ROBOT INTELLIGENCE LAB

Self-Supervised Learning

Good-old-fashioned SSL

Sungjoon Choi, Korea University

kakaobrain

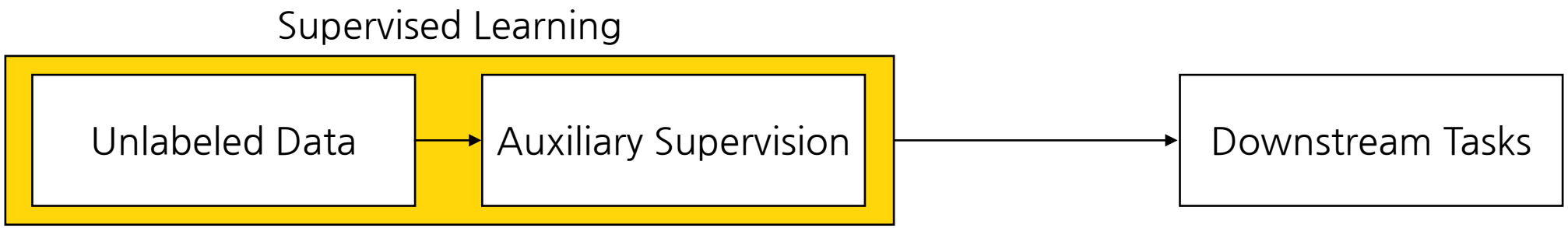
Introduction

ROBOT INTELLIGENCE LAB

What is self-supervised learning?

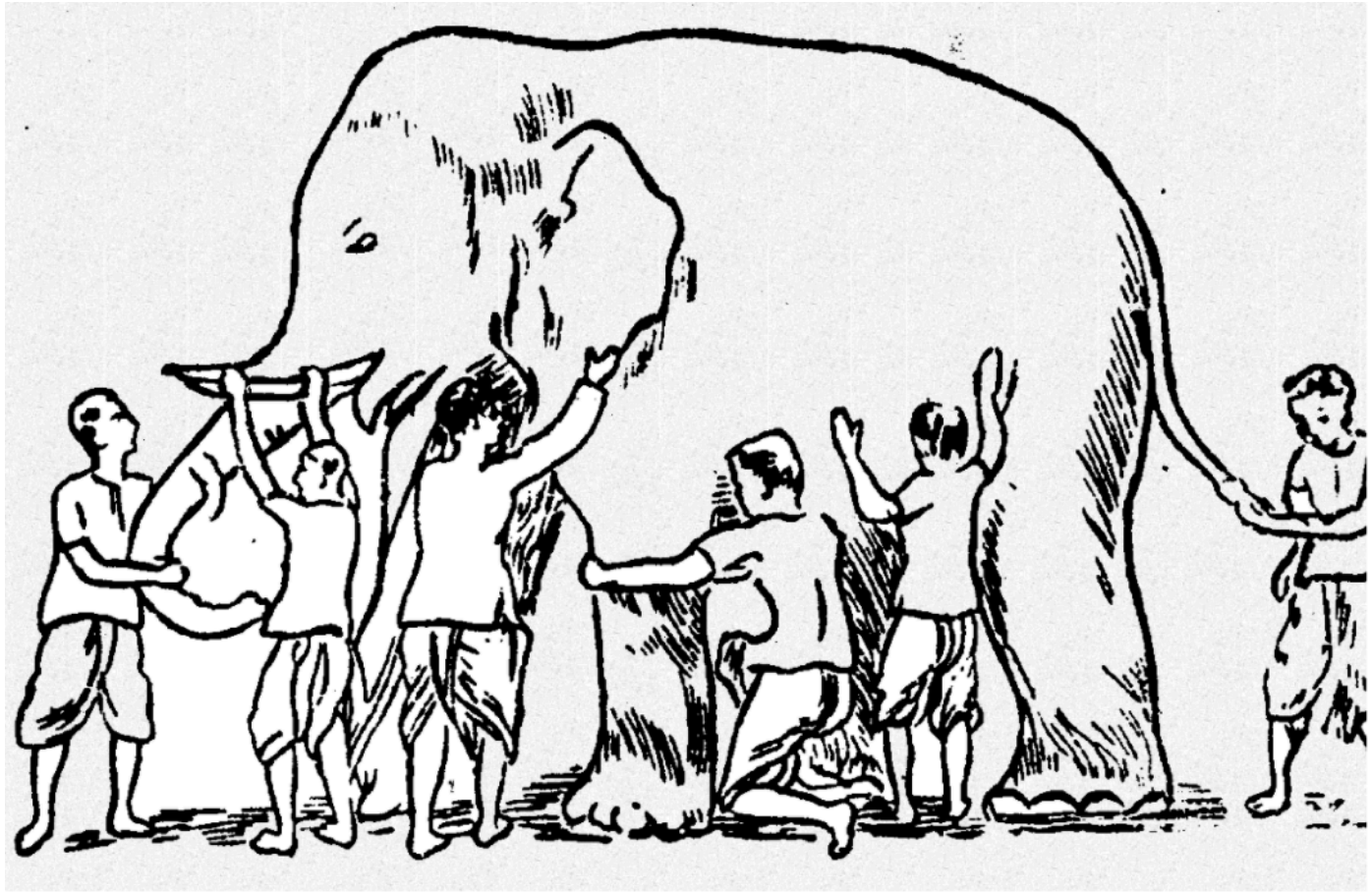
Self-Supervised Learning

ROBOT INTELLIGENCE LAB



Caution

ROBOT INTELLIGENCE LAB



Contents

ROBOT INTELLIGENCE LAB

ROBOT INTELLIGENCE LAB

Jigsaw

"Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles," 2017

Jigsaw Puzzles

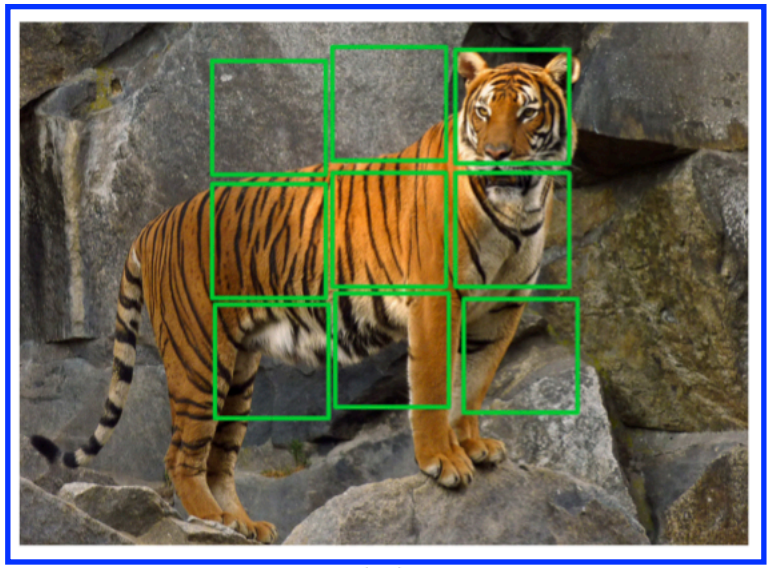
ROBOT INTELLIGENCE LAB

https://www.t3.com/news/where-to-buy-jigsaw-puzzles

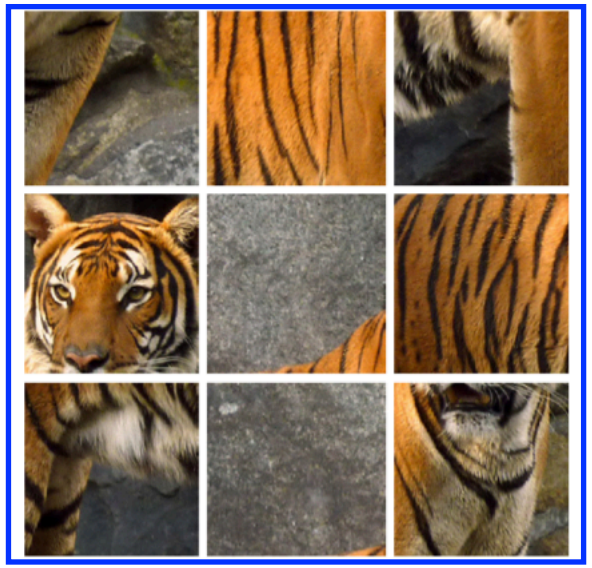
Jigsaw Puzzle Reassembly

ROBOT INTELLIGENCE LAB

Select **patches** from an image



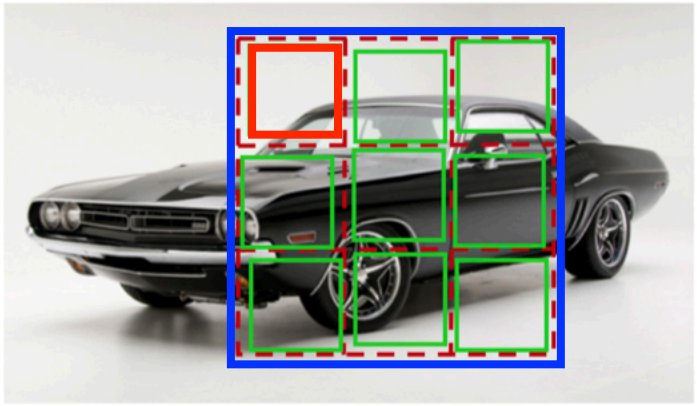
Shuffle **patches**



Reassemble **patches**

Context-Free Architecture

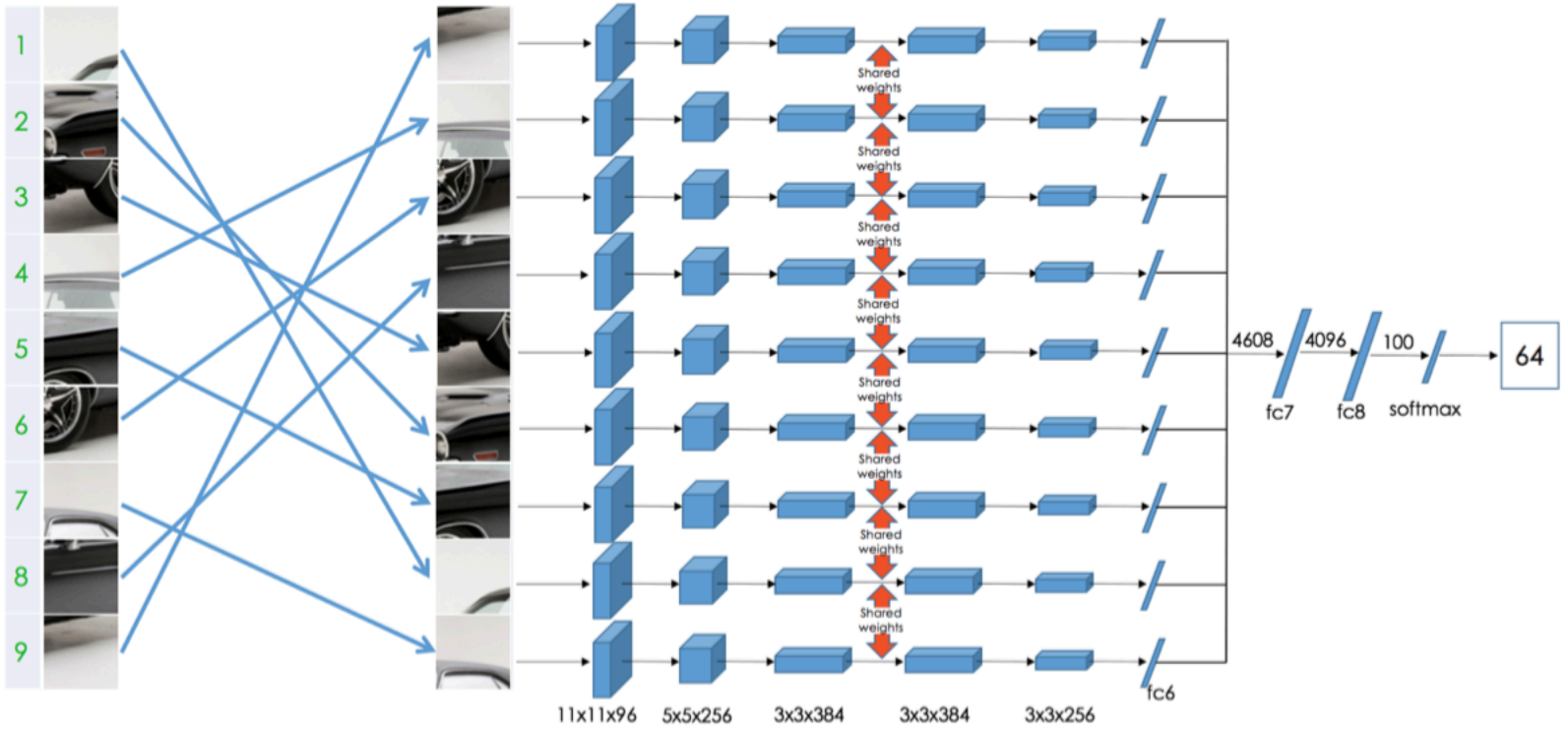
ROBOT INTELLIGENCE LAB



Permutation Set

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

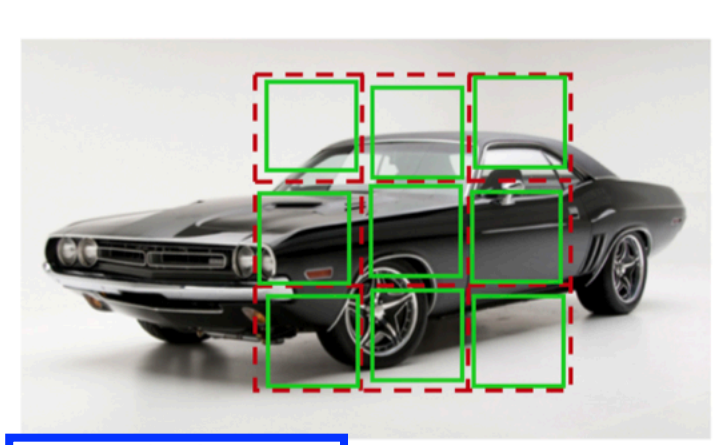
Reorder patches according to the selected permutation



Given an image, context-free architecture randomly crop 225×225 **pixel window**, and randomly pick a 64×64 **pixel tiles** from each 75×75 pixel cell.

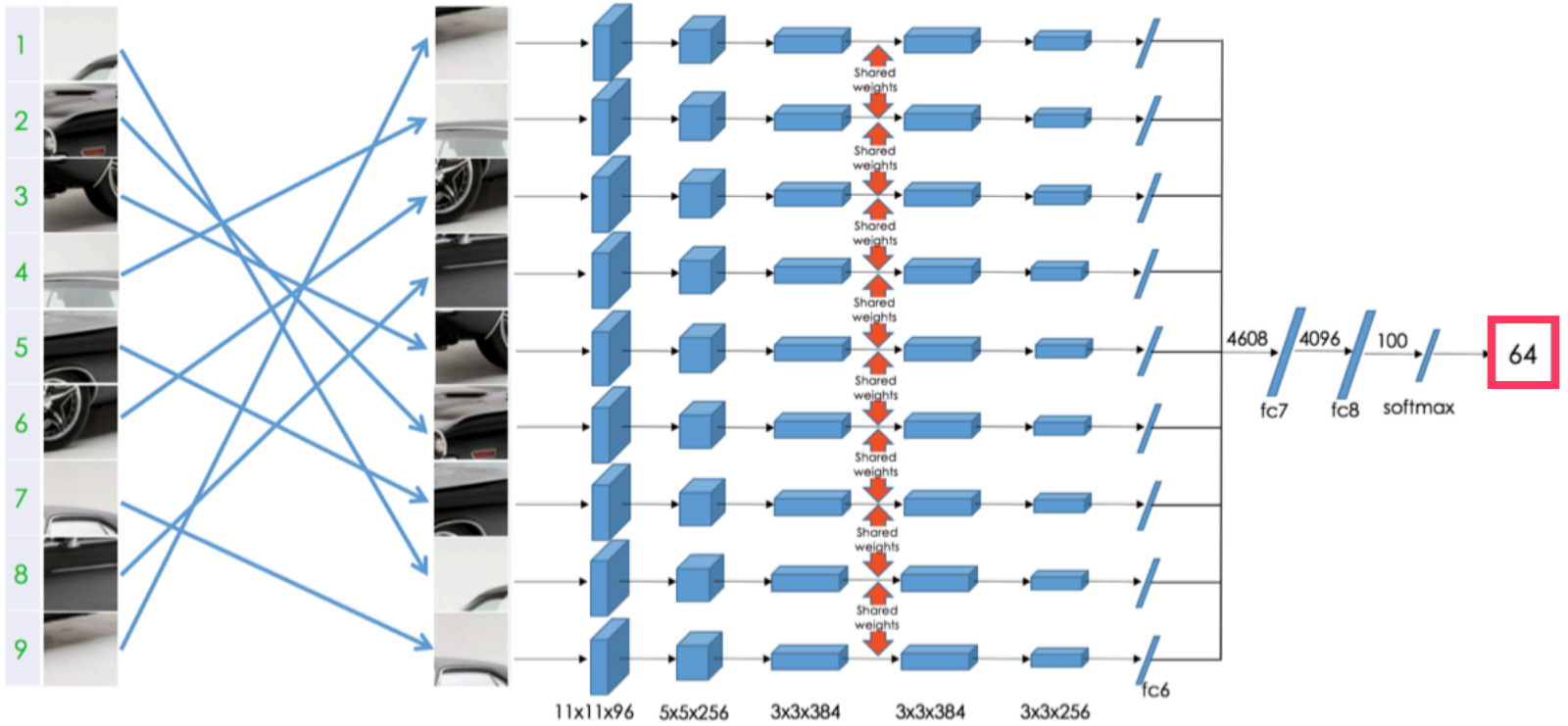
Context-Free Architecture

ROBOT INTELLIGENCE LAB



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

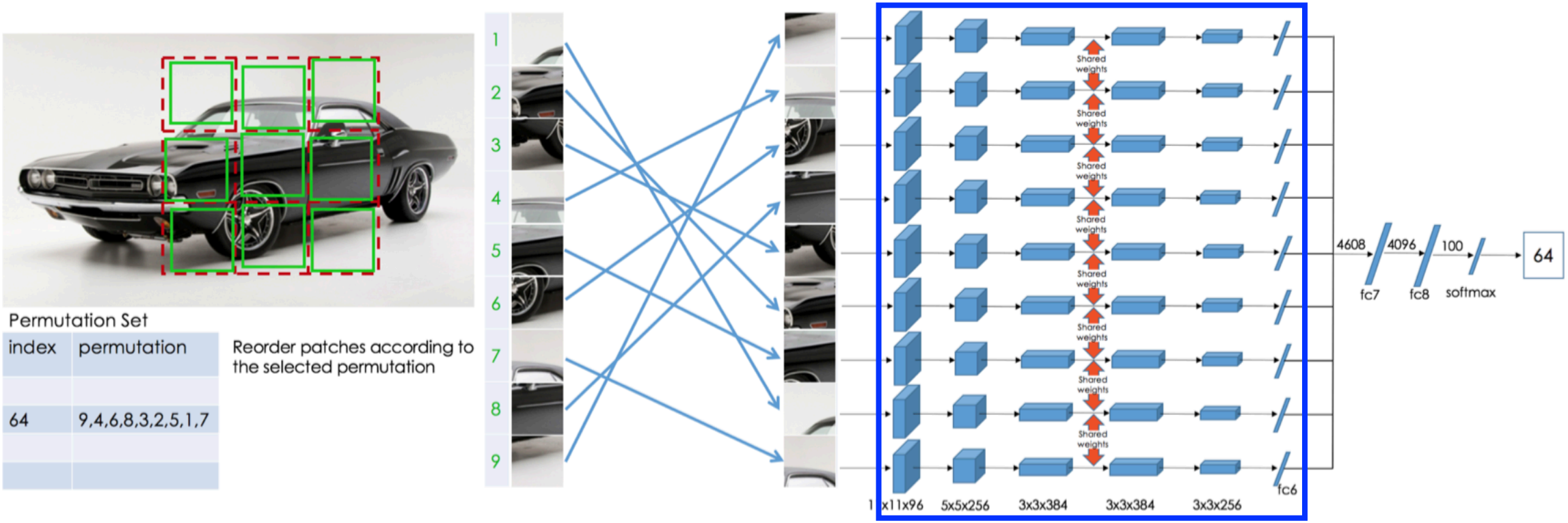
Reorder patches according to the selected permutation



The number of total permutation is $9! = 362{,}880$. However, a subset of a possible permutation set is defined (of size 64) and cast the reassembly problem as a classification problem.

Context-Free Architecture

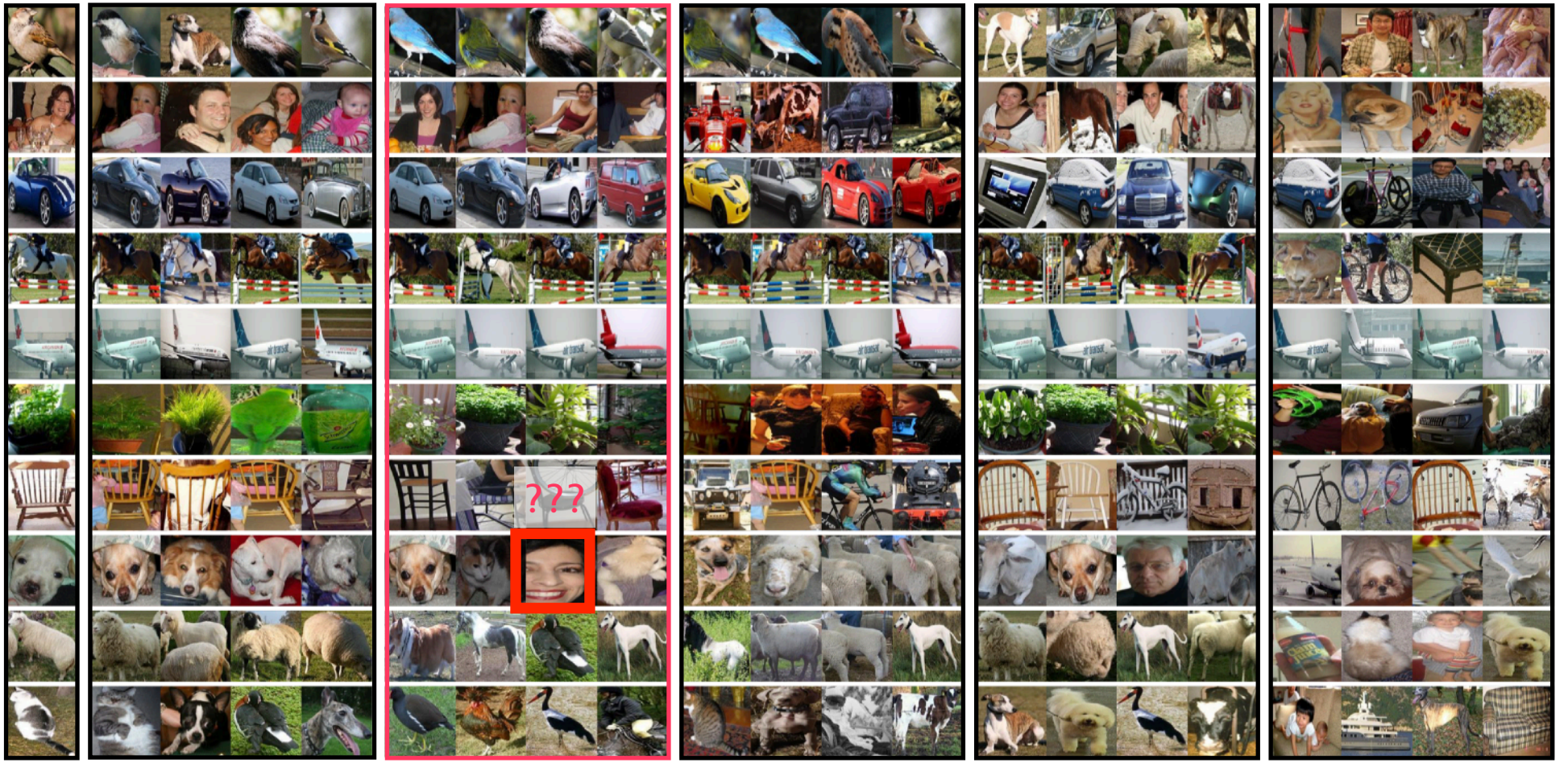
ROBOT INTELLIGENCE LAB



FCN utilizes AlexNet architecture with **weight-sharing** up-to the $fc6$ layer. Nine feature representation vectors are concatenated to predict the permutation class.

Image Retrieval

ROBOT INTELLIGENCE LAB



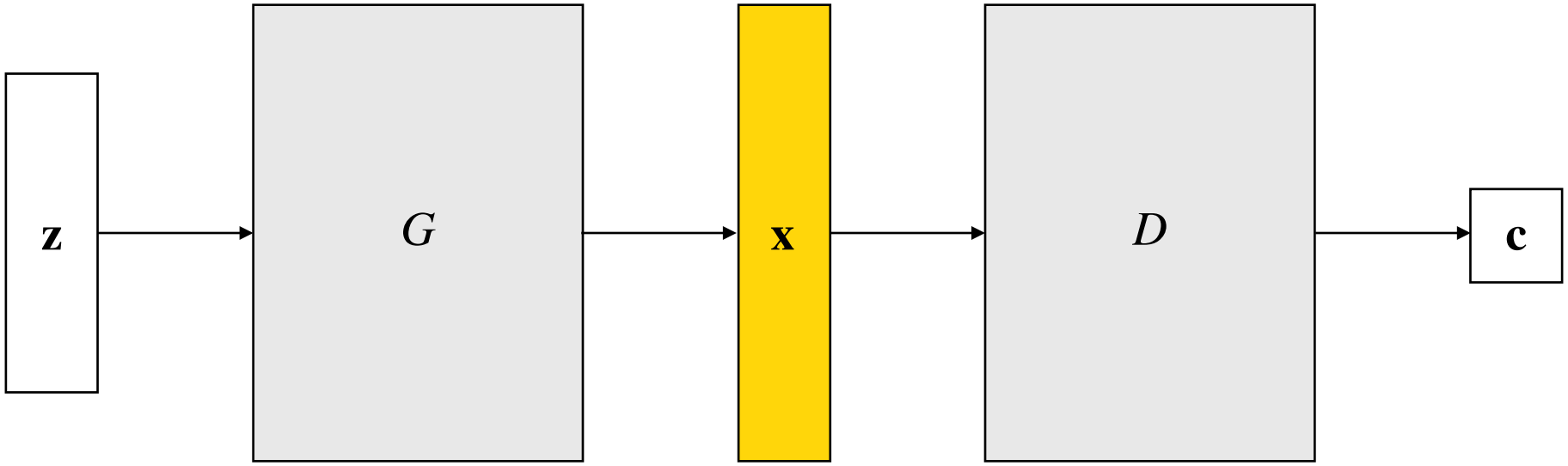
ROBOT INTELLIGENCE LAB

BiGAN

"ADVERSARIAL FEATURE LEARNING," 2017

Generative Adversarial Net

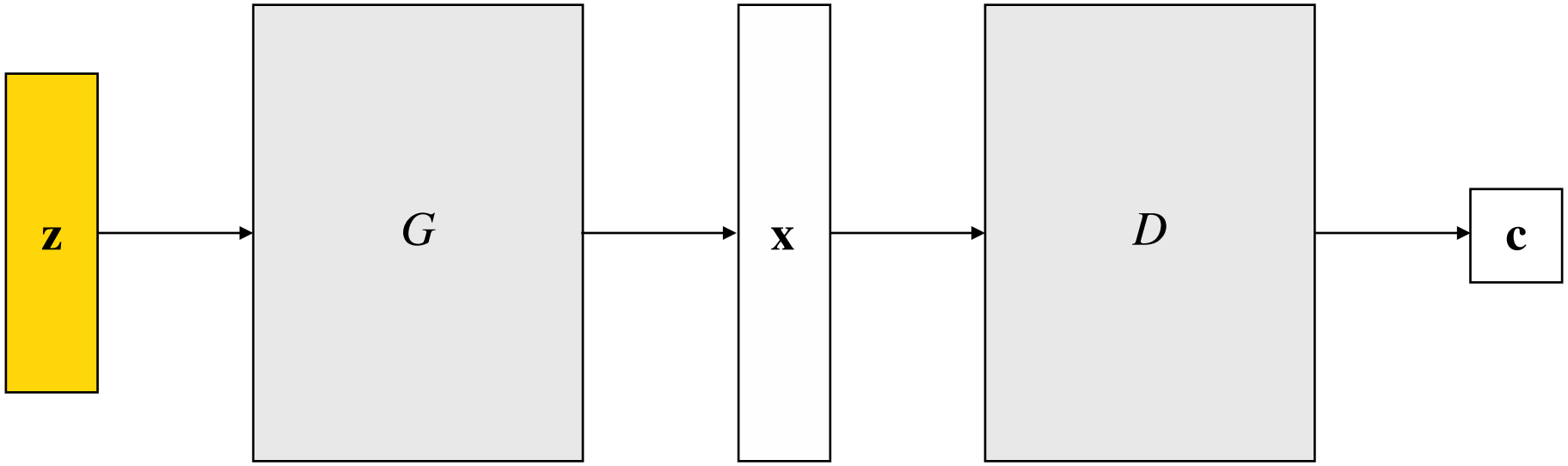
ROBOT INTELLIGENCE LAB



The objective of GAN is to minimize the distance between $p(\mathbf{x})$ and $G(\mathbf{z})$ where $\mathbf{z}$ follows a simple distribution.

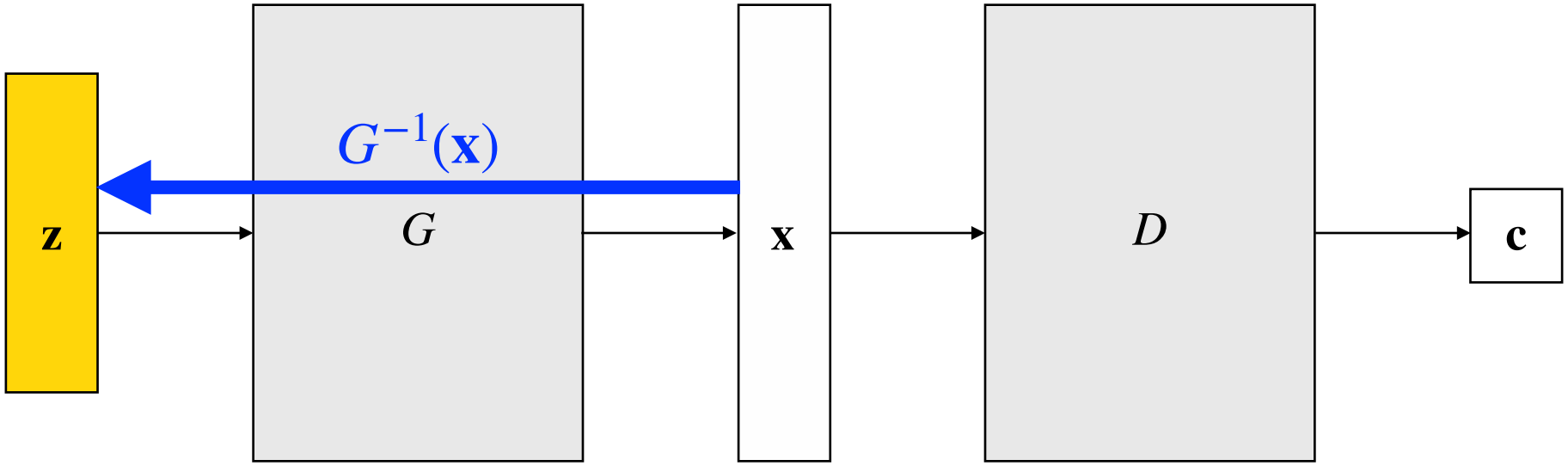
Generative Adversarial Net

ROBOT INTELLIGENCE LAB



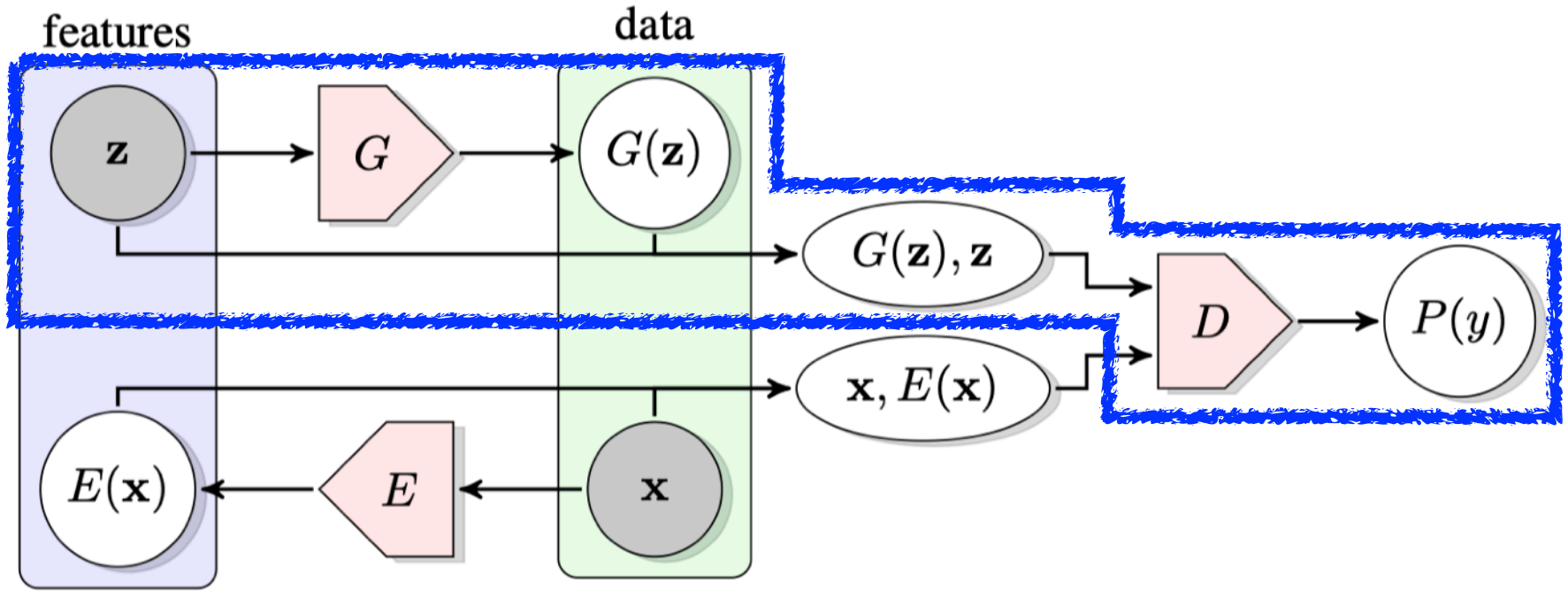
A latent vector $\mathbf{z}$ corresponds to an input $\mathbf{x}$, which can be thought as meaningful representation of $\mathbf{x}$.

Generative Adversarial Net



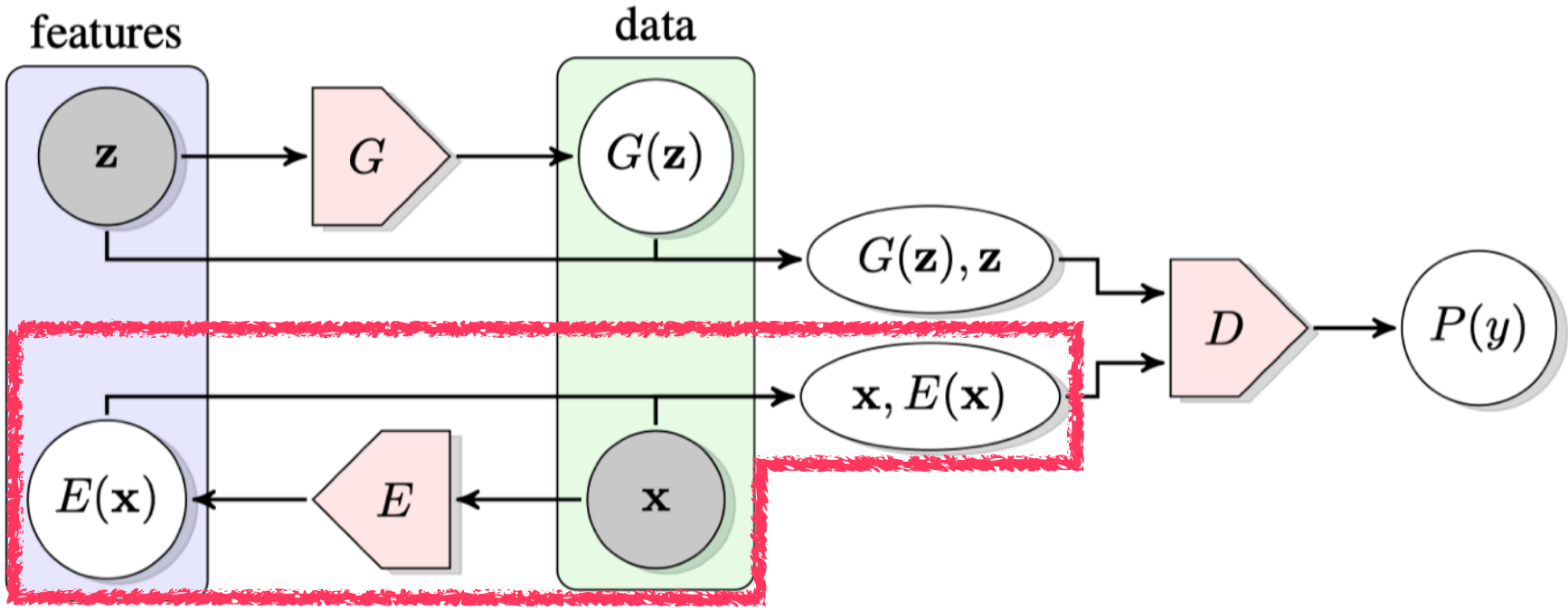
However, there is one remaining problem. G^{-1} does not exist in general.

Bidirectional GAN (BiGAN)



This part is the same as the original GAN

BiGAN



This part is newly added in BiGAN.

BiGAN

ROBOT INTELLIGENCE LAB

The BiGAN training objective is defined as a minimax objective

$$\min_{G,E} \max_{D} V(D, E, G) \tag{2}$$

where

$$V(D, E, G) := \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}}} \big[\underbrace{\mathbb{E}_{\mathbf{z} \sim p_E(\cdot|\mathbf{x})} [\log D(\mathbf{x}, \mathbf{z})]}_{\log D(\mathbf{x}, E(\mathbf{x}))} \big] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{Z}}} \big[\underbrace{\mathbb{E}_{\mathbf{x} \sim p_G(\cdot|\mathbf{z})} [\log (1 - D(\mathbf{x}, \mathbf{z}))]}_{\log(1 - D(G(\mathbf{z}), \mathbf{z}))} \big]. \tag{3}$$

The inputs of the **discriminator** are both $\mathbf{x}$ and $\mathbf{z}$, $D(\mathbf{x}, \mathbf{z})$, (the original discriminator only takes $\mathbf{x}$, i.e., $D(\mathbf{x})$).

BiGAN

ROBOT INTELLIGENCE LAB

The BiGAN training objective is defined as a minimax objective

$$\min_{G,E} \max_{D} V(D, E, G) \tag{2}$$

where

$$V(D, E, G) := \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}}} \Big[\underbrace{\mathbb{E}_{\mathbf{z} \sim p_E(\cdot|\mathbf{x})} [\log D(\mathbf{x}, \mathbf{z})]}_{\log D(\mathbf{x}, E(\mathbf{x}))} \Big] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{Z}}} \Big[\underbrace{\mathbb{E}_{\mathbf{x} \sim p_G(\cdot|\mathbf{z})} [\log (1 - D(\mathbf{x}, \mathbf{z}))]}_{\log(1 - D(G(\mathbf{z}), \mathbf{z}))} \Big]. \tag{3}$$

The newly added encoder maps an input $\mathbf{x}$ to a latent vector $\mathbf{z}$.

BiGAN

ROBOT INTELLIGENCE LAB

The BiGAN training objective is defined as a minimax objective

$$\min_{G,E} \max_{D} V(D, E, G) \tag{2}$$

where

$$V(D,E,G) := \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\Big[\underbrace{\mathbb{E}_{\mathbf{z}\sim p_E(\cdot|\mathbf{x})}\left[\log D(\mathbf{x},\mathbf{z})\right]}_{\log D(\mathbf{x},E(\mathbf{x}))}\Big] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{Z}}}\Big[\underbrace{\mathbb{E}_{\mathbf{x}\sim p_G(\cdot|\mathbf{z})}\left[\log\left(1-D(\mathbf{x},\mathbf{z})\right)\right]}_{\log(1-D(G(\mathbf{z}),\mathbf{z}))}\Big]. \tag{3}$$

To maximize $V(\cdot)$ w.r.t. $D(\cdot)$, the pair of a real image $\mathbf{x}$ and the encoded latent vector $\mathbf{z}' = E(\mathbf{x})$ gets higher value while the pair of a fake image $G(\mathbf{z})$ and the underlying latent vector $\mathbf{z}$ get lower values.

BiGAN

ROBOT INTELLIGENCE LAB

The BiGAN training objective is defined as a minimax objective

$$\min_{G,E} \max_{D} V(D, E, G) \tag{2}$$

where

$$V(D,E,G) := \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\Big[\underbrace{\mathbb{E}_{\mathbf{z}\sim p_E(\cdot|\mathbf{x})}[\log D(\mathbf{x},\mathbf{z})]}_{\log D(\mathbf{x},E(\mathbf{x}))}\Big] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{Z}}}\Big[\underbrace{\mathbb{E}_{\mathbf{x}\sim p_G(\cdot|\mathbf{z})}[\log(1-D(\mathbf{x},\mathbf{z}))]}_{\log(1-D(G(\mathbf{z}),\mathbf{z}))}\Big]. \tag{3}$$

To maximize $V(\cdot)$ w.r.t. $D(\cdot)$, the pair of a real image $\mathbf{x}$ and the encoded latent vector $\mathbf{z}' = E(\mathbf{x})$ gets higher value while the pair of a fake image $G(\mathbf{z})$ and the underlying latent vector $\mathbf{z}$ get lower values.

To minimize $V(\cdot)$ w.r.t. $E(\cdot)$, the encoder $E(\mathbf{x})$ mimics the inverse of $G(\mathbf{z})$, i.e., $G^{-1}(\mathbf{x})$.

BiGAN

ROBOT INTELLIGENCE LAB

The BiGAN training objective is defined as a minimax objective

$$\min_{G,E} \max_{D} V(D, E, G) \tag{2}$$

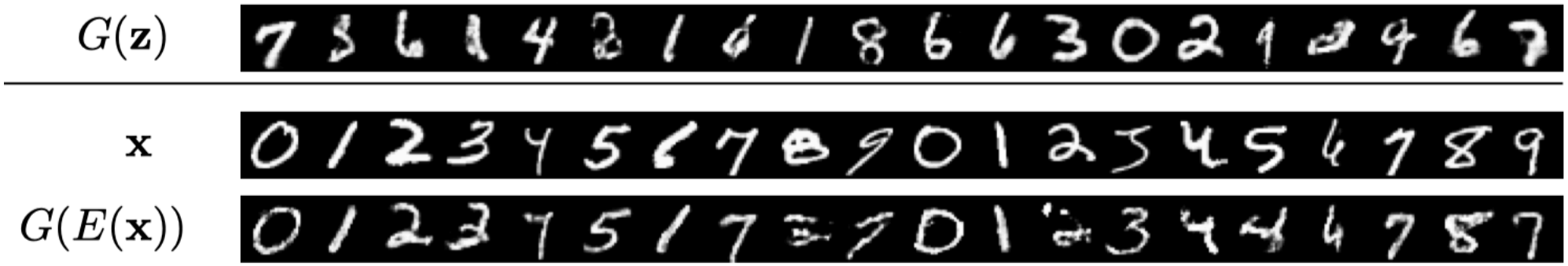
where

$$V(D, E, G) := \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}}} \big[\underbrace{\mathbb{E}_{\mathbf{z} \sim p_E(\cdot|\mathbf{x})} [\log D(\mathbf{x}, \mathbf{z})]}_{\log D(\mathbf{x}, E(\mathbf{x}))} \big] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{Z}}} \big[\underbrace{\mathbb{E}_{\mathbf{x} \sim p_G(\cdot|\mathbf{z})} [\log (1 - D(\mathbf{x}, \mathbf{z}))]}_{\log(1 - D(G(\mathbf{z}), \mathbf{z}))} \big]. \tag{3}$$

An interesting idea of BiGAN is that the training objective of the encoder $E(\mathbf{x})$ is **NOT** directily minimizing $\|E\left(G(\mathbf{z})\right) - \mathbf{z}\|_2^2$.

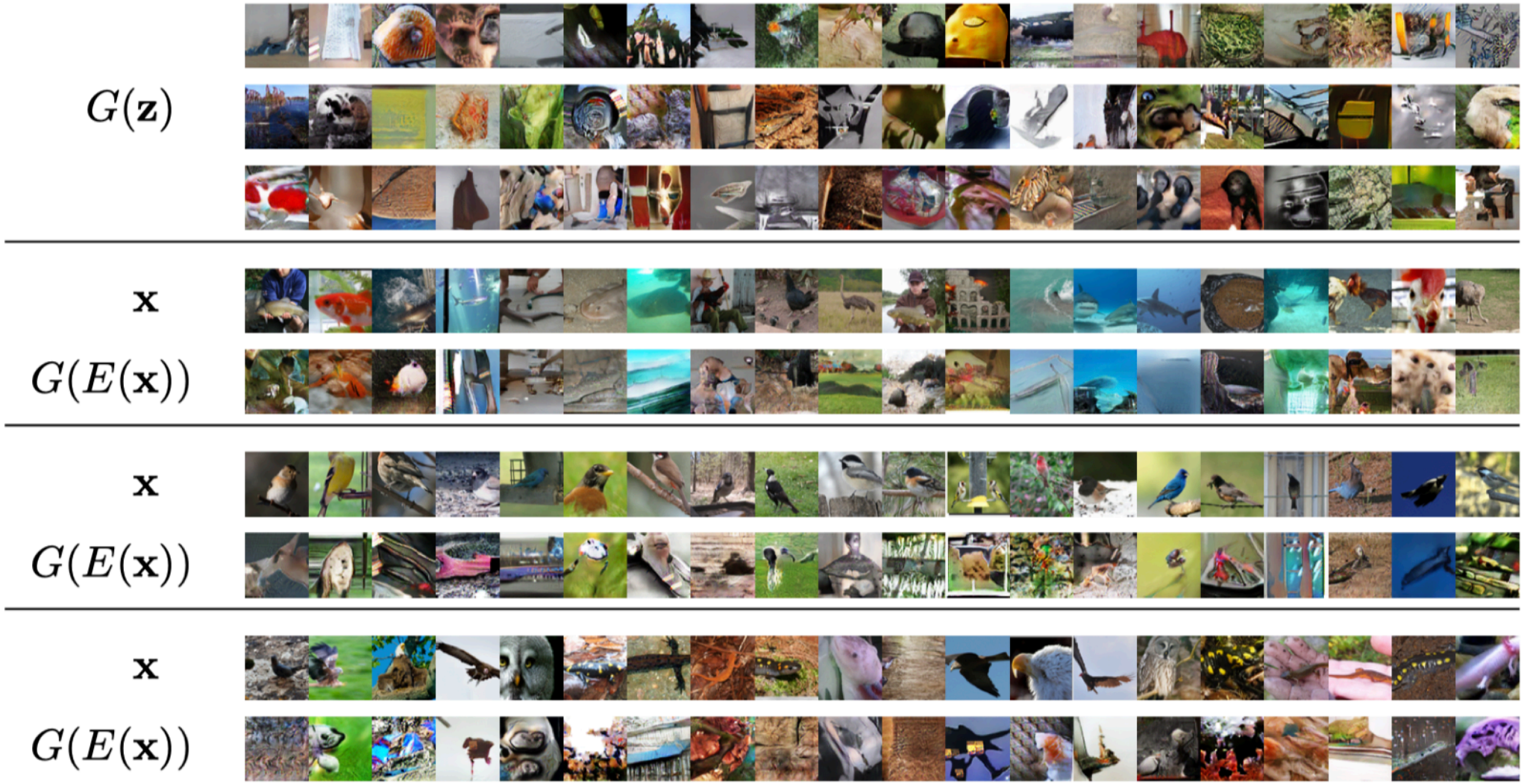
Reconstruction

ROBOT INTELLIGENCE LAB



Reconstruction

ROBOT INTELLIGENCE LAB



ROBOT INTELLIGENCE LAB

RotNet

"UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS," 2018

Intuition

ROBOT INTELLIGENCE LAB

What kind of **supervision** can get out of these images?

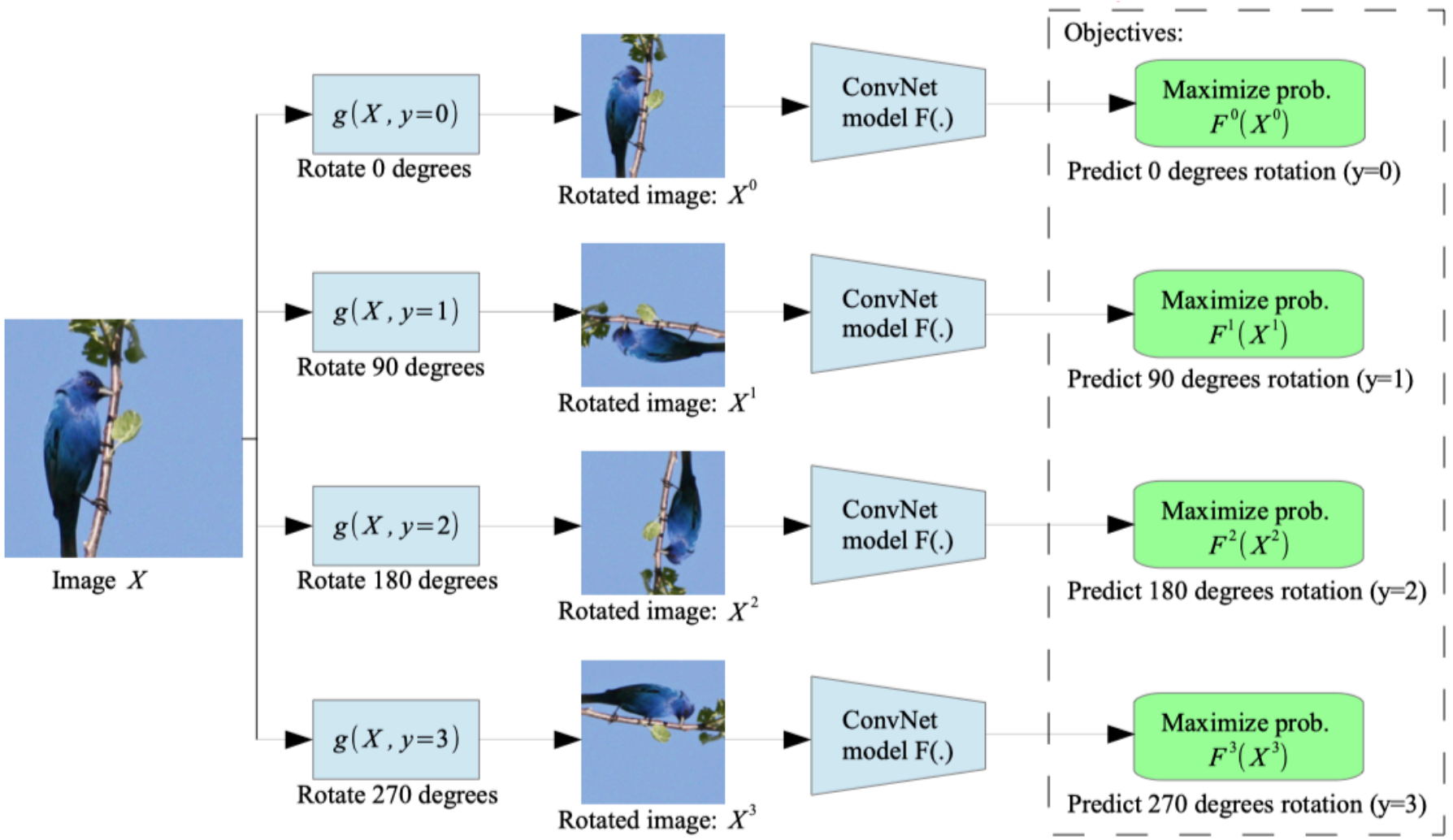
Intuition

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

Predict the **rotated angle** of the original image!

Self-supervised Task

ROBOT INTELLIGENCE LAB



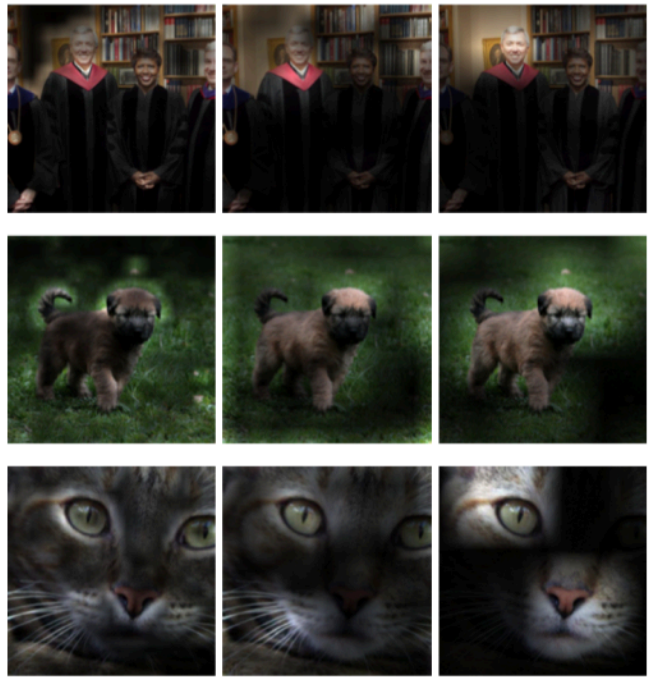
Attention Maps

ROBOT INTELLIGENCE LAB

Input images on the models

Conv1 27 × 27 Conv3 13 × 13 Conv5 6 × 6

(a) **Attention maps of supervised model**



Conv1 27 × 27 Conv3 13 × 13 Conv5 6 × 6

(b) **Attention maps of our self-supervised model**

*Attention map is computed based on the magnitude of activations at each spatial cell.

Experiments

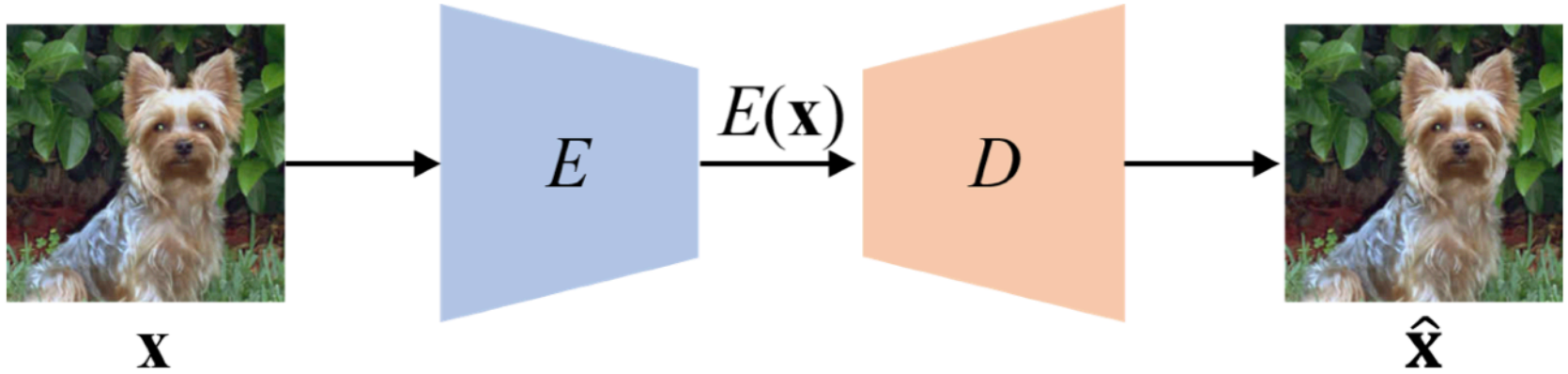
ROBOT INTELLIGENCE LAB

- Multiple experiments including unsupervised feature learning for
  - ImageNet classification
  - PASCAL classification
  - PASCAL detection
  - PASCAL segmentation
  - CIFAR-10 classification

Auto-Encoding Transformations

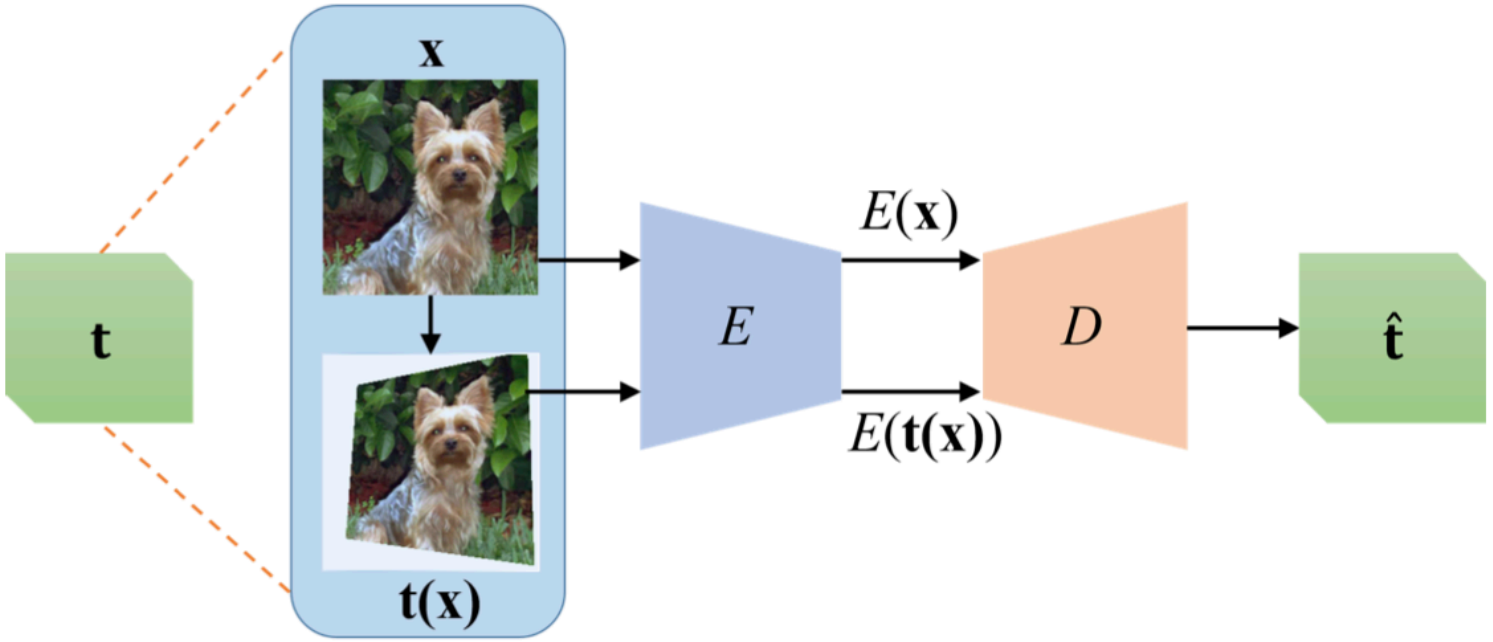
"AET vs. AED: Unsupervised Representation Learning by Auto-Encoding Transformations rather than Data," 2019

Auto-Encoders



Auto-Encoding Data (AED) has been widely used for unsupervised learning.

Auto-Encoding Transformation



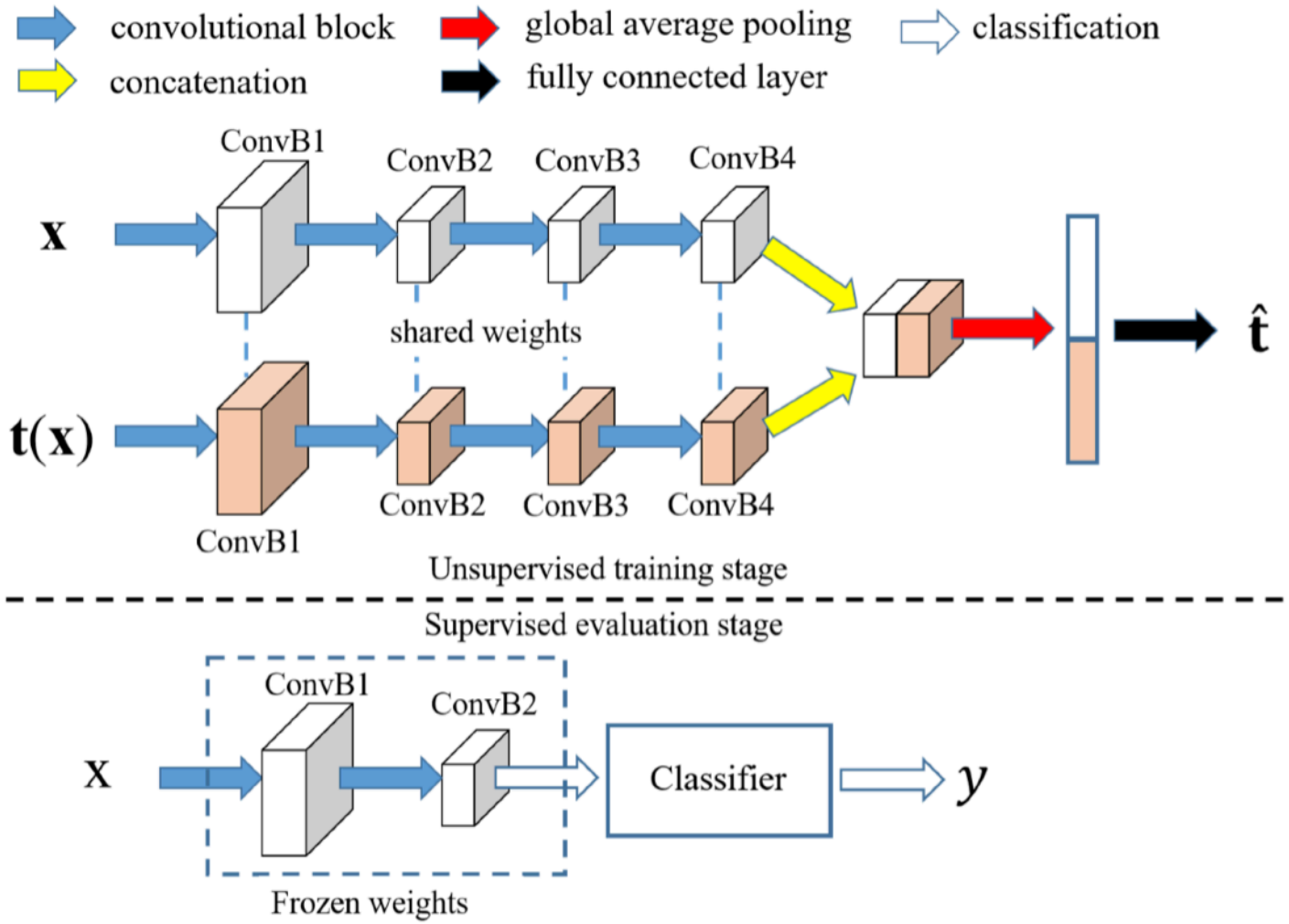
This paper presents **auto-encoding transformation (AET)** as a mean of self-supervised learning.

Problem Formulation

ROBOT INTELLIGENCE LAB

- **Encoder**
  - $E : \mathbf{x} \mapsto E(\mathbf{x})$ extracts the **representation** of a sample (image)
- **Transformation**
  - $\mathbf{t} : \mathbf{x} \mapsto \mathbf{t}(\mathbf{x})$ where a transformation is sampled from a distribution $\mathcal{T}$
  - The transform distribution $\mathcal{T} = \{\mathbf{t}_\theta | \theta \sim \Theta\}$ is parametrized by θ.
    - Projective and affine transformations: $M(\theta) \in \mathbb{R}^{3\times3}$ where the loss becomes $l(\mathbf{t}_\theta, \mathbf{t}_{\hat{\theta}}) = \frac{1}{2}\|M(\theta) - M(\hat{\theta})\|_2^2$
    - GAN-Induced transformations: $\mathbf{t}_\mathbf{z}(\mathbf{x}) = G(\mathbf{x}, \mathbf{z})$ where the loss becomes $l(\mathbf{t}_\mathbf{z} - \mathbf{t}_{\hat{\mathbf{z}}}) = \frac{1}{2}\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$
- **Decoder**
  - $D : \left[E(\mathbf{x}), E(\mathbf{t}(\mathbf{x}))\right] \mapsto \hat{t}$ **predicts the input transformation** from the original and transformed images.

Network Architectures



CIFAR-10

ROBOT INTELLIGENCE LAB

Method	Error rate
Supervised NIN (Lower Bound)	7.20
Random Init. + conv (Upper Bound)	27.50
Roto-Scat + SVM [22]	17.7
ExamplarCNN [7]	15.7
DCGAN [26]	17.2
Scattering [21]	15.3
RotNet + FC [10]	10.94
RotNet + conv [10]	8.84
(Ours) AET-affine + FC	9.77
(Ours) AET-affine + conv	8.05
(Ours) AET-project + FC	**9.41**
(Ours) AET-project + conv	**7.82**

ImageNet (Top-1 Accuracy)

ROBOT INTELLIGENCE LAB

Method	Conv4	Conv5
ImageNet Labels [3](Upper Bound)	59.7	59.7
Random [20] (Lower Bound)	27.1	12.0
Tracking [29]	38.8	29.8
Context [5]	45.6	30.4
Colorization [31]	40.7	35.2
Jigsaw Puzzles [19]	45.3	34.6
BiGAN [6]	41.9	32.2
NAT [3]	-	36.0
DeepCluster [4]	-	44.0
RotNet [10]	50.0	43.8
(Ours) AET-project	**53.2**	**47.0**

Reconstruction Results

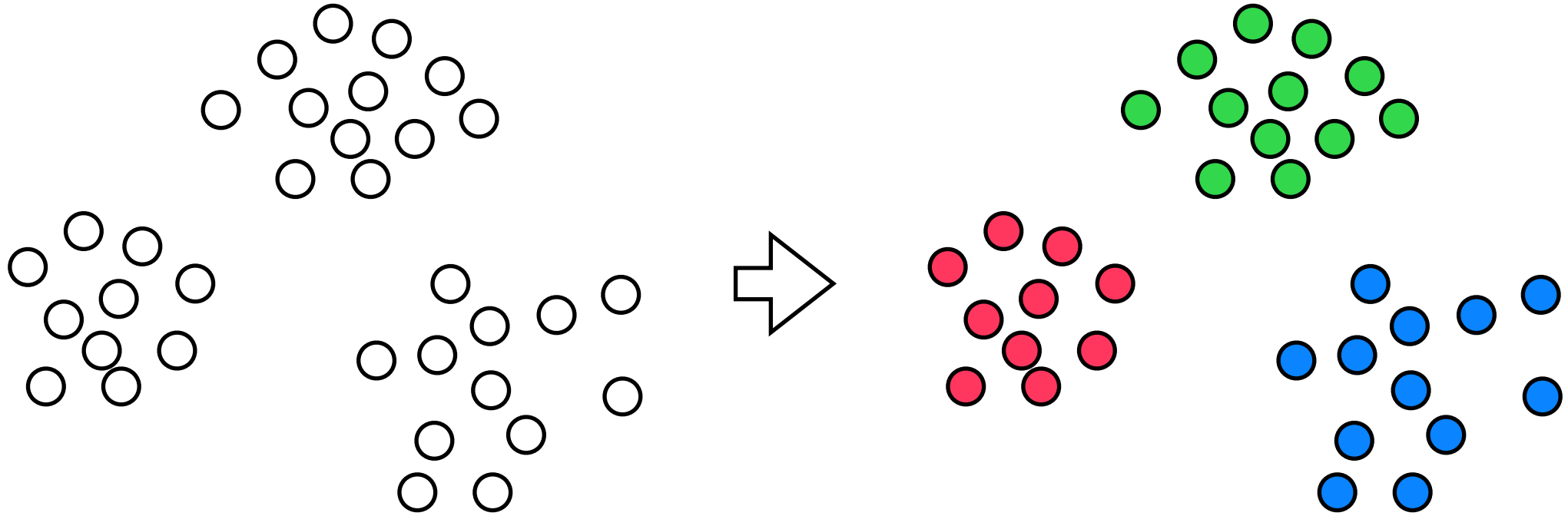
ROBOT INTELLIGENCE LAB

DeepCluster

"Deep Clustering for Unsupervised Learning of Visual Features," 2019

Clustering

ROBOT INTELLIGENCE LAB



Conventional clustering methods work fairly well on **low-dimensional data** where **distance measures** can easily be defined and evaluated.

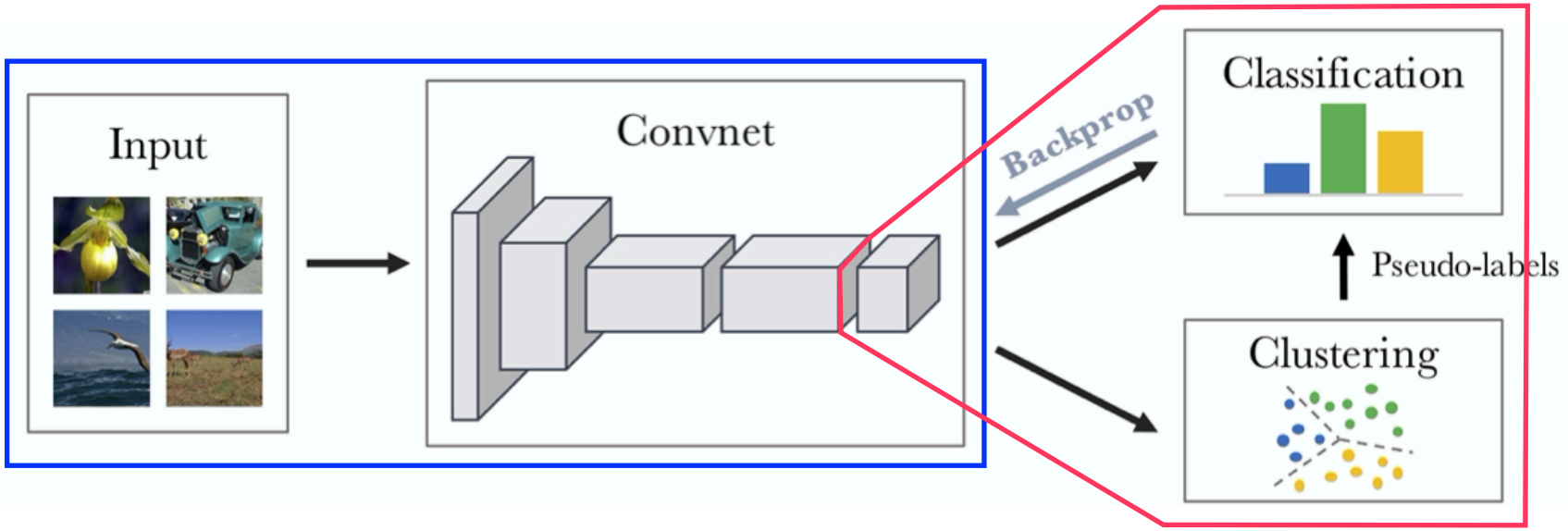
Distance Measure

ROBOT INTELLIGENCE LAB

However, in high-dimensional space, computing distances between instances becomes **unintuitive**.

Deep Clustering

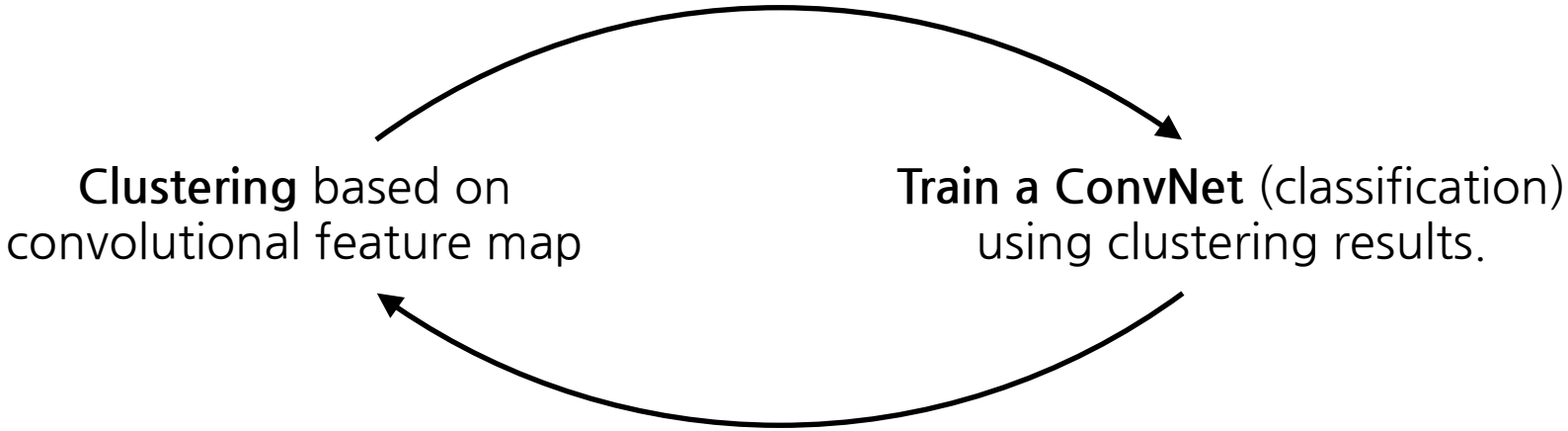
ROBOT INTELLIGENCE LAB



This paper combines the **representation learning power** of a Convnet with an **auxiliary learning objectives** from clustering results.

Deep Clustering

ROBOT INTELLIGENCE LAB



Results

ROBOT INTELLIGENCE LAB

Method	ImageNet					Places				
	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels	–	–	–	–	–	22.1	35.1	40.2	43.3	44.6
ImageNet labels	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
Pathak *et al.* [38]	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
Doersch *et al.* [25]	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Zhang *et al.* [28]	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
BiGAN Donahue *et al.* [20]	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
Jigsaw Noroozi and Favaro [26]	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Noroozi *et al.* [45]	18.0	**30.6**	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
Zhang *et al.* [43]	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
DeepCluster	12.9	29.2	**38.2**	**39.8**	**36.1**	18.6	30.8	**37.0**	**37.5**	**33.1**

ROBOT INTELLIGENCE LAB

Single Image SSL

"A CRITICAL ANALYSIS OF SELF-SUPERVISION , WHAT WE CAN LEARN FROM A SINGLE IMAGE," 2020

Question

ROBOT INTELLIGENCE LAB

Is self-supervision **able to exploit** the information contained in a large number of images in order to learn different parts of a neural network?

Findings

ROBOT INTELLIGENCE LAB

1. **A single image** is sufficient, when combined with self-supervision and data augmentation, to learn the **first few layers** of standard deep network as well as using millions of images and full supervision.

2. For the deeper layers of the network, self-supervision remains inferior to supervision even if millions of images are used for training.

SSL Methods

- BiGAN ("ADVERSARIAL FEATURE LEARNING," 2017)
- RotNet ("UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS," 2018)
- DeepCluster ("Deep Clustering for Unsupervised Learning of Visual Features," 2019)

Data

- **Baseline**: trained on d source images (e.g., $d = 1{,}281{,}167$ for ImageNet)
- **SSL methods**: $N \ll d$ source images (i.i.d. samples) and the remaining $d - N$ images are augmentations of the source images
  - N controls the **amount of information** in the training data, and if $N = 1$, we are using a single image for training.
- **Augmentations**: cropping, scaling, rotation, contrast changes, and adding noise

Real Samples

ROBOT INTELLIGENCE LAB

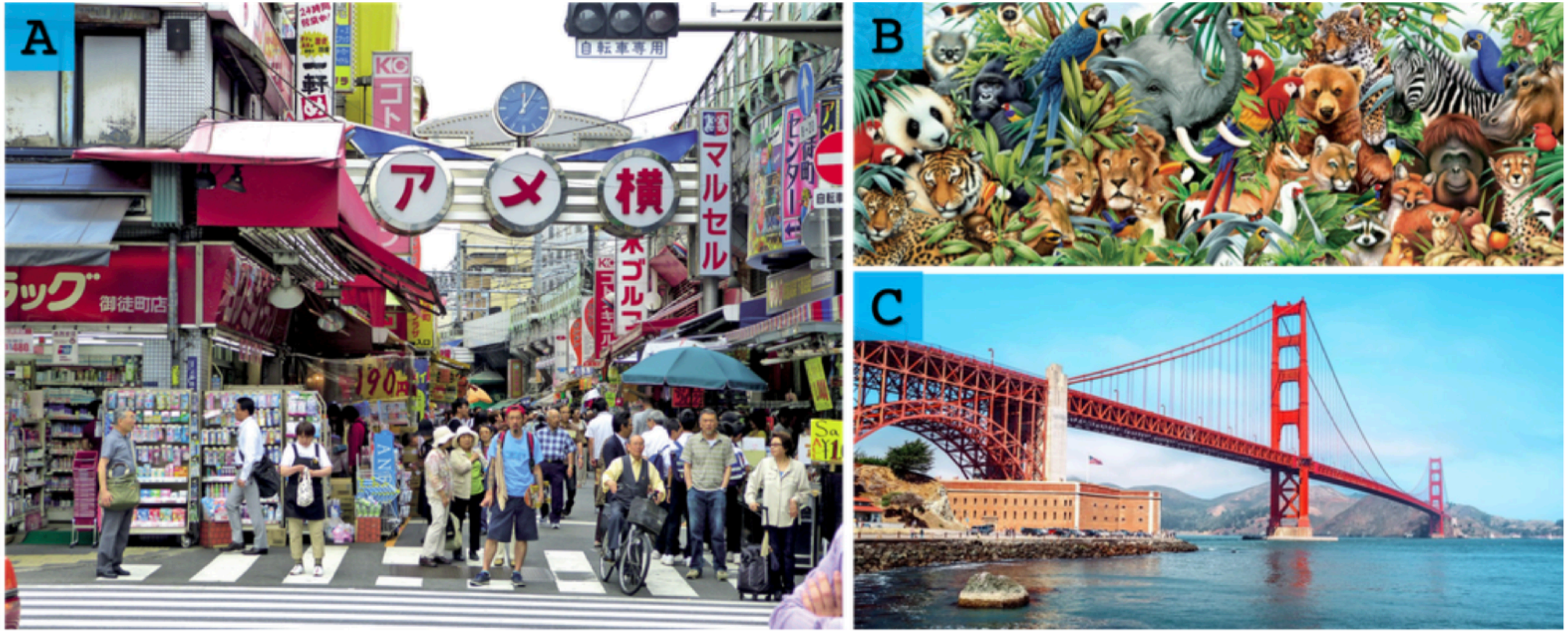


Image **A** and **B**: contain rich texture
Image **C**: ablates "crowdness" as it contains large areas covering no objects

Linear Probes

ROBOT INTELLIGENCE LAB

Solve a difficult task such as ImageNet classification by training **a linear classifier** on top of pre-trained feature representations, which are kept fixed.

ImageNet

ROBOT INTELLIGENCE LAB

	Method, Reference	#images	ILSVRC-12 conv1	conv2	conv3	conv4	conv5
(a)	Full-supervision‡	1,281,167	19.3	36.3	44.2	48.3	50.5
(b)	(Oyallon et al., 2017): Scattering	0	-	18.9	-	-	-
(c)	Random‡	0	11.6	17.1	16.9	16.3	14.1
(d)	(Krähenbühl et al., 2016):k-means‡	≈160	17.5	23.0	24.5	23.2	20.6
(e)	(Donahue et al., 2017): BiGAN‡	1,281,167	17.7	24.5	31.0	29.9	28.0
(f)	mono, Image A	1	20.4	30.9	33.4	28.4	16.0
(g)	mono, Image B	1	20.5	30.4	31.6	27.0	16.8
(h)	deka	10	16.2	16.5	16.5	13.1	7.5
(i)	kilo	1,000	16.1	17.7	18.3	17.6	13.5
(j)	(Gidaris et al., 2018): RotNet	1,281,167	18.8	31.7	38.7	38.2	36.5
(k)	mono, Image A	1	19.9	30.2	30.6	27.6	21.9
(l)	mono, Image B	1	17.8	27.6	27.9	25.4	20.2
(m)	deka	10	19.6	30.7	32.6	28.9	22.6
(n)	kilo	1,000	21.0	33.5	36.5	34.0	29.4
(o)	(Caron et al., 2018): DeepCluster	1,281,167	18.0	32.5	39.2	37.2	30.6
(p)	mono, Image A	1	20.7	31.5	32.5	28.5	21.0
(q)	mono, Image B	1	19.7	30.1	31.6	28.5	20.4
(r)	mono, Image C	1	18.9	29.2	31.5	28.9	23.5
(s)	deka	10	18.5	29.0	31.1	28.2	21.9
(t)	kilo	1,000	19.5	29.8	33.0	31.7	26.8

On lower layers, the performance gap between fully-supervised and self-supervised methods is small.

ImageNet

ROBOT INTELLIGENCE LAB

	Method, Reference	#images	ILSVRC-12 conv1	conv2	conv3	conv4	conv5
(a)	Full-supervision‡	1,281,167	19.3	36.3	44.2	48.3	50.5
(b)	(Oyallon et al., 2017): Scattering	0	-	18.9	-	-	-
(c)	Random‡	0	11.6	17.1	16.9	16.3	14.1
(d)	(Krähenbühl et al., 2016):k-means‡	≈160	17.5	23.0	24.5	23.2	20.6
(e)	(Donahue et al., 2017): BiGAN‡	1,281,167	17.7	24.5	31.0	29.9	28.0
(f)	mono, Image A	1	20.4	30.9	33.4	28.4	16.0
(g)	mono, Image B	1	20.5	30.4	31.6	27.0	16.8
(h)	deka	10	16.2	16.5	16.5	13.1	7.5
(i)	kilo	1,000	16.1	17.7	18.3	17.6	13.5
(j)	(Gidaris et al., 2018): RotNet	1,281,167	18.8	31.7	38.7	38.2	36.5
(k)	mono, Image A	1	19.9	30.2	30.6	27.6	21.9
(l)	mono, Image B	1	17.8	27.6	27.9	25.4	20.2
(m)	deka	10	19.6	30.7	32.6	28.9	22.6
(n)	kilo	1,000	21.0	33.5	36.5	34.0	29.4
(o)	(Caron et al., 2018): DeepCluster	1,281,167	18.0	32.5	39.2	37.2	30.6
(p)	mono, Image A	1	20.7	31.5	32.5	28.5	21.0
(q)	mono, Image B	1	19.7	30.1	31.6	28.5	20.4
(r)	mono, Image C	1	18.9	29.2	31.5	28.9	23.5
(s)	deka	10	18.5	29.0	31.1	28.2	21.9
(t)	kilo	1,000	19.5	29.8	33.0	31.7	26.8

However, on higher layers, SSL methods show inferior performance.

Conclusion

ROBOT INTELLIGENCE LAB

SSL with a single image can successfully capture the simplest image statistics (lower level features), but a gap exists with full supervision even when using larget datasets.

Conclusion

Good old fashioned SSL methods, more to come next week!

Thank You