Unsupervised Learning

Computer Vision – Lecture 18

Further Reading

- Slides from S Savarese, A Zamir
- Slides from F Li
- Slides from Y Asano

Basics: Supervised Learning

Dataset $D = \{(x_i, y_i) | 1 \leq i \leq N\}$

Inputs x_i

Outputs y_i

Training/Validation/Testing $D = D_T \cup D_V \cup D_*$

Learn $f(x) = y$ by minimizing $\sum_{(x_i,y_i)\in D_T} \mathcal{L}(f(x_i), y_i)$

Hope to generalise: $\sum_{(x_i,y_i)\in D_*} \mathcal{L}(f(x_i), y_i)$

Basics: Unsupervised Learning

Dataset $D = \{x_i | 1 \leq i \leq N\}$

Inputs x_i

~~Outputs y_i~~

Training/Validation~~/Testing~~ $D = D_T \cup D_V$

Testing on downstream task $T = \{(\chi_i, v_i) | 1 \leq i \leq M\}$

Learn $f(x) =$? by minimizing ??

Hope to generalise to another task $\sum_{(\chi_i, v_i) \in T} \mathcal{L}(f(\chi_i), v_i)$

What we need to do

- Trick the model to learn the downstream task without labels

- Build inductive biases into the model
- Find a learning signal for the training scheme
- Often: prevent trivial solutions and cheating

Learning Signal Toolbox

- Learning signals are general ideas to incorporate priors
- Priors model general assumptions about the world/task

This lecture contains a wide set of tools for learning from priors

- Can be used in all settings un/weakly/fully-supervised
- Not specific to tasks, apply to many areas
- For free*! (no annotations needed)

*you still need to cite some papers

Learning Signal Toolbox

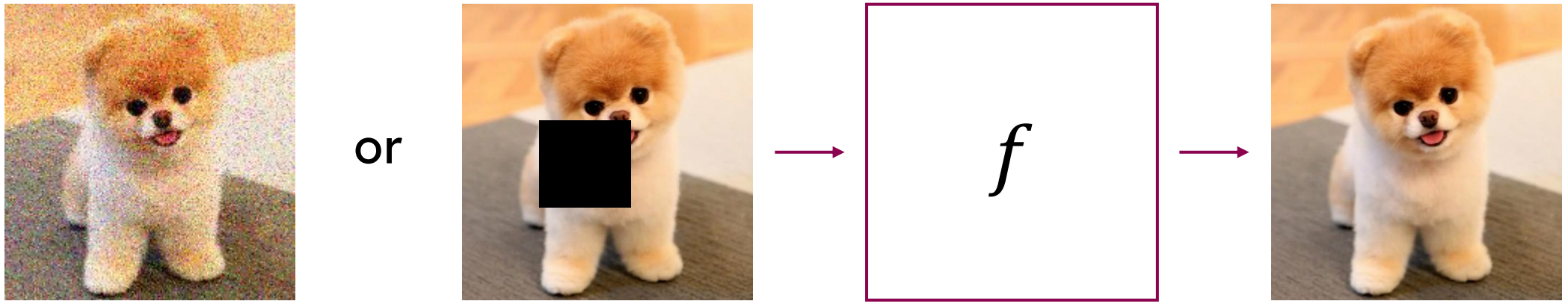
- Recovery: $f\big(M(x)\big)$ <-> x
- Bottleneck: $f\big(g(x)\big)$ <-> x with restriction on $g(x)$
- Dataset: $f(x_1)$ <-> $f(x_2)$
- Invariance: $f\big(\pi(x)\big)$ <-> $f(x)$
- Equivariance: $f\big(\pi(x)\big)$ <-> $\pi'\big(f(x)\big)$ often $\pi \equiv \pi'$ but not always
- Transformation estimation: $f\big(\pi(x,\theta)\big)$ <-> θ
- Generative: $f(z)$ <-> D
- Task-specific: $f(x)$ <-> priors
- Uncertainty: $f(x)$ <-> own error
- Many more!

Learning Signal Toolbox

Reconstruction

Learning Signal Toolbox

Recovery

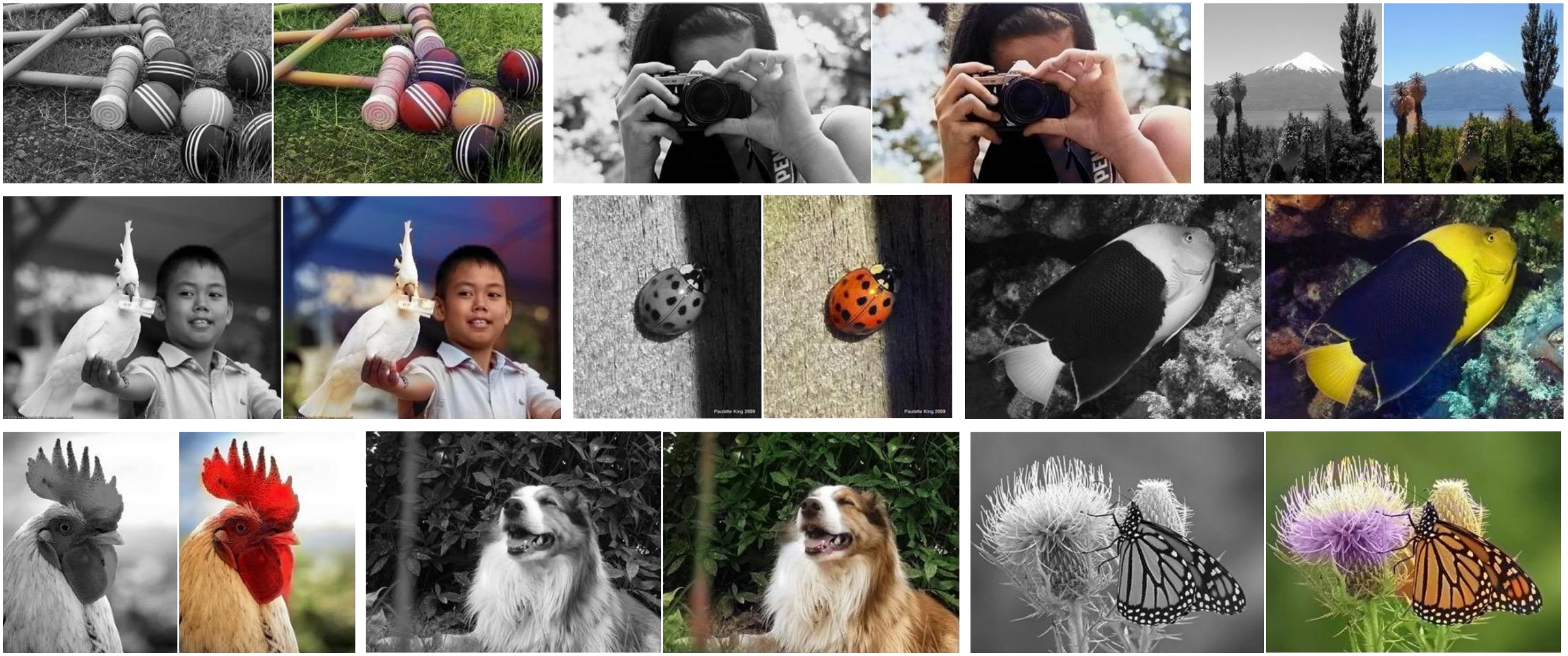


(Context autoencoder [Pathak et al.; CVPR '17],
denoising autoencoder [Vincent et al.; ICML '08],
Diffusion models [Sohl-Dickstein et al., ICML '15], etc.)

Rcovery: Inpainting

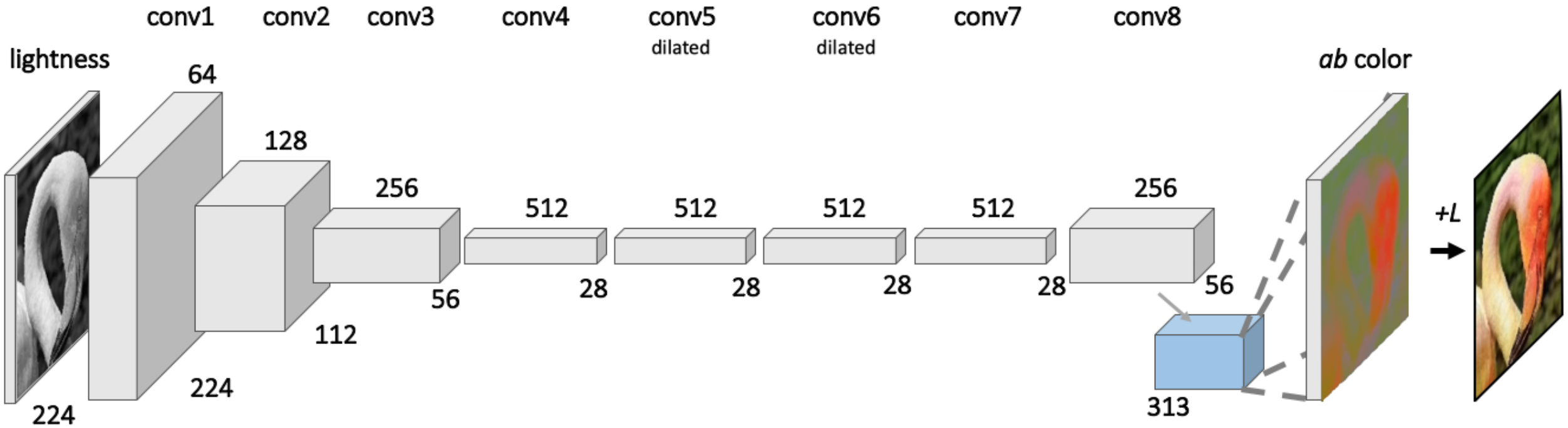
D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. Efros. Context Encoders: Feature Learning by Inpainting. CVPR 2016

Recovery: Colorization



R. Zhang, P. Isola, and A. Efros, Colorful Image Colorization, ECCV 2016

Colorization: Architecture



R. Zhang, P. Isola, and A. Efros, Colorful Image Colorization, ECCV 2016

Failure Cases

Source: A. Efros, R. Zhang

Inherent Ambiguity

Grayscale

Source: A. Efros, R. Zhang

Inherent Ambiguity

Prediction

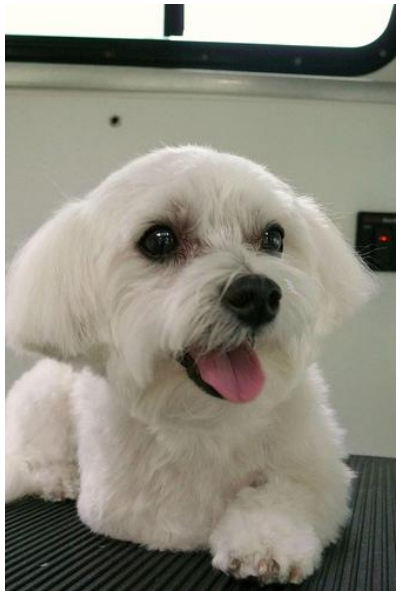
Ground Truth

Source: A. Efros, R. Zhang

Biases

Source: A. Efros, R. Zhang

Biases

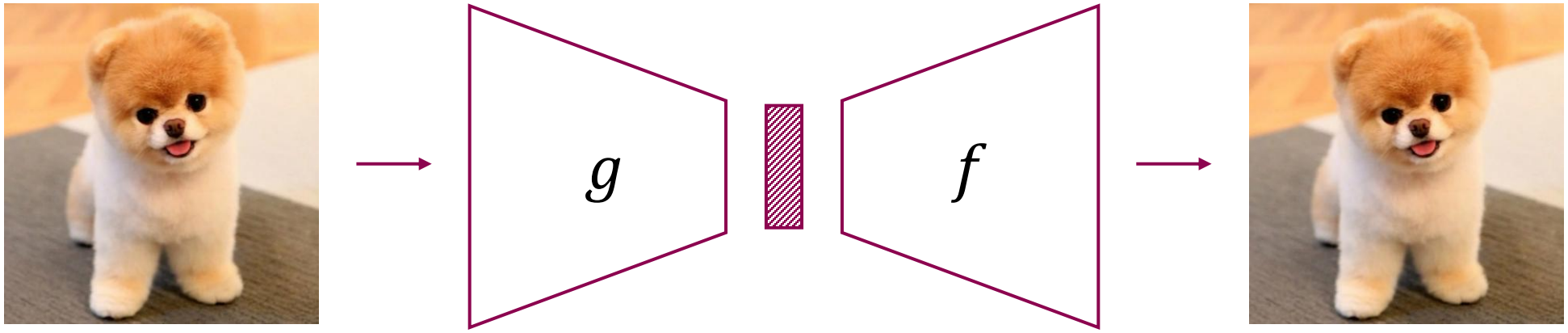




Source: A. Efros, R. Zhang

Learning Signal Toolbox

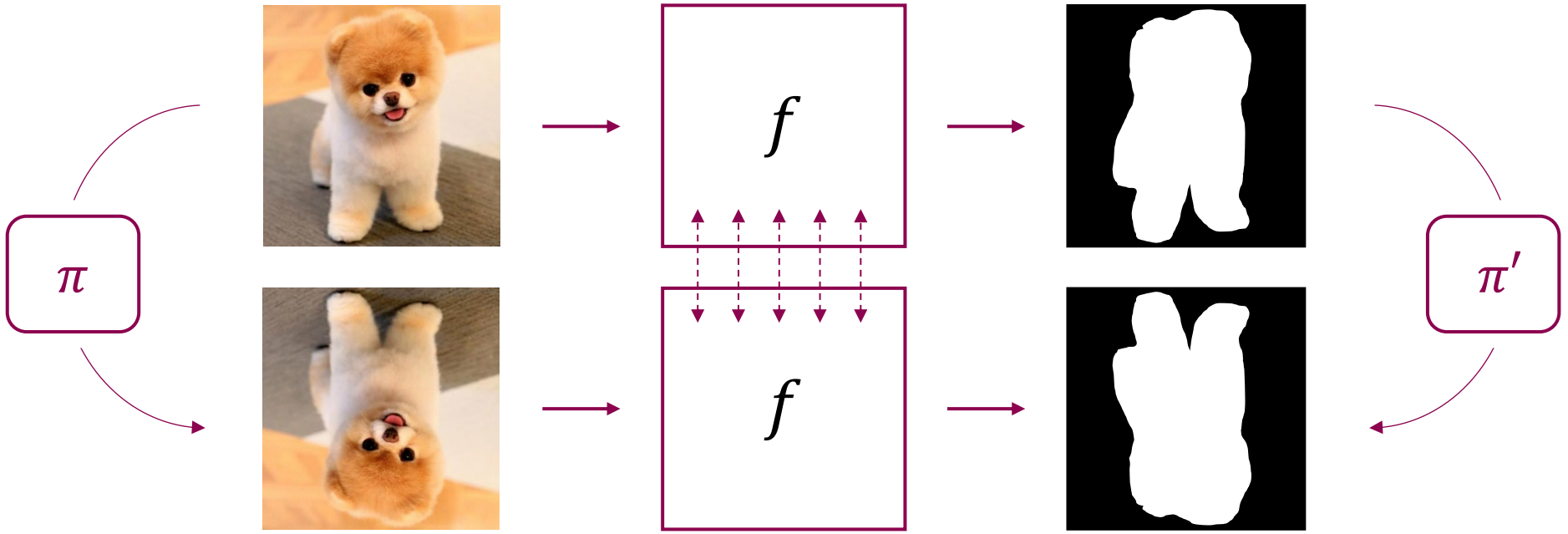
Bottleneck



Low dimensionality
Sparsity
Dictionary
Task-specific
…

Learning Signal Toolbox

Equivariance



$$\pi'^{-1}\Big(f\big(\pi(x)\big)\Big) = f(x)$$

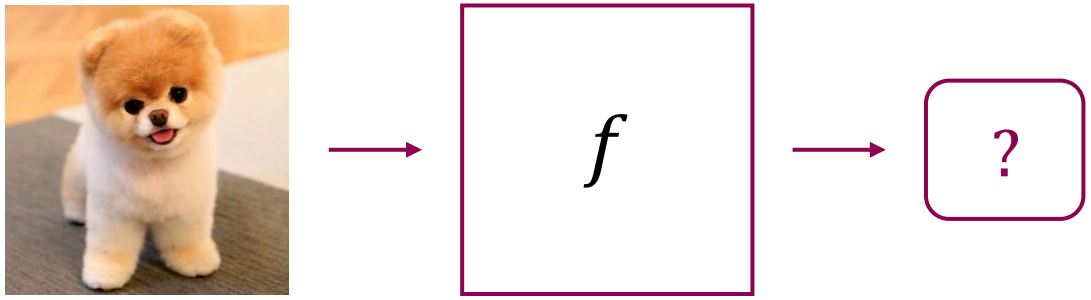
is often more difficult because inverse is hard

$$f\big(\pi(x)\big) = \pi'\big(f(x)\big)$$

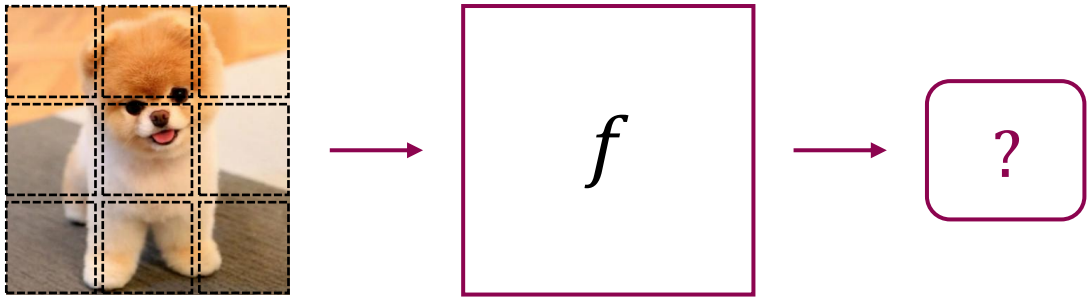
* Often $\pi \equiv \pi'$

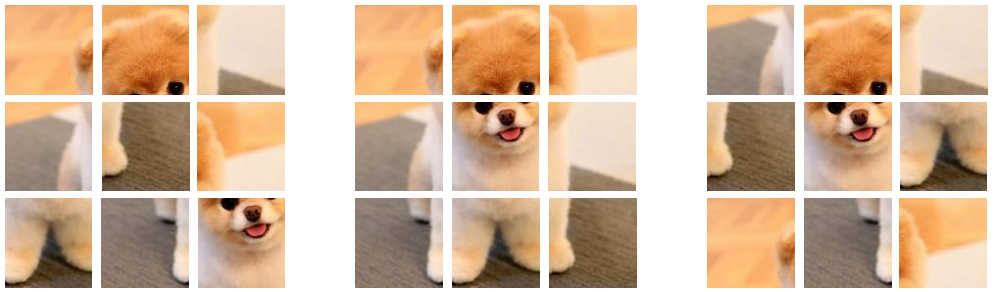
Learning Signal Toolbox

Transformation Estimation



Rotation [Gidaris et al.; ICLR ‘18]

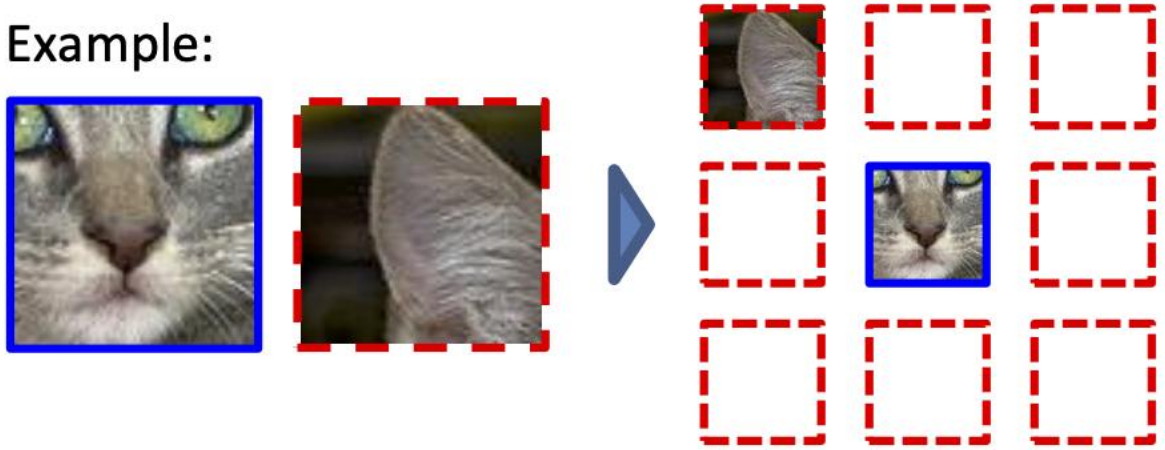




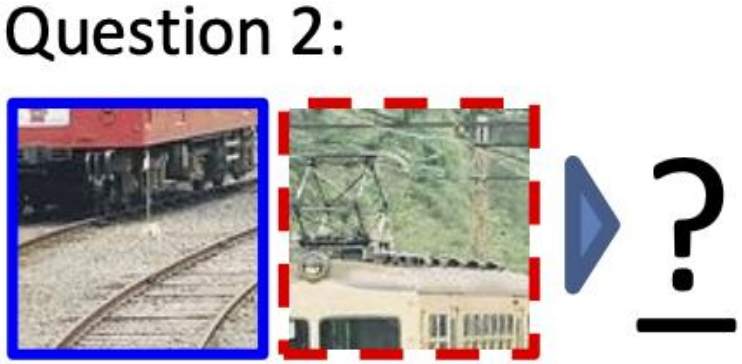
Jigsaw puzzle [Noroozi & Favaro; ECCV ‘16]

Context prediction

- *Pretext task*: randomly sample a patch and one of 8 neighbors
- Guess the spatial relationship between the patches



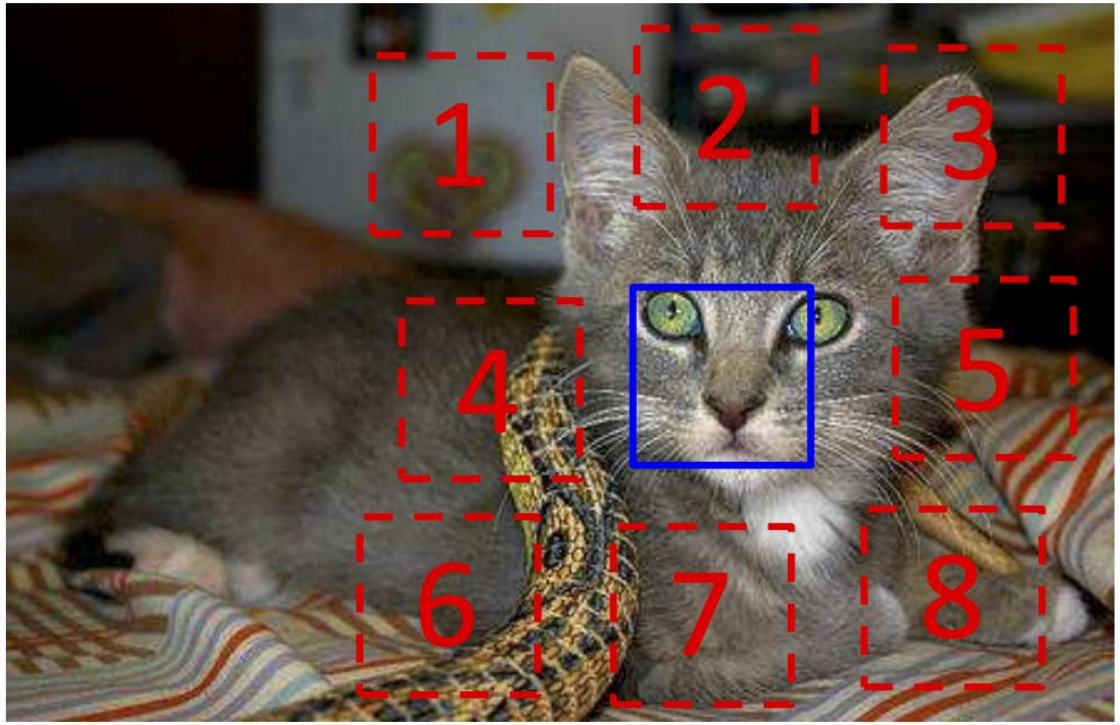
A: Bottom right



A: Top center

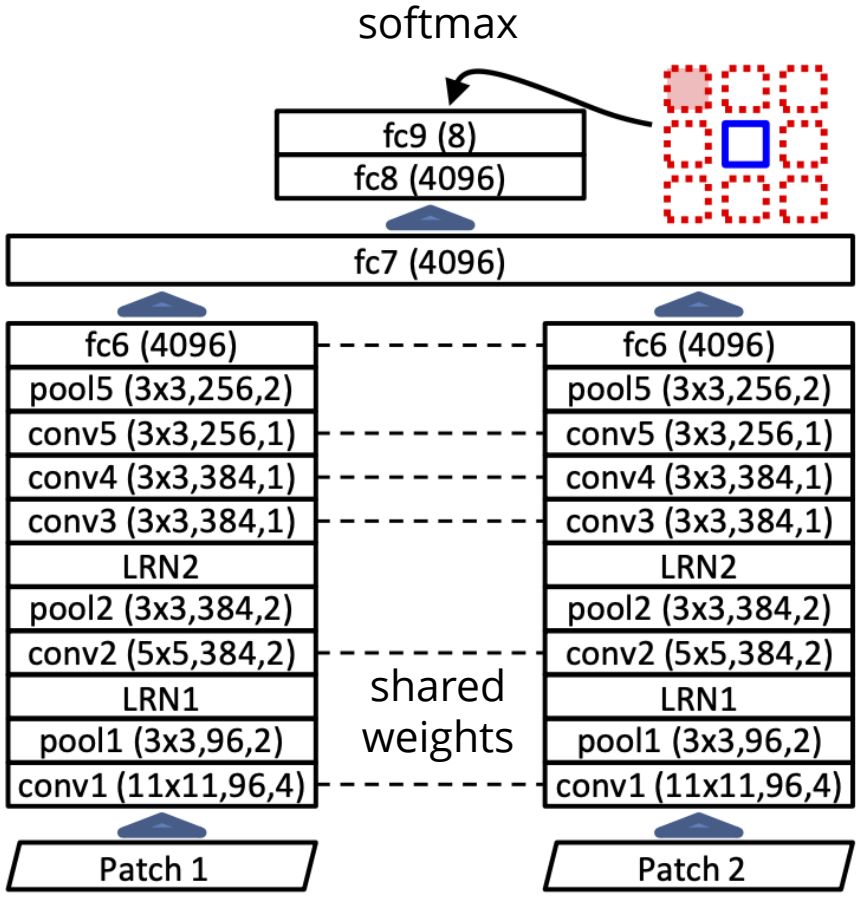
C. Doersch, A. Gupta, A. Efros. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Context prediction: Details



Prevent “cheating”: sample patches with gaps, pre-process to overcome chromatic aberration

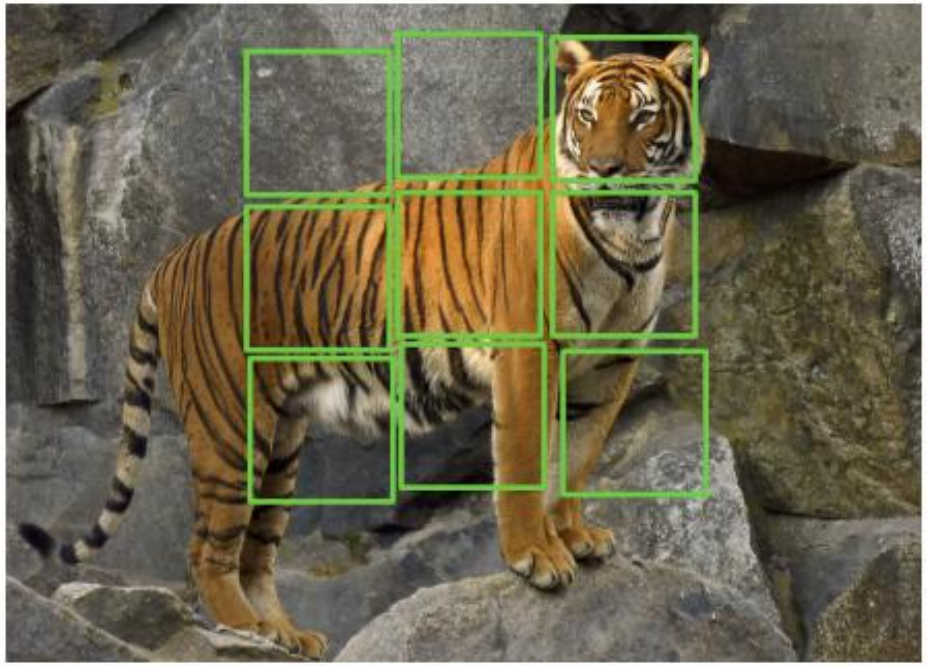
AlexNet-like architecture



C. Doersch, A. Gupta, A. Efros. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Jigsaw puzzle solving

Crop out tiles



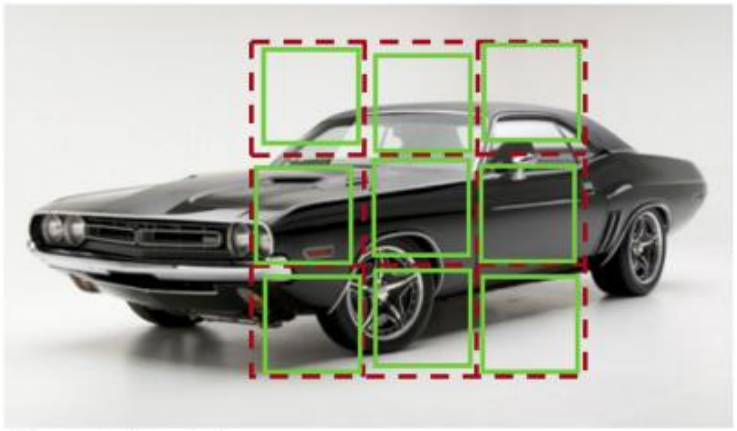
Shuffle

Pretext task: reassemble

Claim: jigsaw solving is easier than context prediction, trains faster, transfers better

M. Noroozi and P. Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV 2016

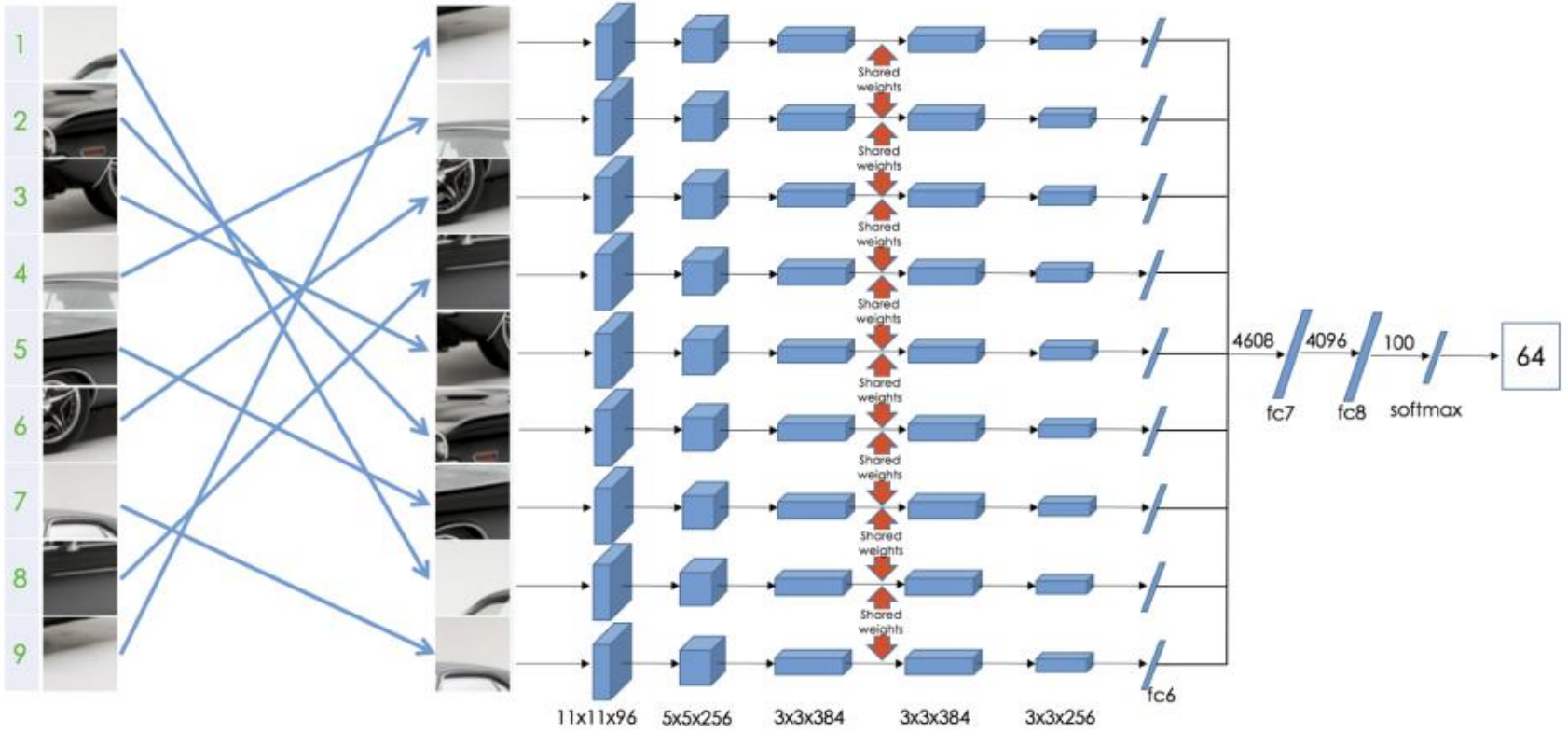
Jigsaw puzzle solving: Details



Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation



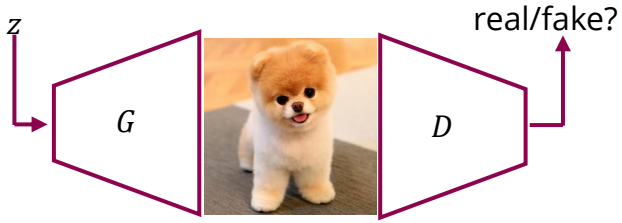
Predetermined set of
1000 permutations
(out of 362,880 possible)

M. Noroozi and P. Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV 2016

Learning Signal Toolbox

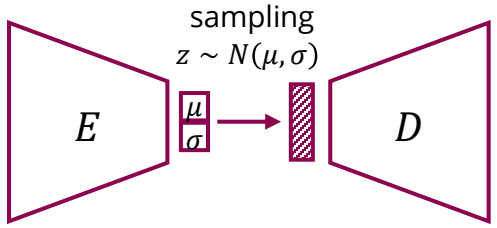
Generative

GANs



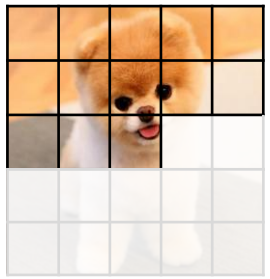
[Goodfellow et al.; NeurIPS ‘14]

Variational Autoencoder



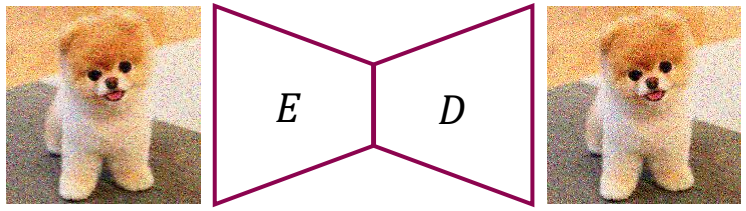
[Kingma et al.; ICLR ‘14]

Autoregressive



[van den Oord et al.;
ICML ’16, NeurIPS ‘16]

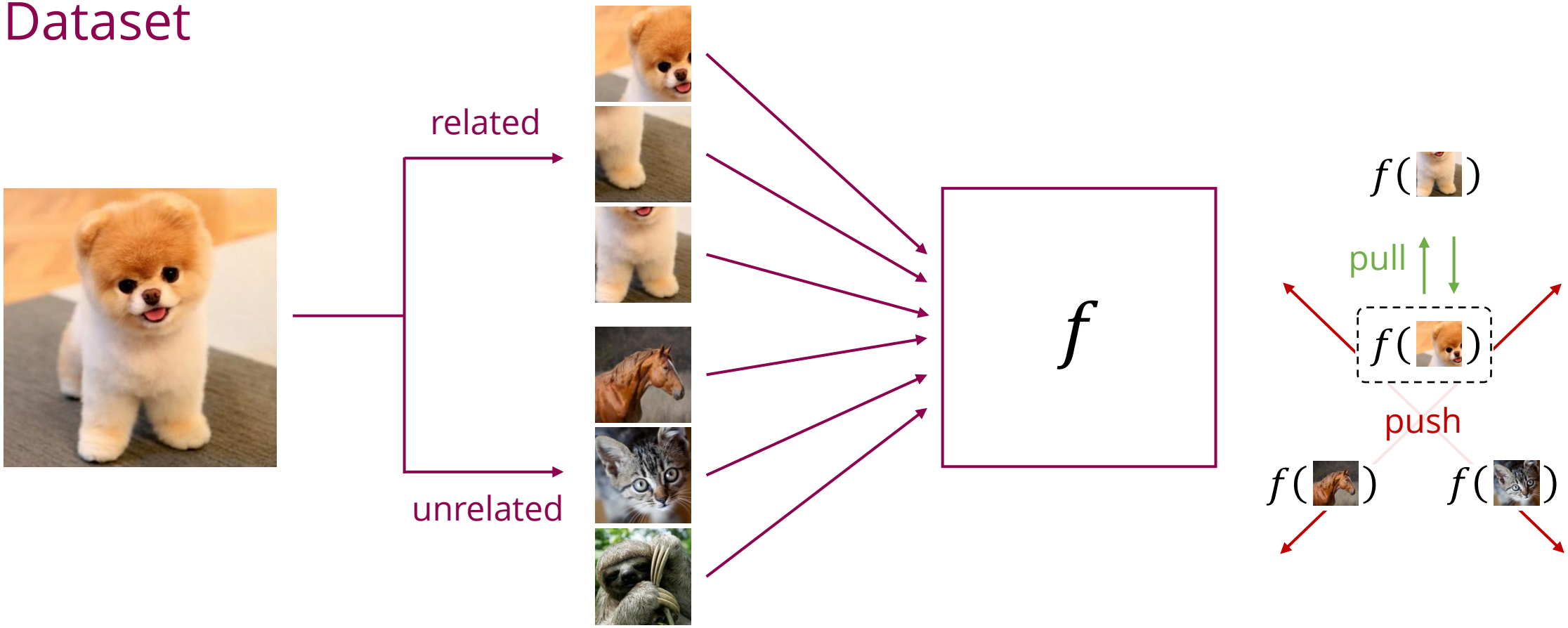
Diffusion Model



[Sohl-Dickstein et al.; ICML ‘15]

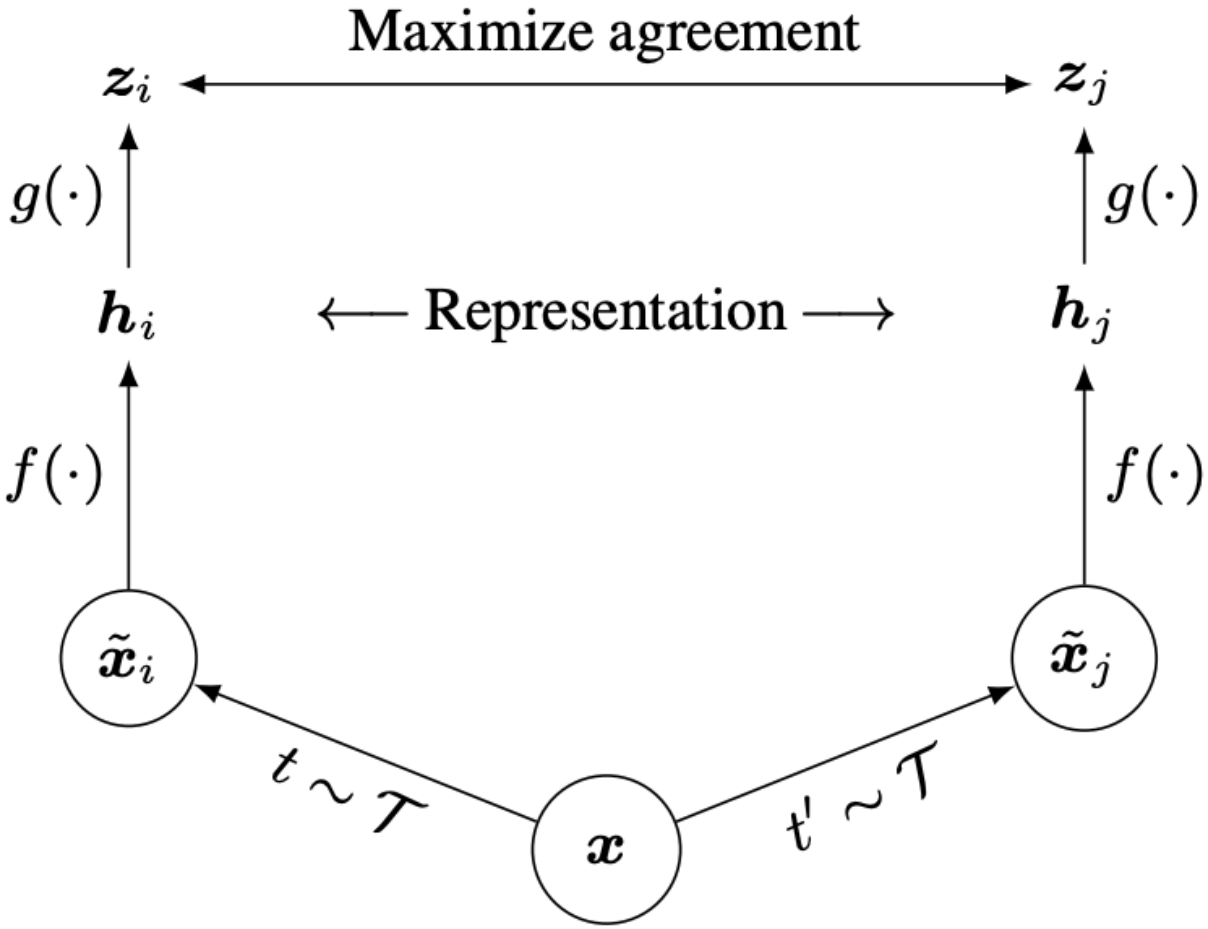
Learning Signal Toolbox

Dataset



* Different definitions of “related” and “unrelated” samples exist “Contrastive”

SimCLR

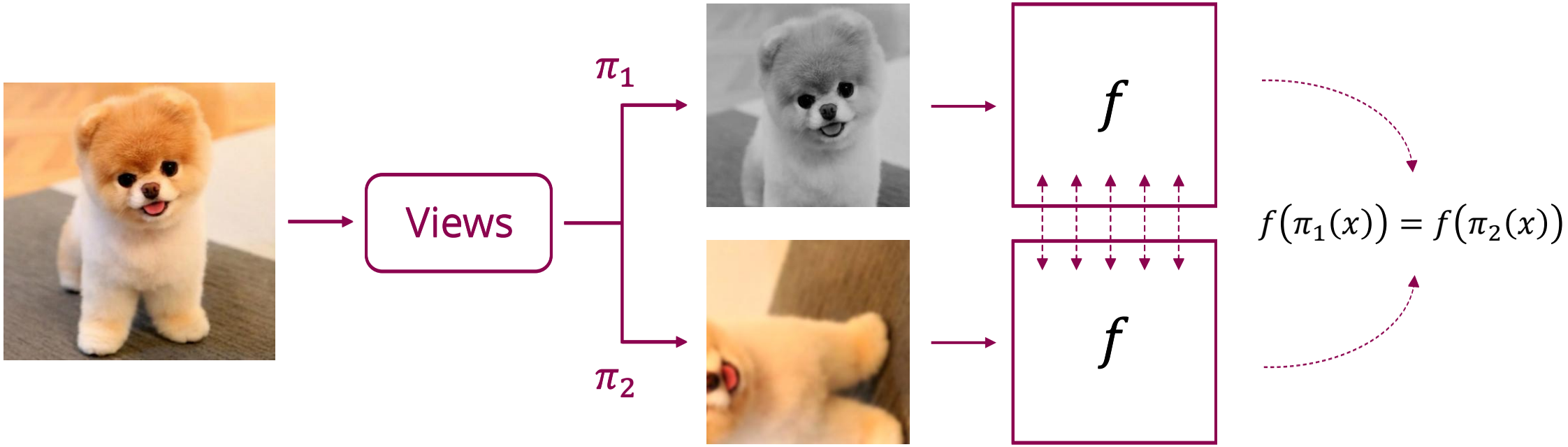


- Introduce nonlinear *projection* (g) between representation (h) and feature used for computing contrastive loss (z).
- Use large mini-batch size.

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

Learning Signal Toolbox

Invariance



* Often with strong augmentations, but without changing the identity of the image

SimCLR: Augmentations

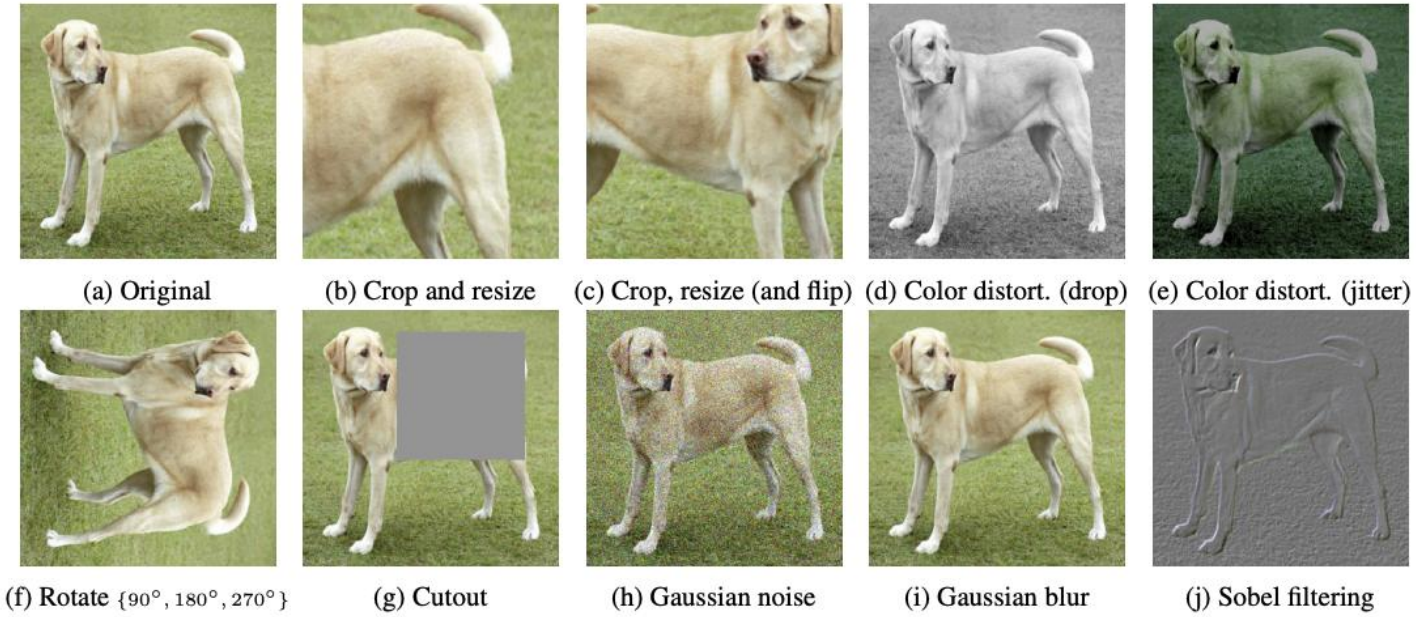


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

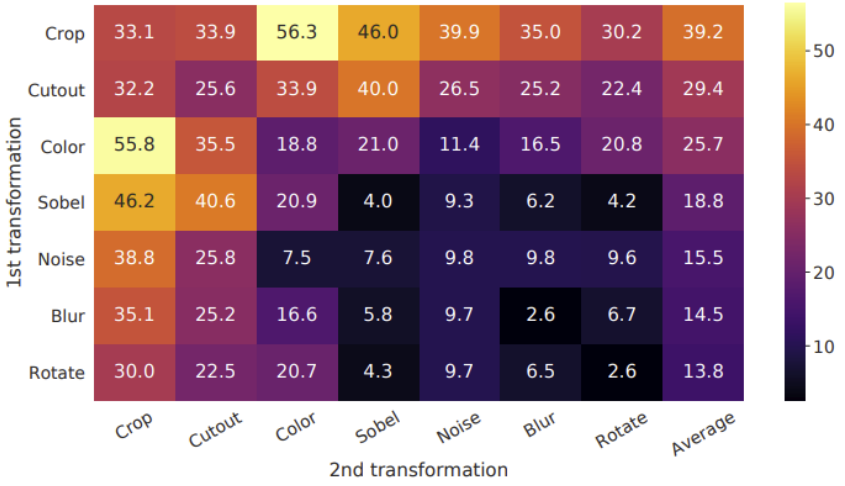


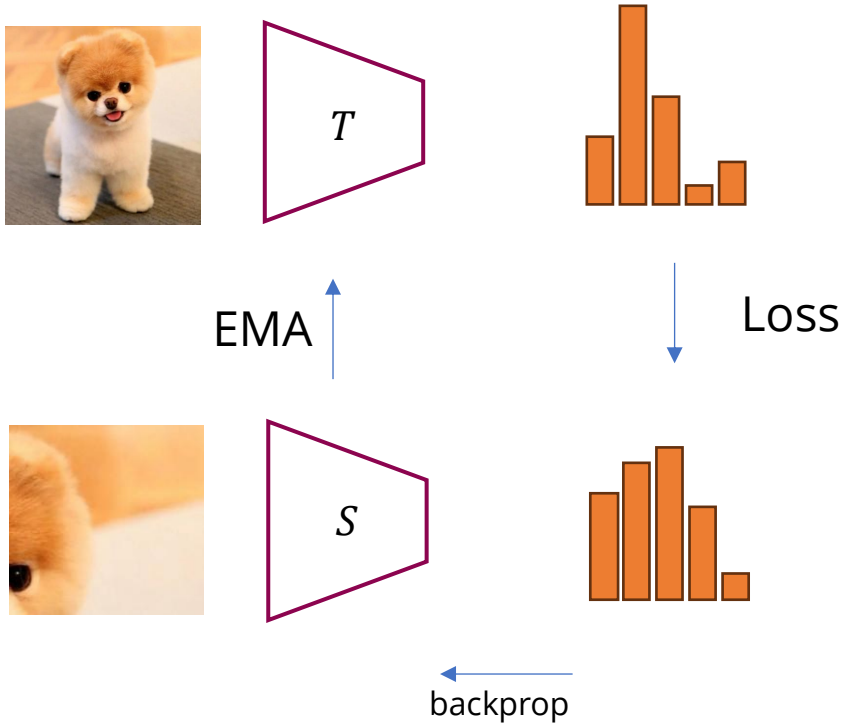
Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

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

DINO: Self-Distillation with No Labels

- Student-Teacher training.
- Teacher's weights are exponential moving average (EMA) of the student.
- Teacher sees global view.
- Student sees local view.

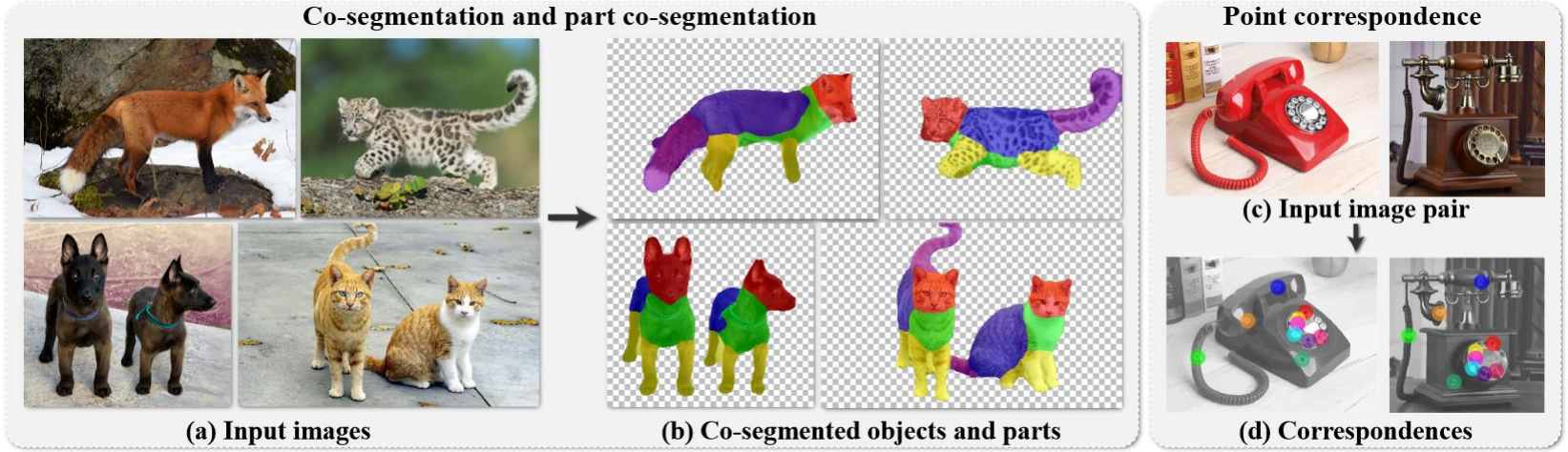
- Student tries to predict teacher's distribution.



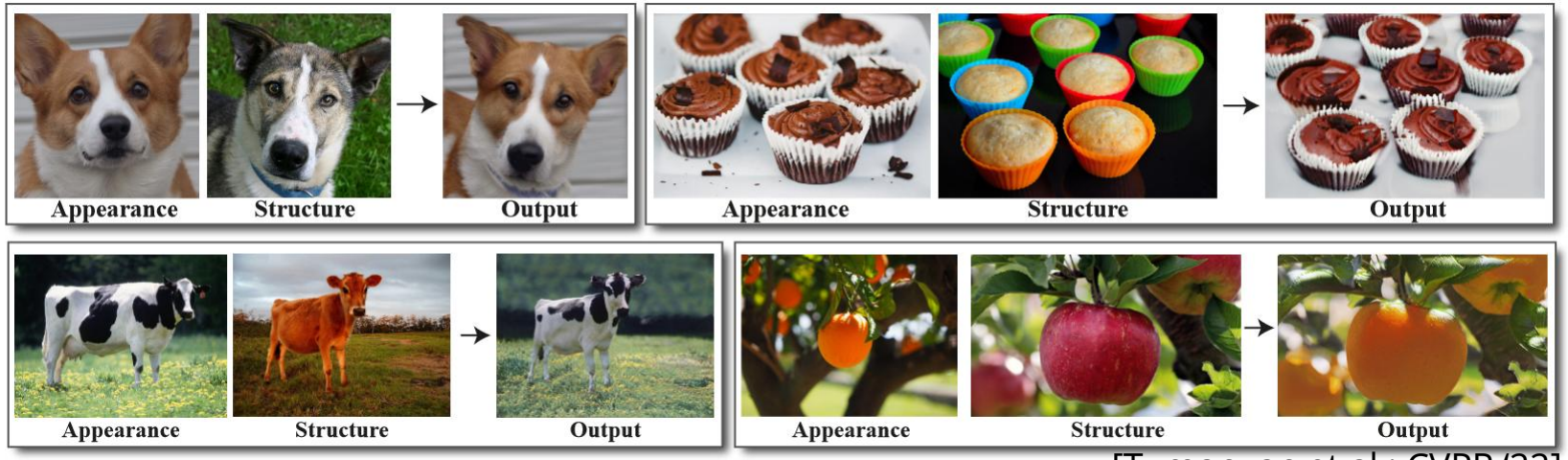
DINO – sharpening and centring

- Collapse: same prediction for all samples.
- Centring: $f_T(x_i) - \frac{1}{N}\sum_{j=1}^{N} f_T(x_j)$
- Sharpening: low temperature for teacher softmax.
- Loss: Entropy between student and teacher distributions.

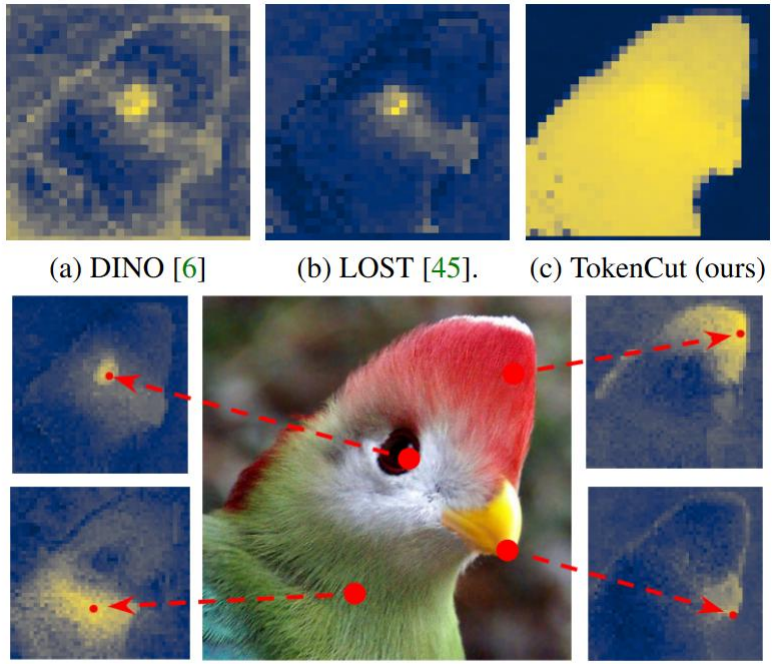
DINO features are popular



[Amir et al.; CVPR '22]



[Tumanyan et al.; CVPR '22]



[Wang et al.; CVPR '22]

Evaluation

- Occasionally simple: when training aligns fully with the task
- Often: some processing is needed
- Bridging the final gap between model and task

- In a practical setting: unsupervised learning is just the beginning

- In research: how far do we get with as little supervision as possible?

Unsupervised vs. Self-Supervised?

Andrew Davison
@AjdDavison

Can anyone explain if there is a difference between unsupervised and self-supervised learning? To me they seem the same and I find myself using both terms interchangeably (I prefer unsupervised), but I feel like I'm confusing people who understand them to mean different things.

7:16 PM · Jun 14, 2022 · Twitter for Android

They are the same,

Self supervised uses a different loss function

To be a bit cute, unsupervised is the term for older techniques that don't work. Self supervised is the term for newer methods that do work.

Unlike SSL, it's difficult to abbreviate unsupervised learning

labels are derived (perhaps implicitly) using problem-specific principles

"self-supervised" when the code looks exactly like supervised

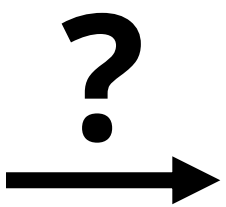
There is no common definition (and someone will always complain)!

Unsupervised vs. Weakly Supervised

Weak supervision means supervision but for a different task

- Segmentation from bounding boxes/captions/classes
- (Dense) depth from stereo
- Captioning from object labels
- Human in the loop annotations
- Objects from sound
- …

Unsupervised Image Classification

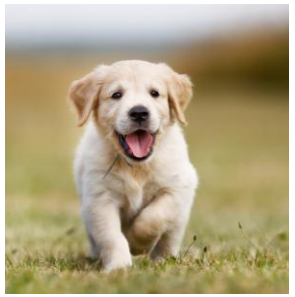


f

Unsupervised Image Classification

f

f



f

Grouping

Class 42

Class 27

Class 42

Assignments found with supervision

Class 1 = Cat

...

Class 27 = Butterfly

...

Class 42 = Dog

...

Semantics

Horse

Cat

Butterfly

Dog

Cow

Hungarian Matching

- Find the lowest cost 1-to-1 assignments between N clusters and N labels.
- Cost: a $N \times N$ matrix C that contains the errors we induce when we match cluster i to label j.

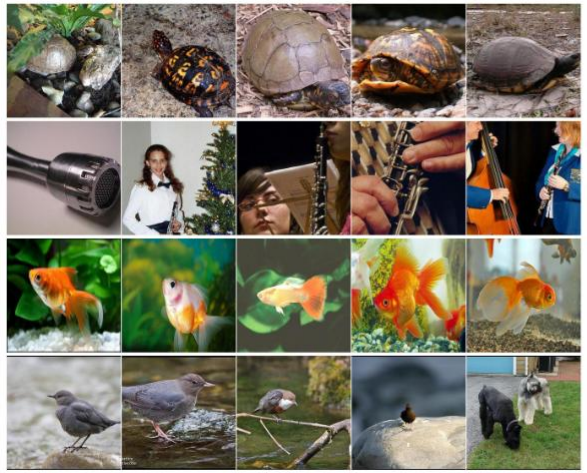
- Find a row permutation matrix P that minimizes the diagonal.

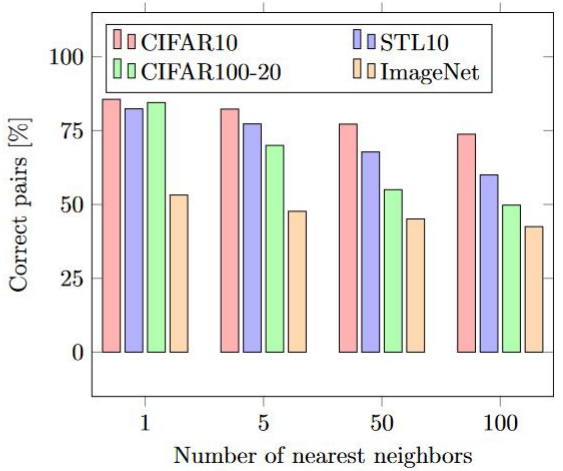
$$\min_P \mathrm{Tr}(PC)$$

- Complexity: $\mathcal{O}(N^3)$

Classifying Images without Labels

- Learn a self-supervised representation
- Loss: Neighbouring images same class + all classes have equal size
- Even better with self-training





Method	Backbone	Labels	Top-1	Top-5
Supervised Baseline	ResNet-50	✓	25.4	48.4
Pseudo-Label	ResNet-50	✓	-	51.6
VAT + Entropy Min. [56]	ResNet-50	✓	-	47.0
InstDisc [51]	ResNet-50	✓	-	39.2
BigBiGAN [15]	ResNet-50(4x)	✓	-	55.2
PIRL [32]	ResNet-50	✓	-	57.2
CPC v2 [20]	ResNet-161	✓	52.7	77.9
SimCLR [7]	ResNet-50	✓	48.3	75.5
SCAN (Ours)	ResNet-50	✗	39.9	60.0

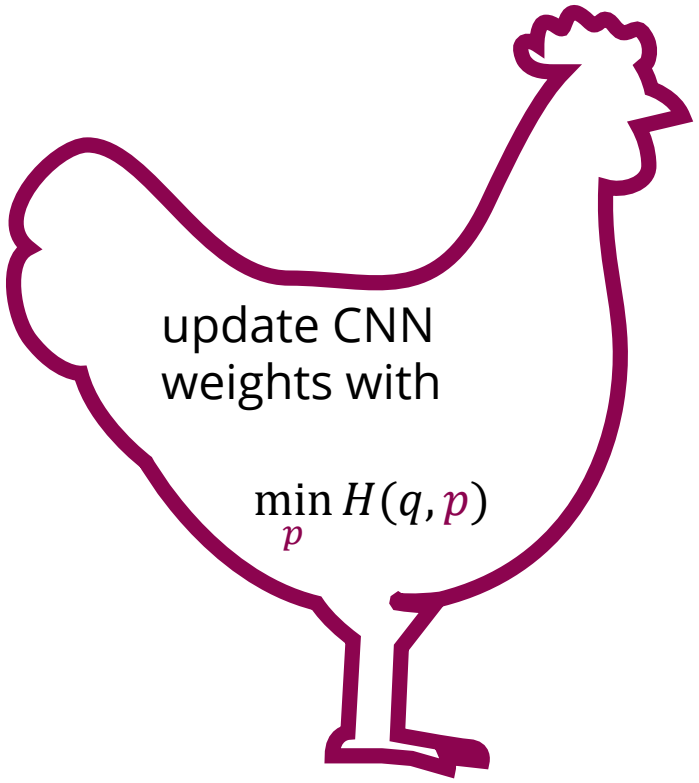
[Gansbeke et al.; ECCV '20]

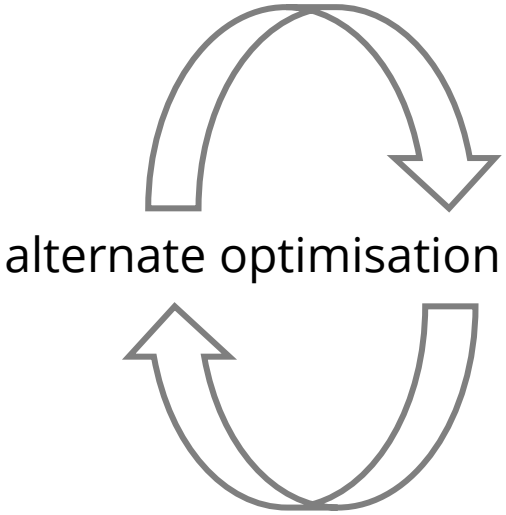
Self-labelling by Clustering

Learning

CNN

"class 37"



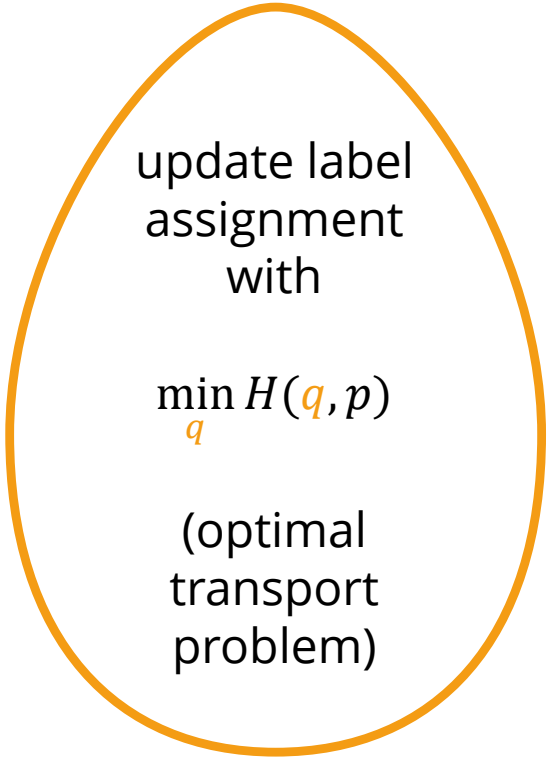


Cross-Entropy:

$$H(q,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{y} \mathrm{q(y|x_i)} \log p(y|x_i, \Theta)$$

Labelling

"class 37"



[Asano et al.; ICLR '20]

Clusters are interpretable

cluster 1908, purity: 0.952

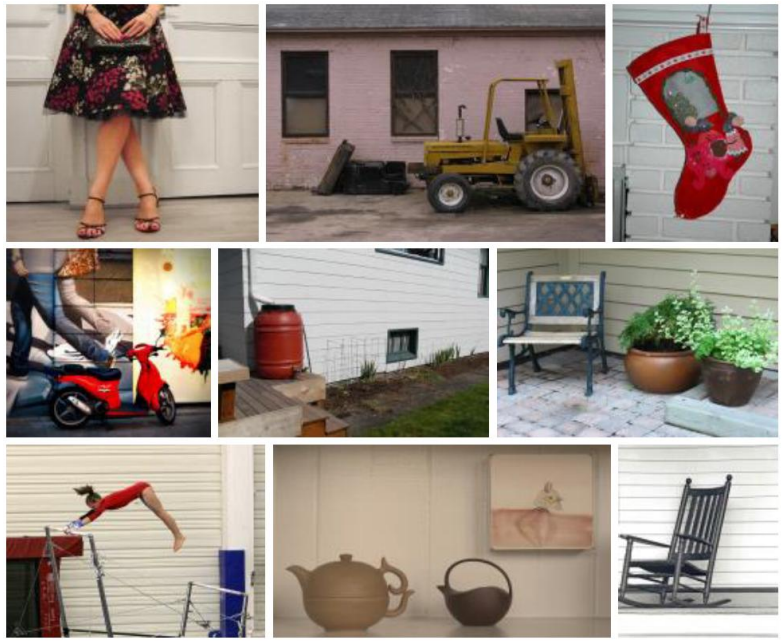
cluster 503, purity: 0.930

cluster 393, purity: 0.668

Clusters are interpretable

cluster 406, purity: 0.455

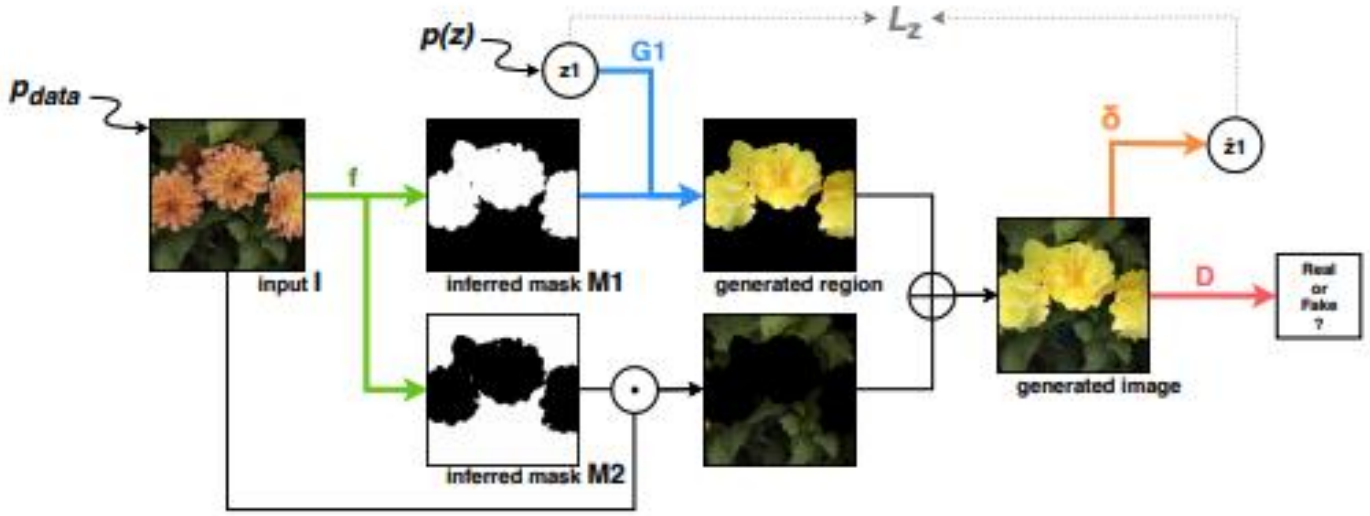
cluster 0, purity: 0.558



cluster 2568, purity: 0.377

GAN-based Segmentation

- ReDo: Layer-wise generative models for unsupervised object discovery

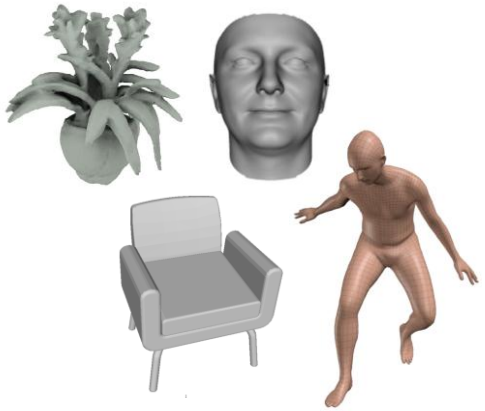


- Advantage: You get segmentation for free!
- Drawbacks: fragile training, difficulty scaling due to custom architecture

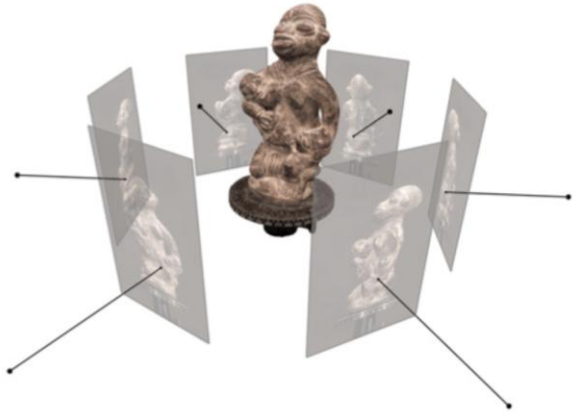
Unsupervised object segmentation by redrawing
M Chen, T Artières, L Denoyer, NeurIPS 2019

Chen et al. 2019

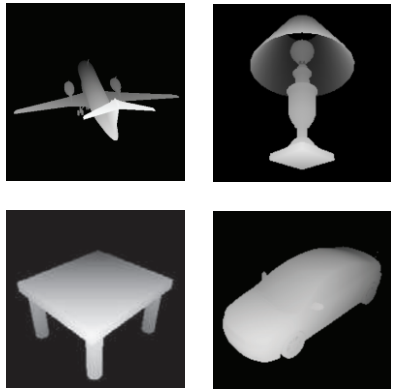
3D from a single image



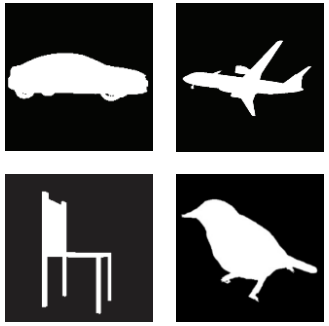
3D ground truth or shape models



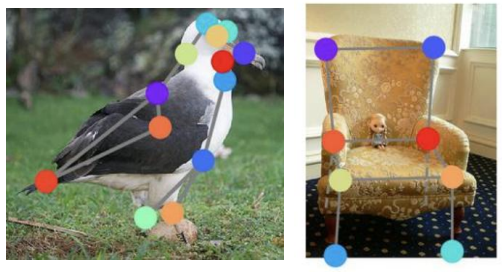
multi-view



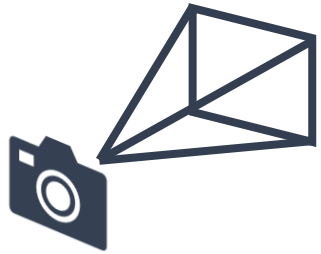
depth maps



silhouettes

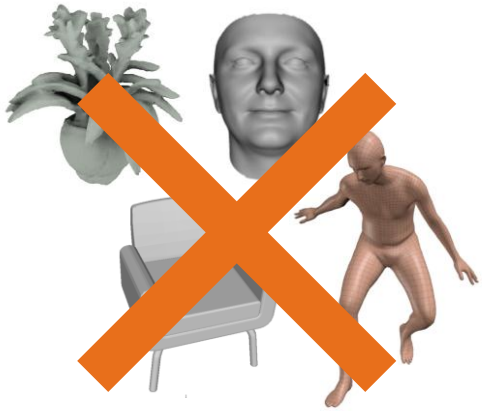


keypoints



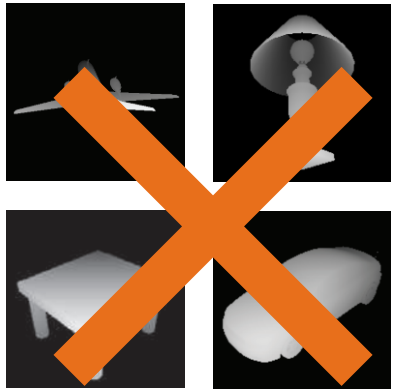
camera viewpoint

3D from a single image

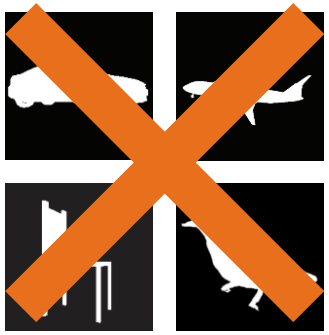


3D ground truth or shape models

multi-view



depth maps



silhouettes

keypoints

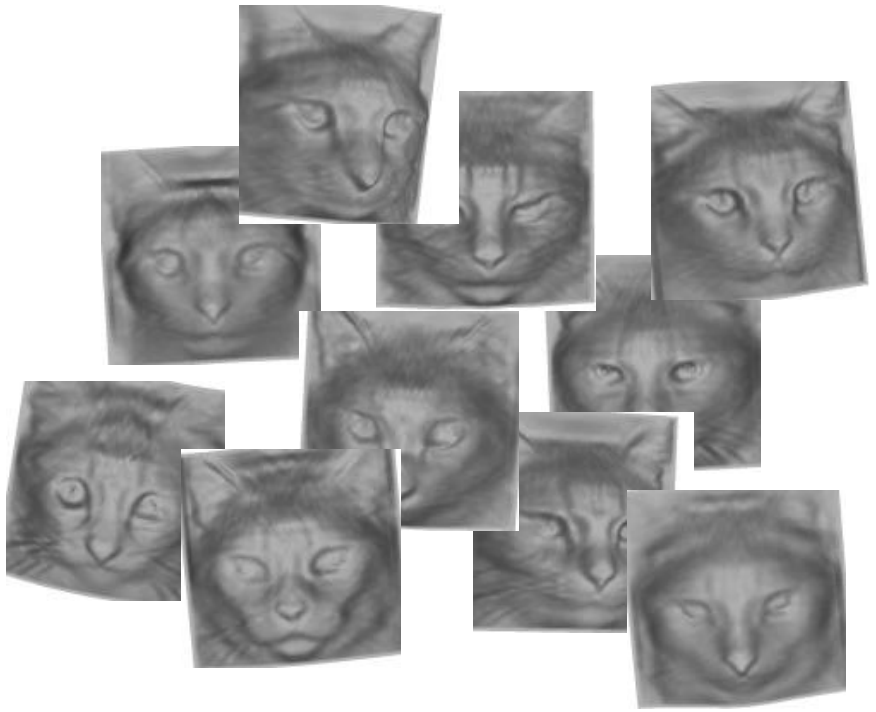
camera viewpoint

Unsupervised Learning of 3D Objects

Training Data

Unsup3D

Output



single-view images of a category

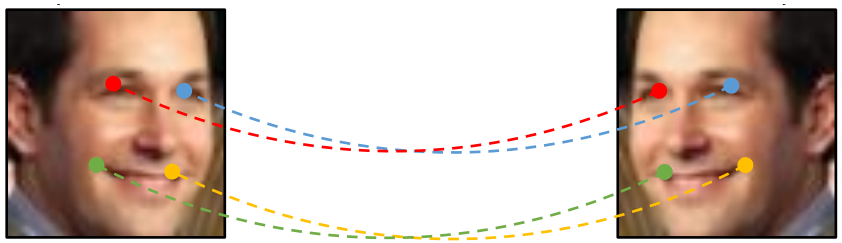
NO other supervision!

instance-specific 3D shapes

Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild, S Wu et al., CVPR 2020

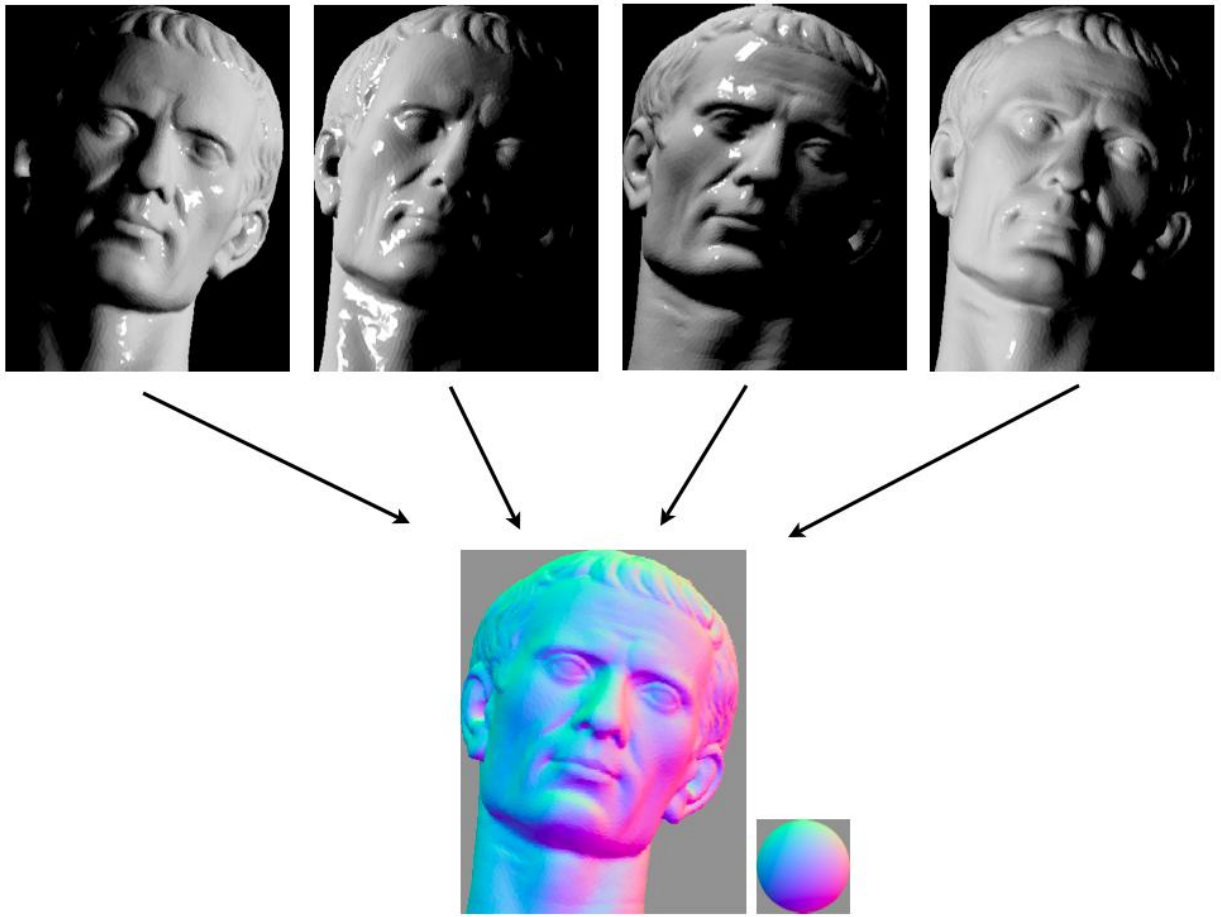
Observation I

- Symmetry is a strong constraint!



Observation II

- Shading is a strong constraint! (Shape from shading)



Photometric method for determining surface orientation from multiple images
Robert J Woodham, Optical engineering, 1980

Photo-Geometric Autoencoding

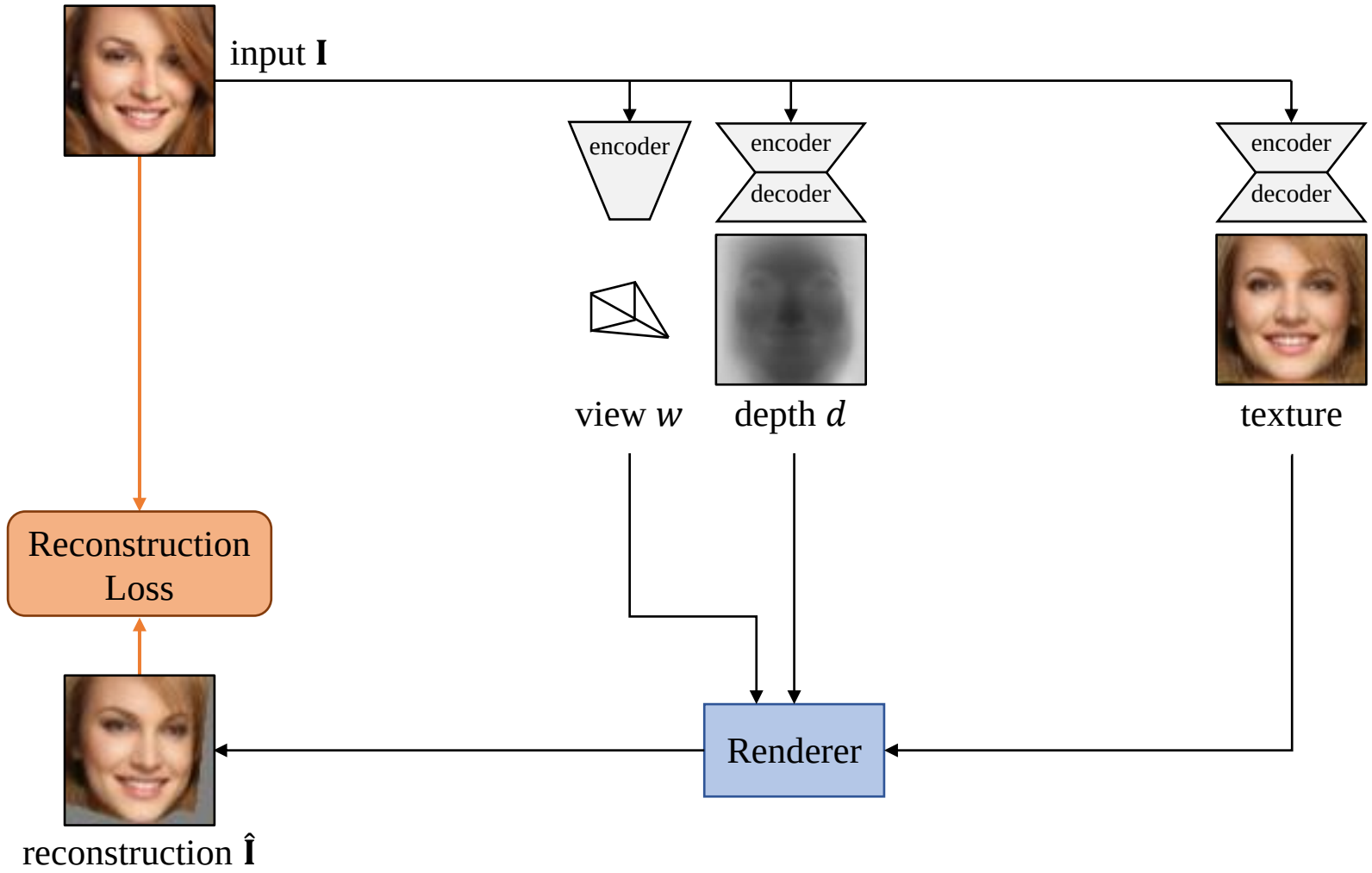


Photo-Geometric Autoencoding

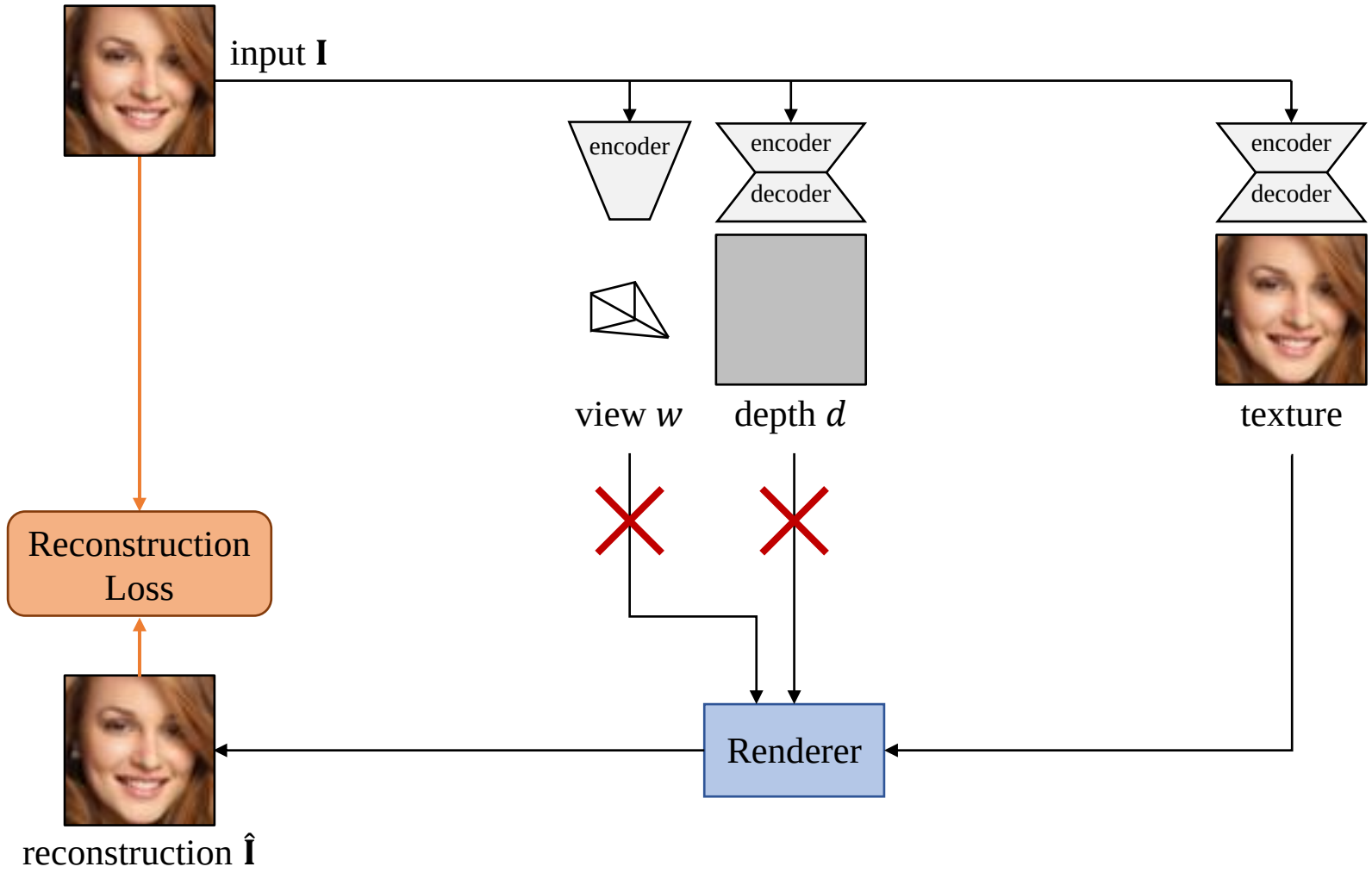


Photo-Geometric Autoencoding

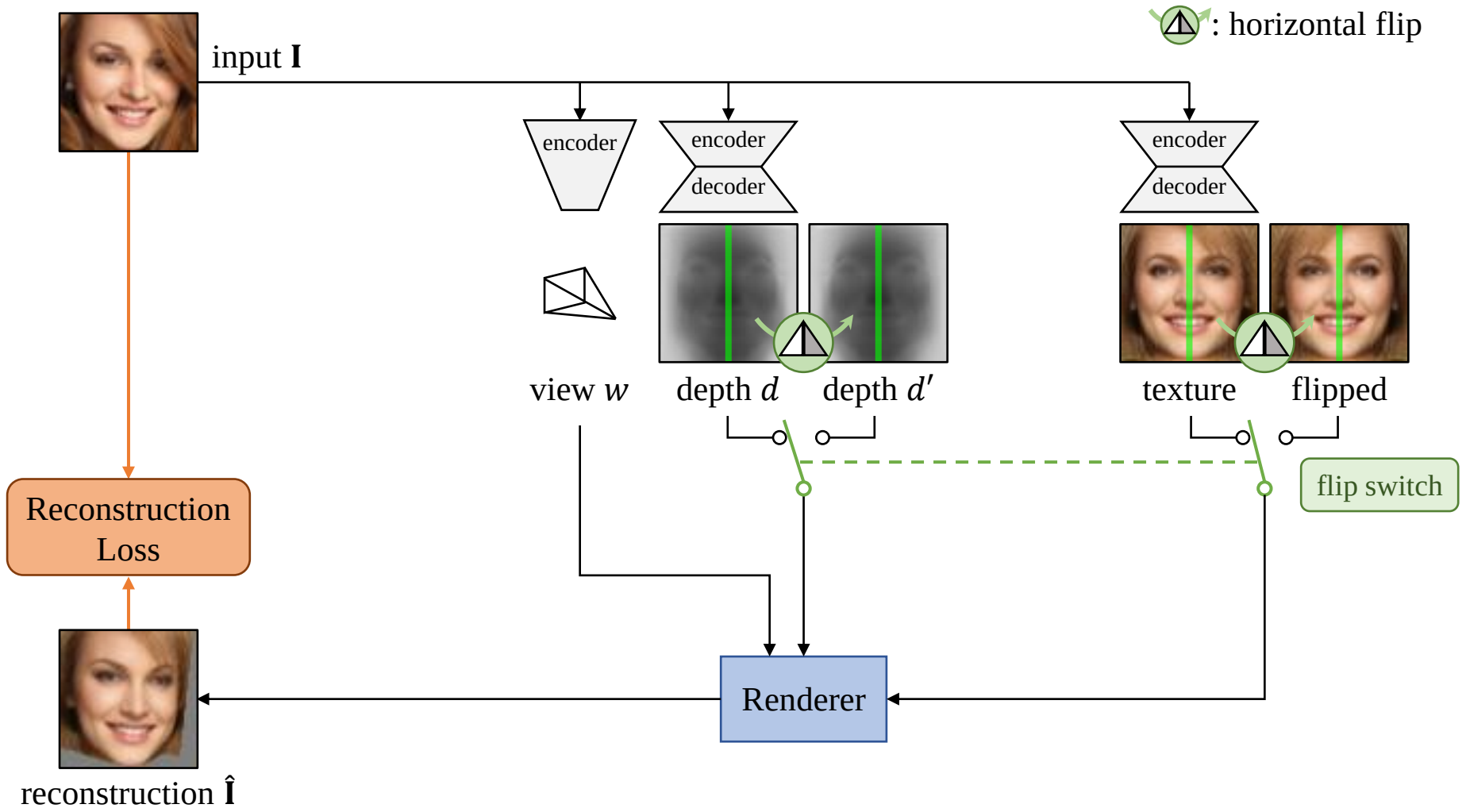


Photo-Geometric Autoencoding

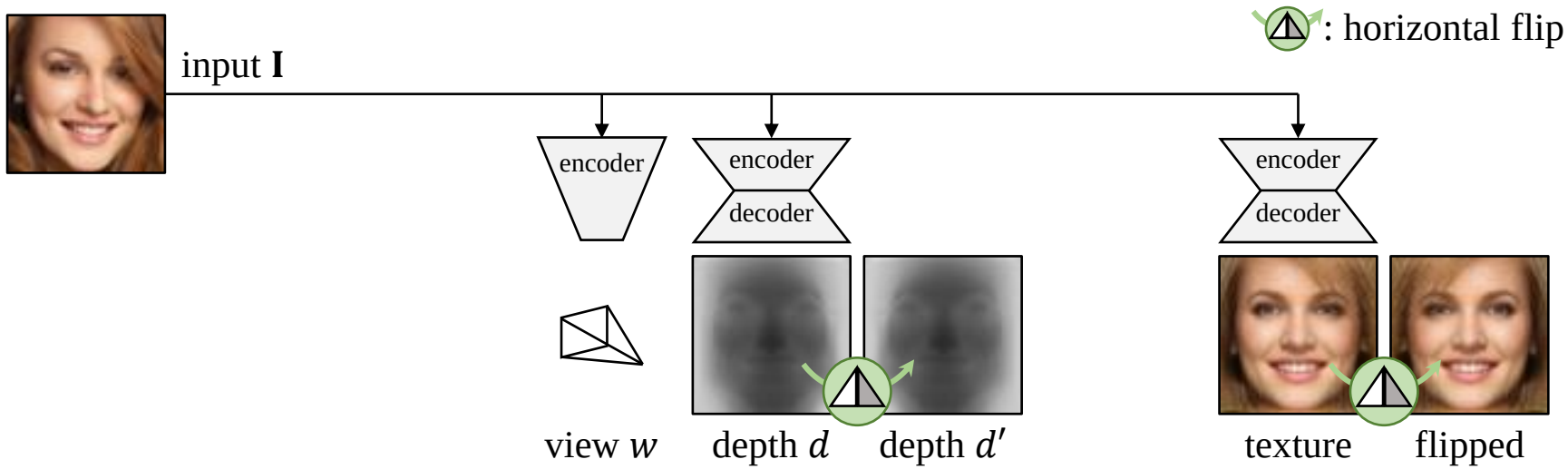


Photo-Geometric Autoencoding

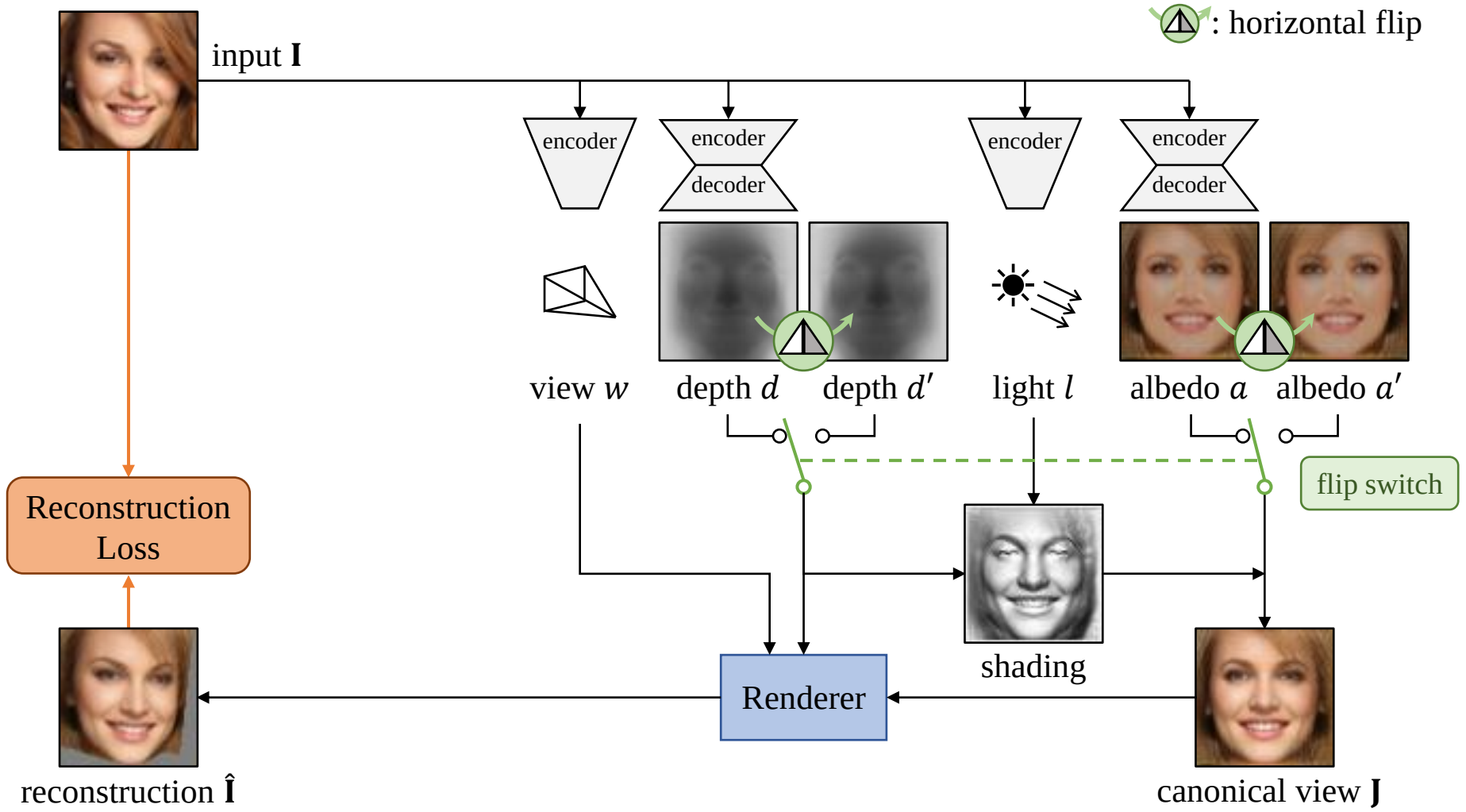


Photo-Geometric Autoencoding

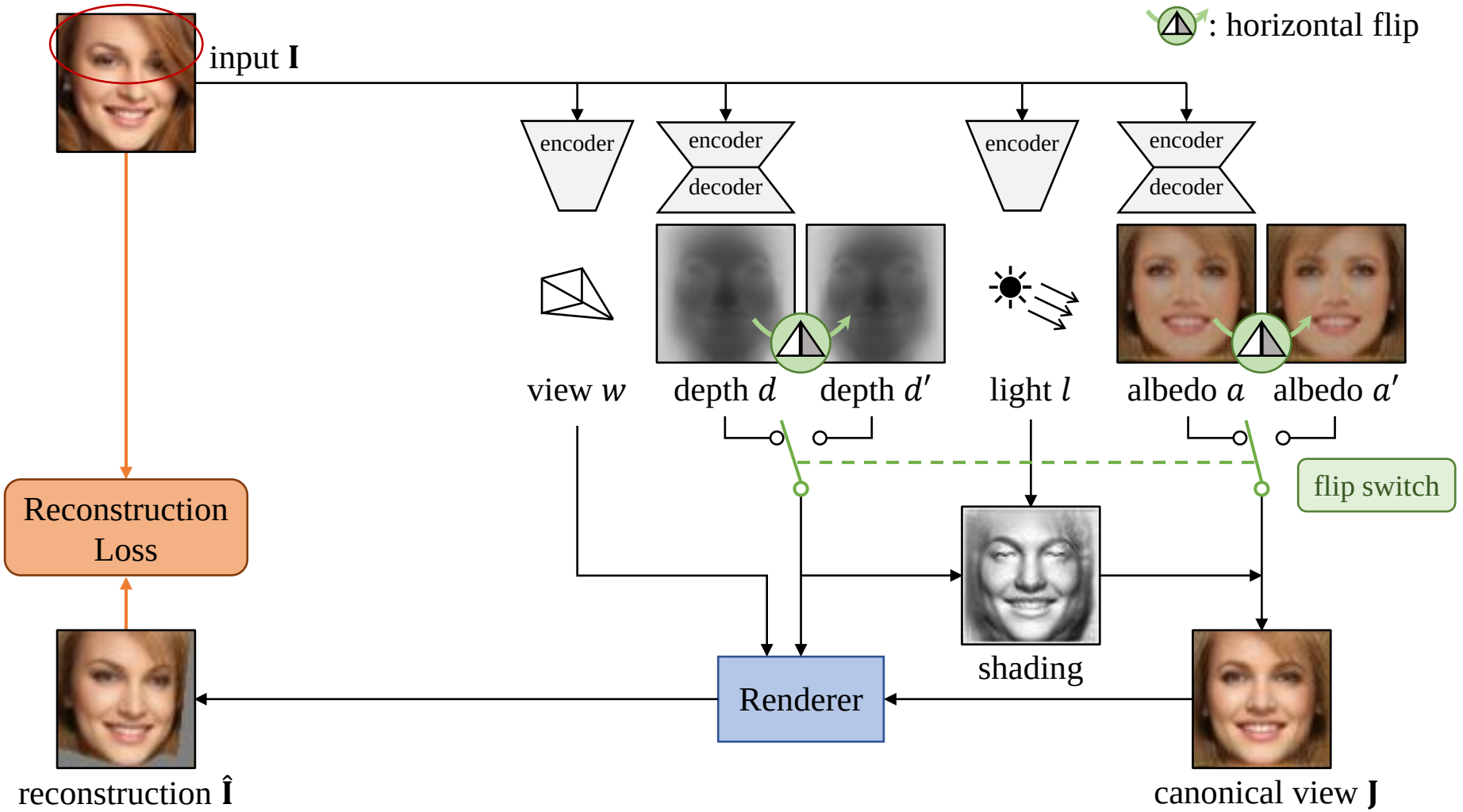


Photo-Geometric Autoencoding

Q3: Non-symmetric albedo, deformation, etc?

A3: Predict uncertainty

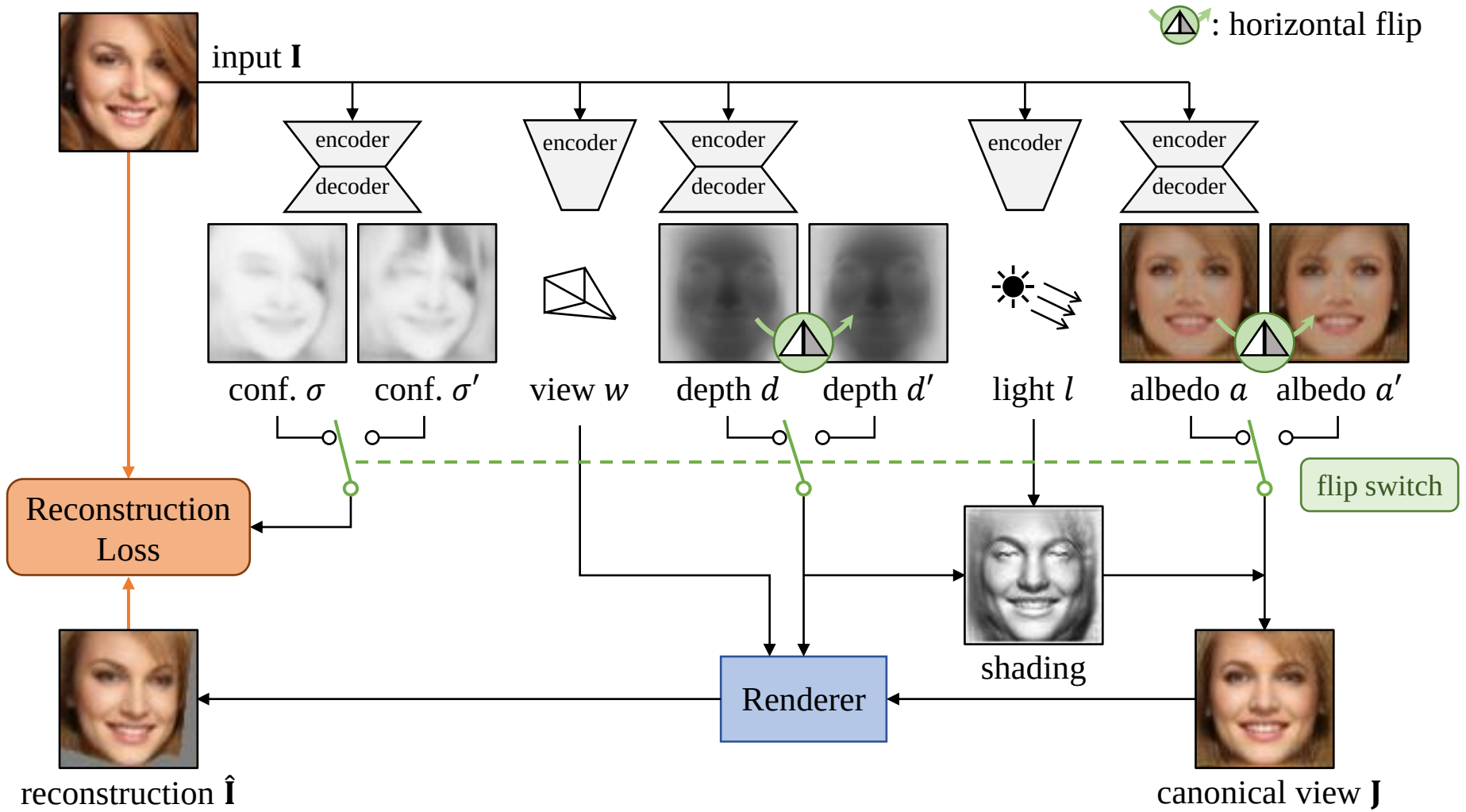
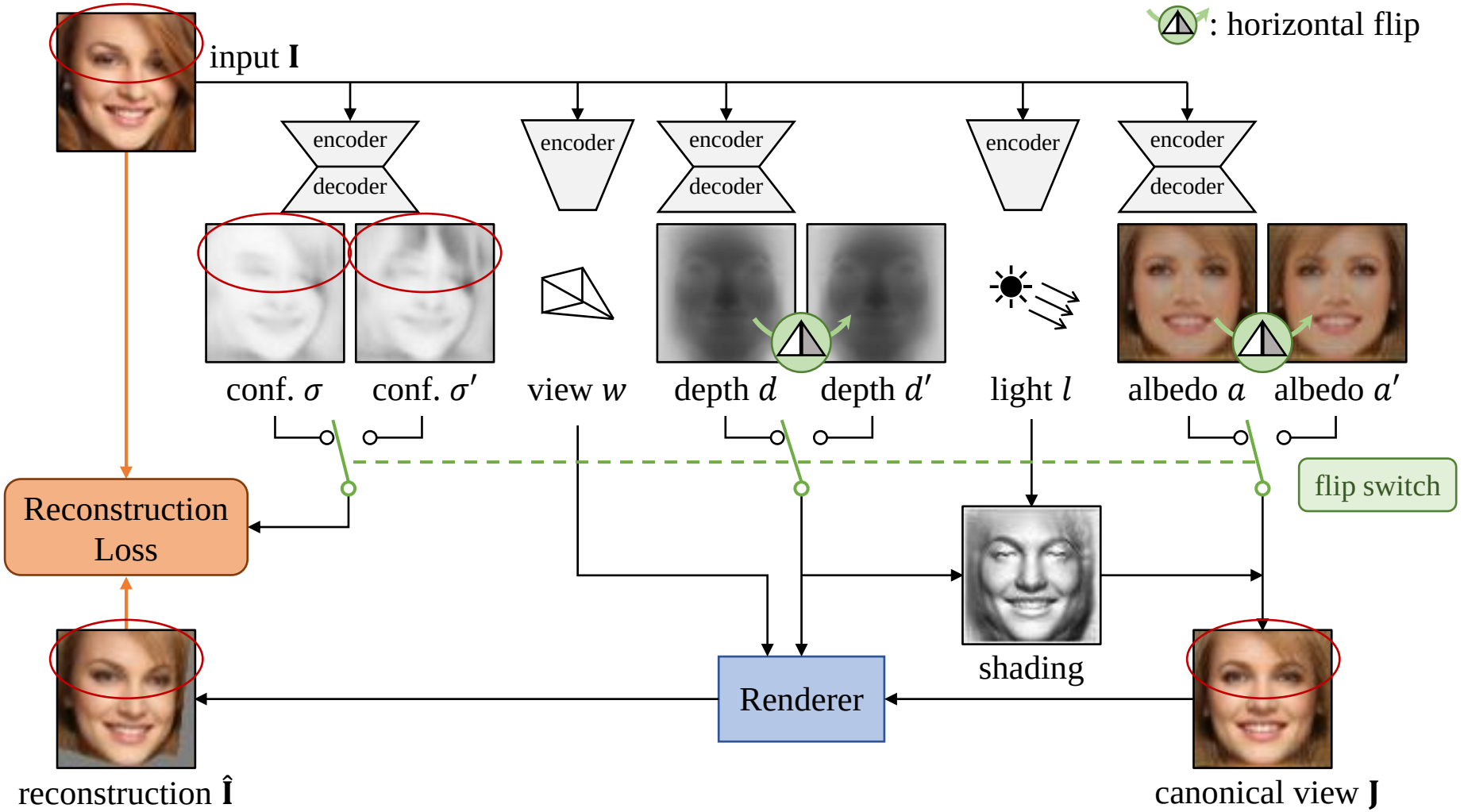
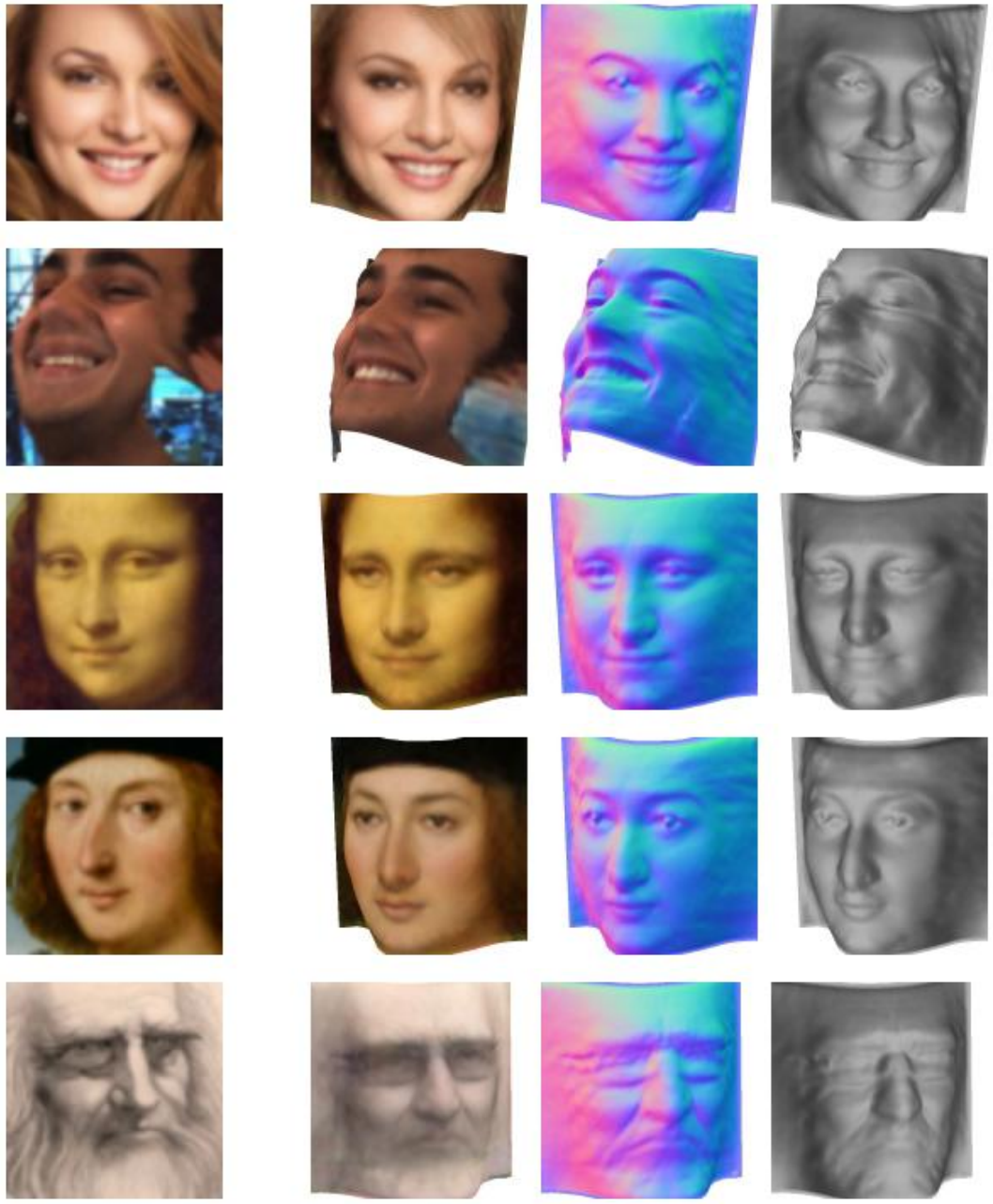
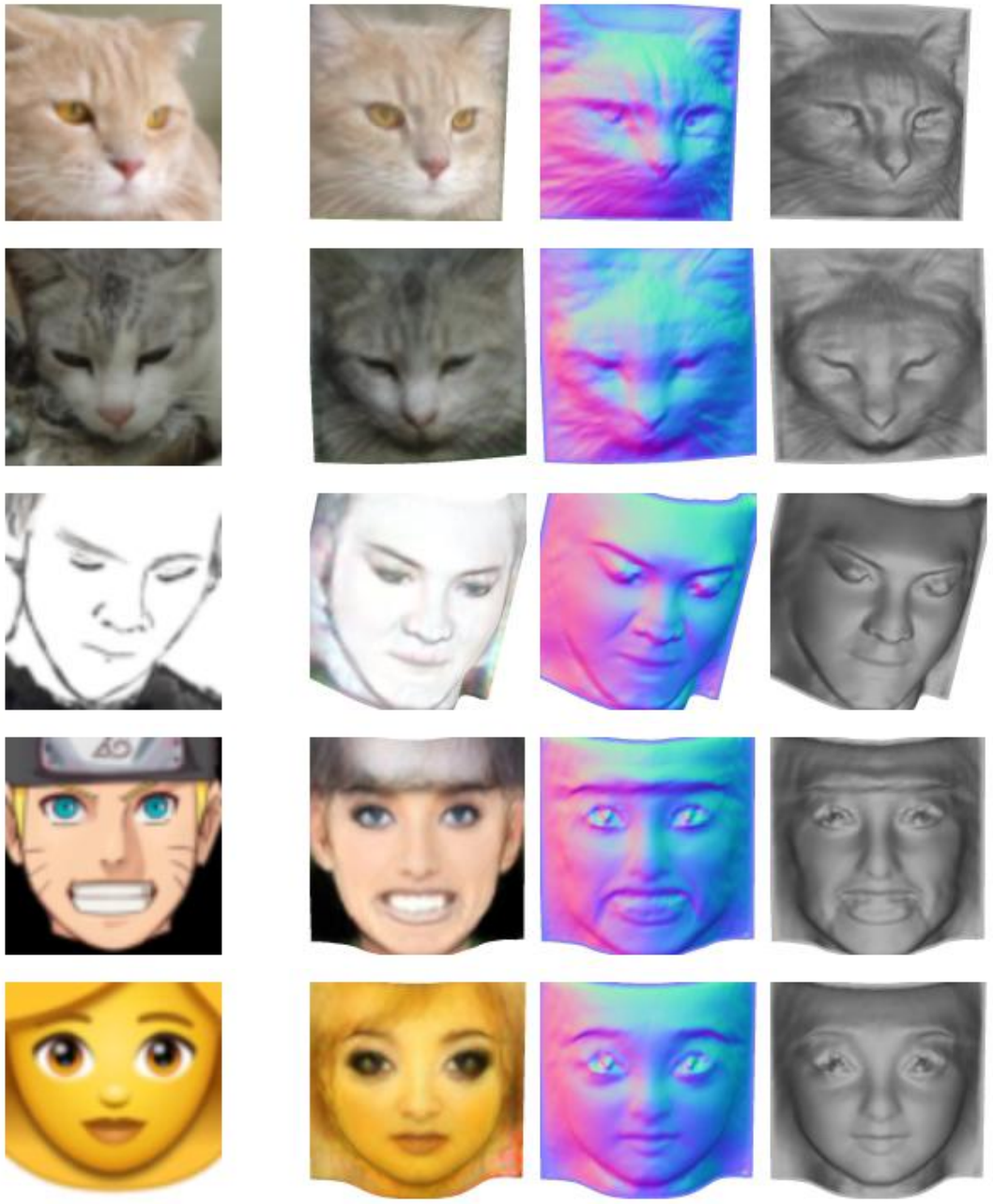


Photo-Geometric Autoencoding





input reconstruction



input reconstruction