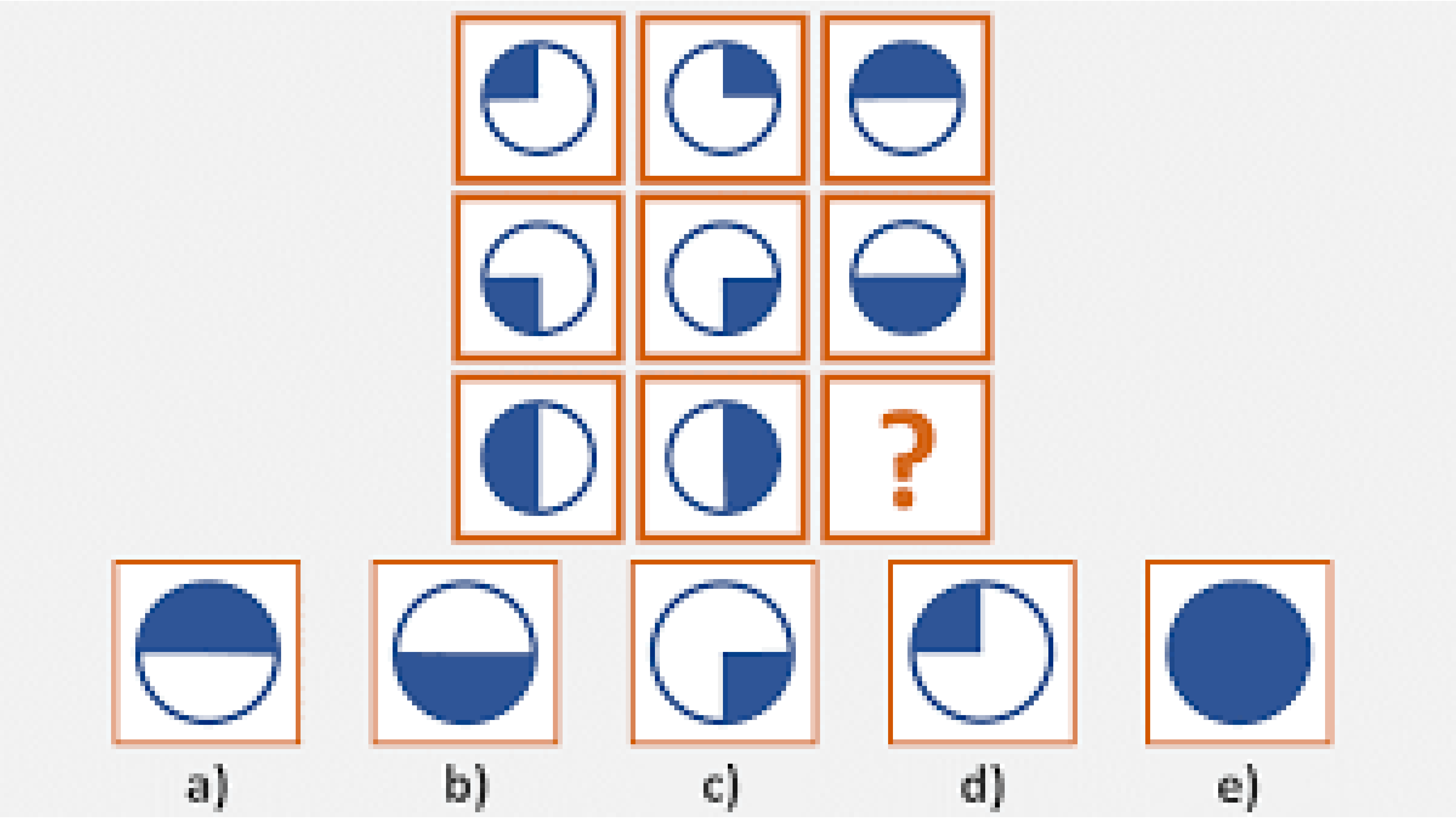
# Not easily describable

# Not easily describable

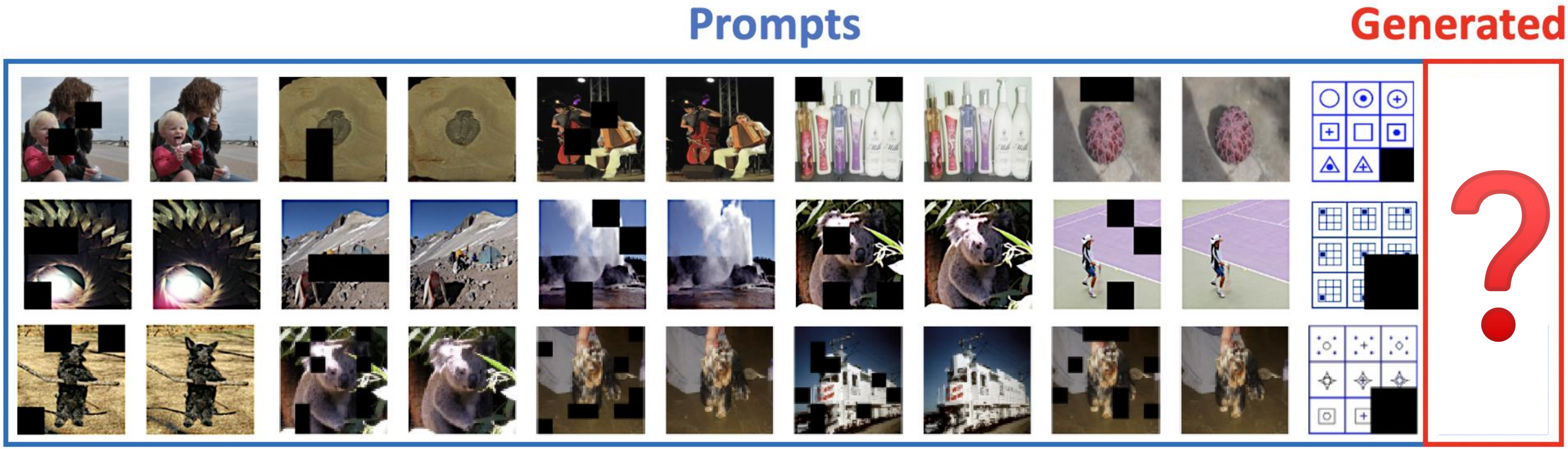
# Not easily describable

# Not easily describable

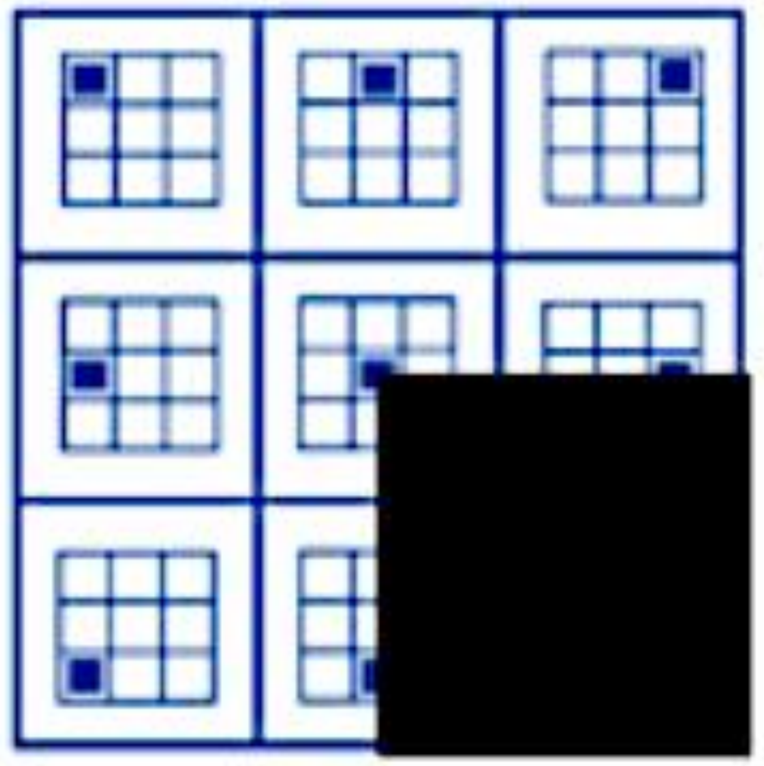
# Raven’s Progressive Test (Non-verbal IQ test)

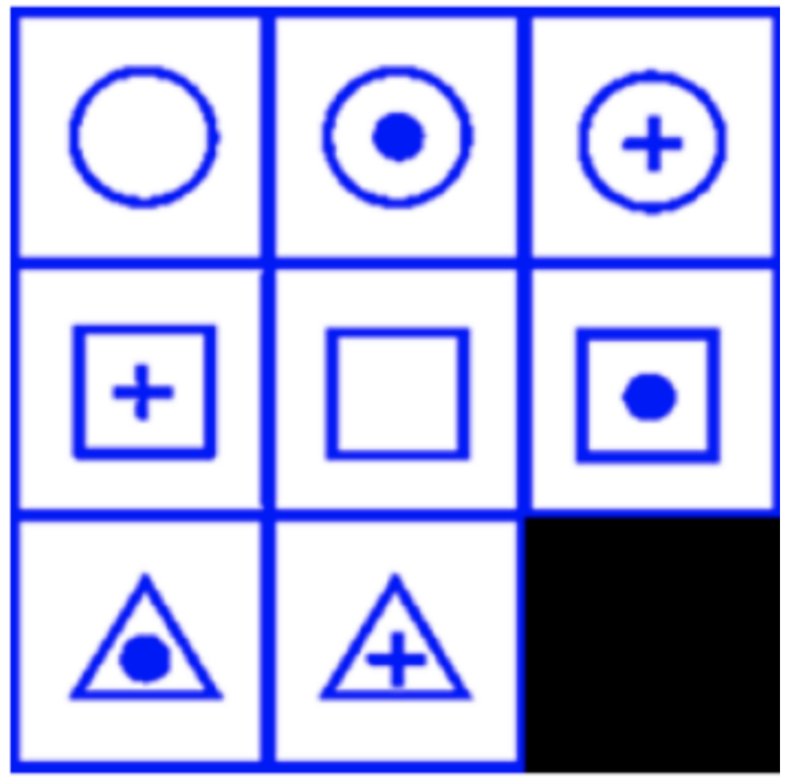


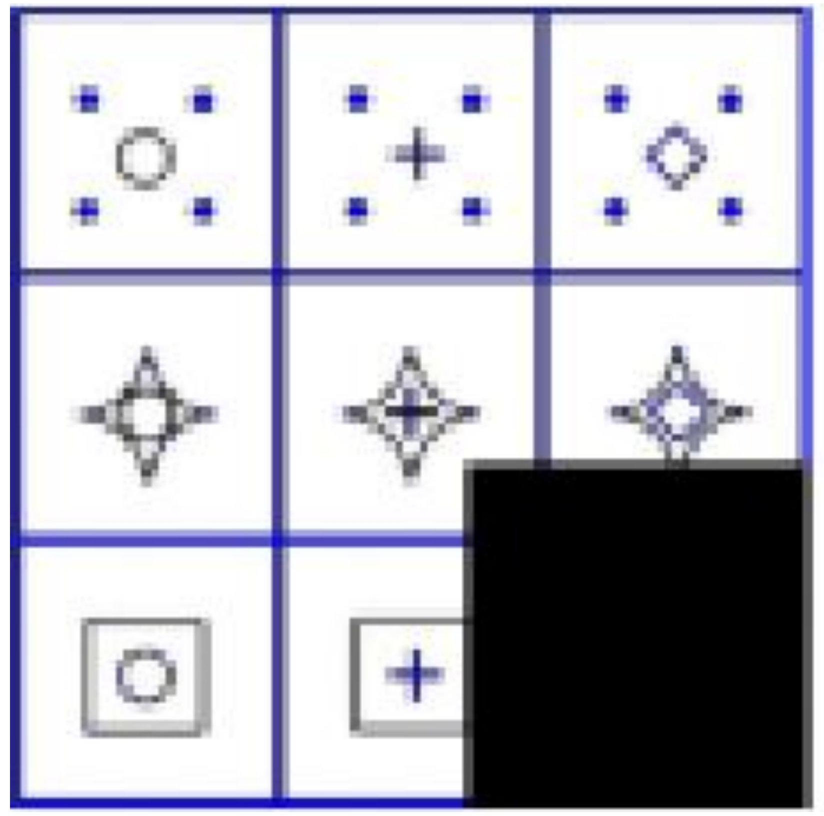

Raven, J. C. (1936)


Prompts
Generated

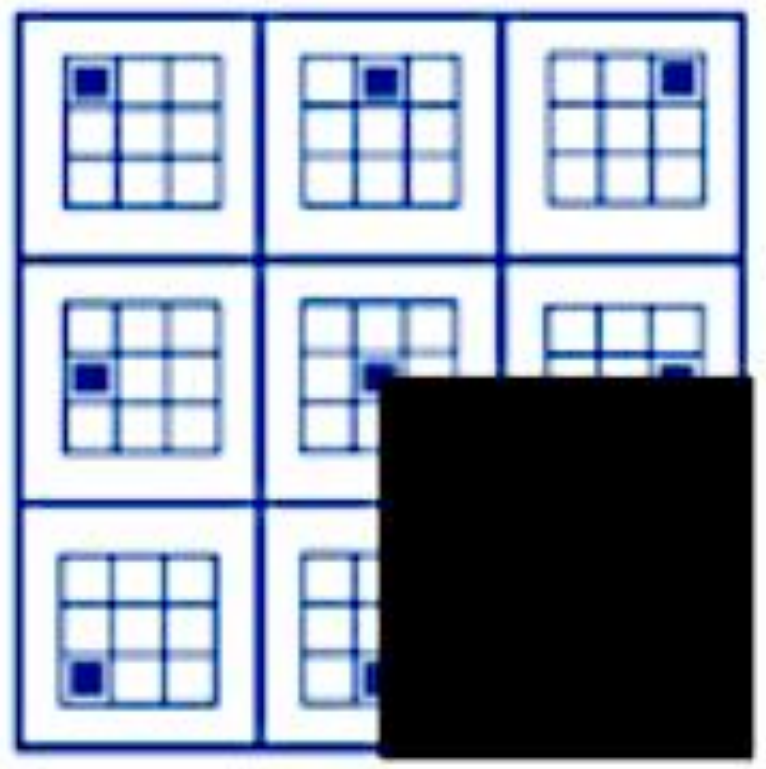
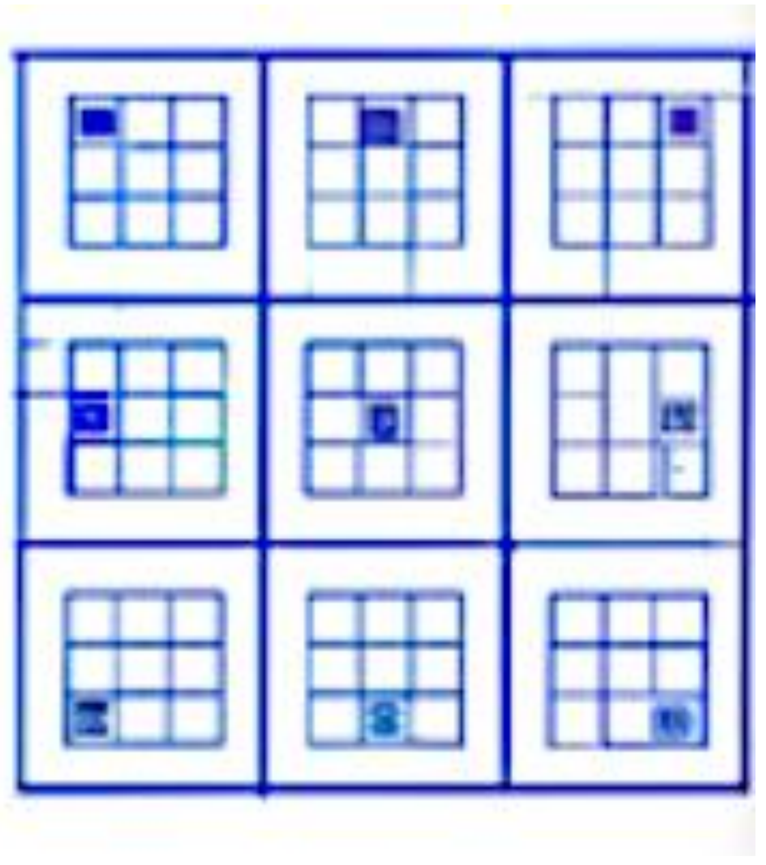
# Non-verbal IQ test

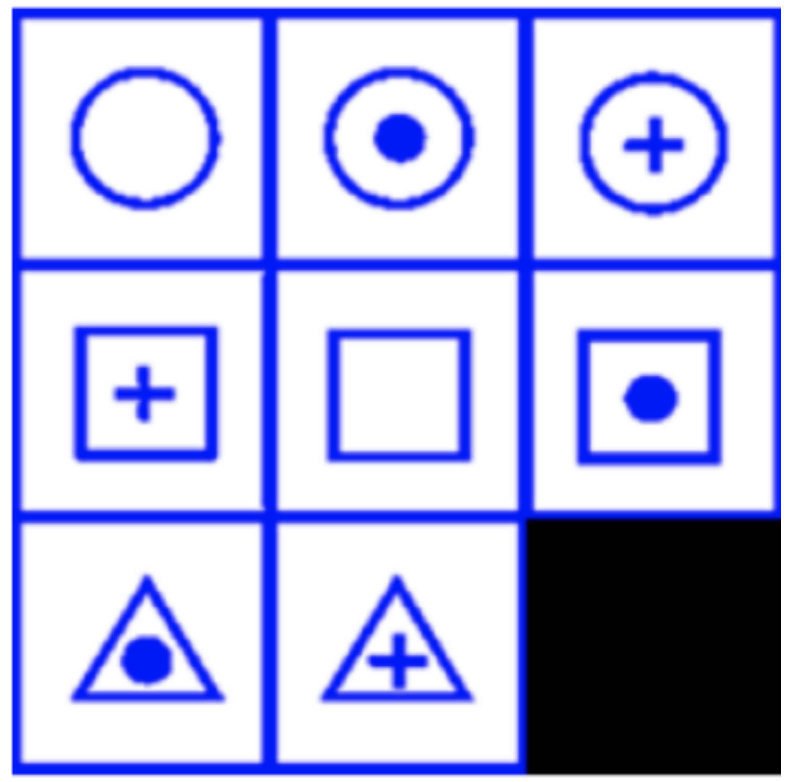
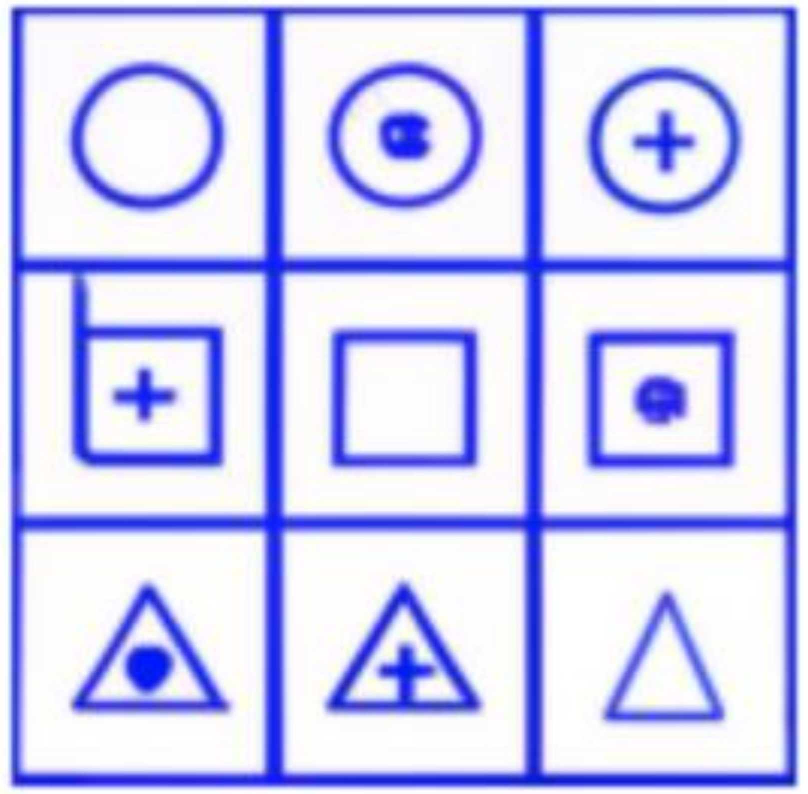


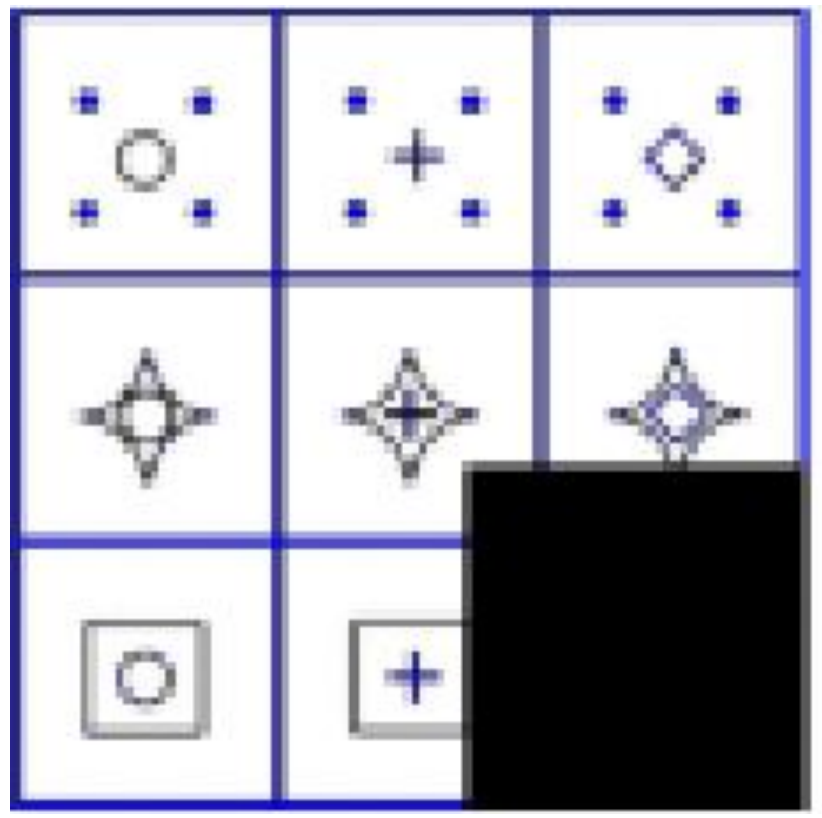
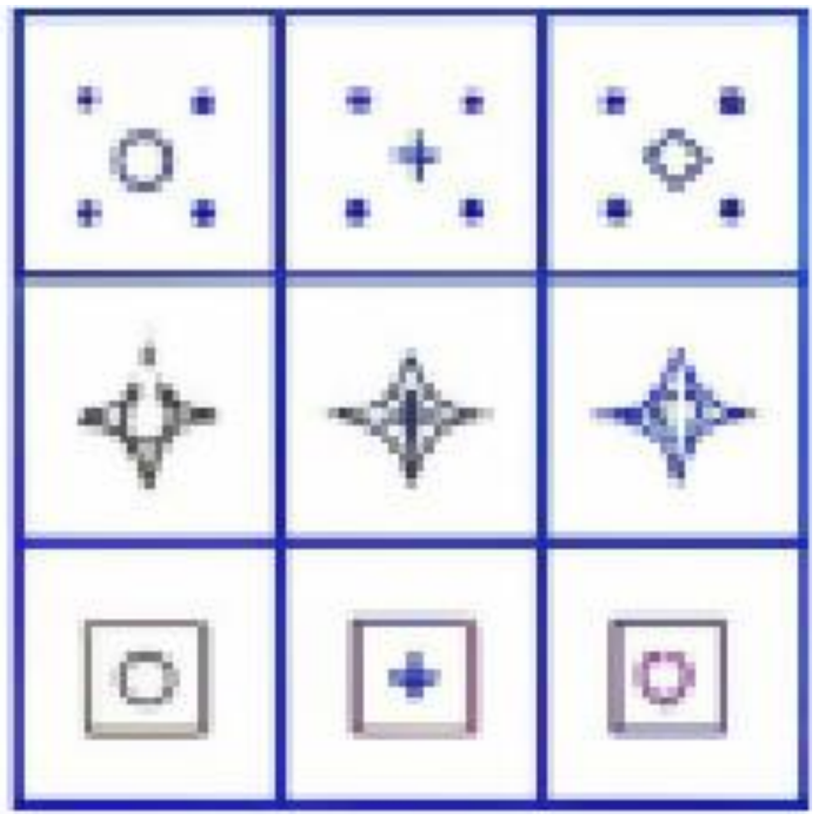




# Non-verbal IQ test

# Summary

- The dream of self-supervising ourselves with **natural world data** (rather than text) is alive and well!
- But we are still at the beginning of the journey
- Plenty left to do!

# Thank you