Deep Generative models

Sewoong Oh

CSE446
University of Washington

Interpretability of Neural Networks

Interpreting neural network with saliency map

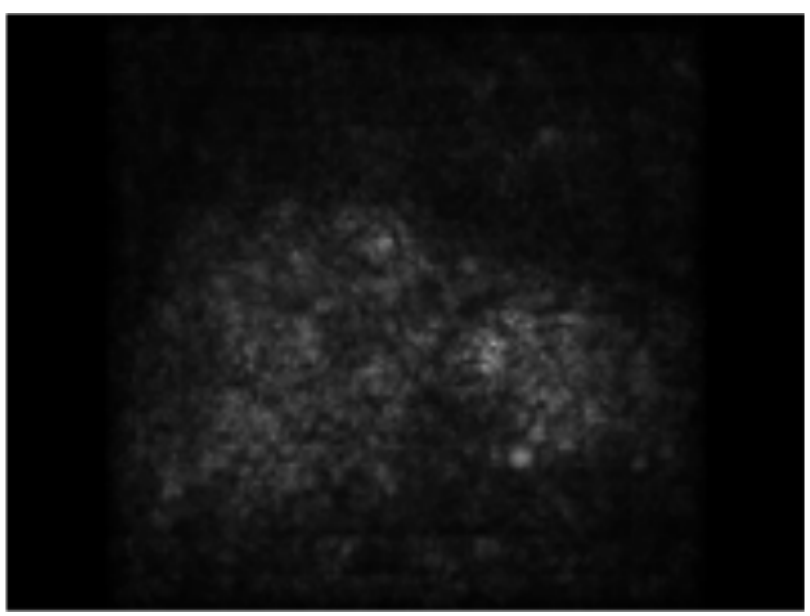
- why did neural network classify the image as a “dog”?

NN is trained to classify an input image

→ CNN → $f(x)$ → $\frac{e^{f_i(x)}}{\sum_j e^{f_j(x)}}$

$$\begin{bmatrix} 0.02 \\ 3.1 \\ -1.2 \\ \vdots \\ \vdots \end{bmatrix} \rightarrow \begin{array}{c} \begin{bmatrix} 0.05 \\ 0.87 \\ 0.01 \\ \vdots \\ \vdots \end{bmatrix} \end{array} \begin{array}{l} \text{“cat”} \\ \text{“dog”} \\ \text{“ant”} \\ \\ \\ \end{array}$$

Saliency map of the image x and the NN model $f(\,\cdot\,)$ is defined as $\nabla_x f_{\text{dog}}(x)$



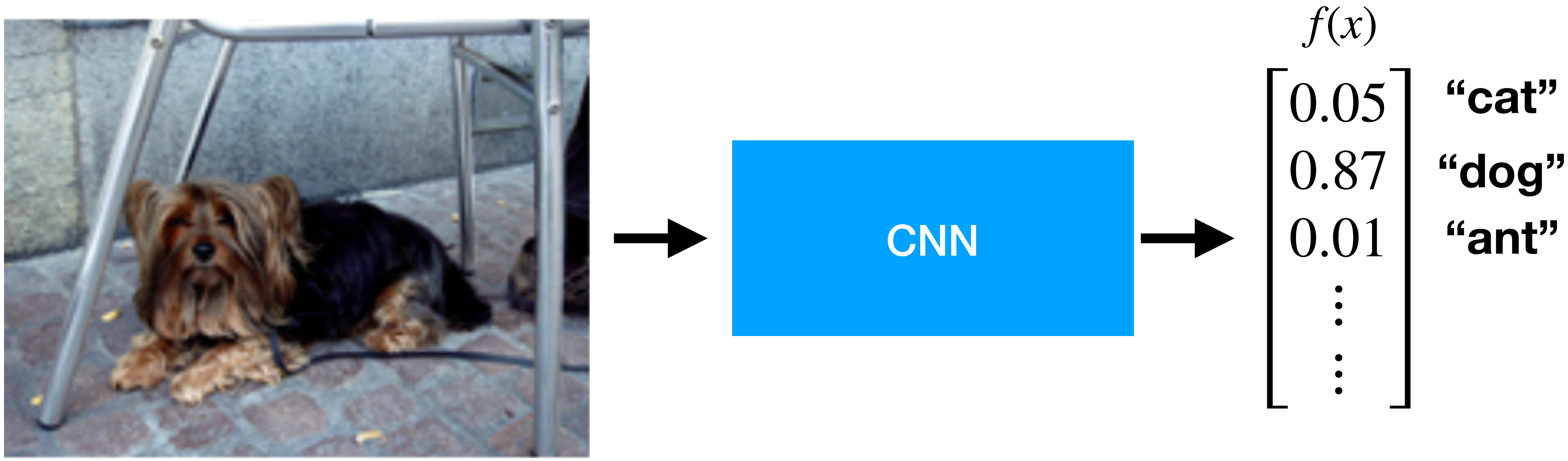
Saliency map

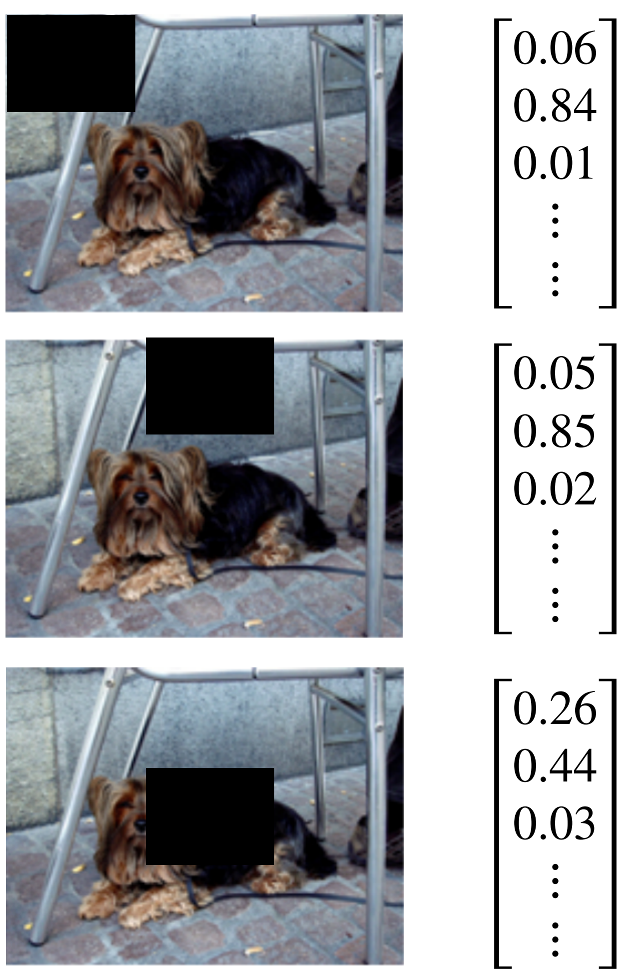
How much does this pixel contribute in classifying the image as a dog?

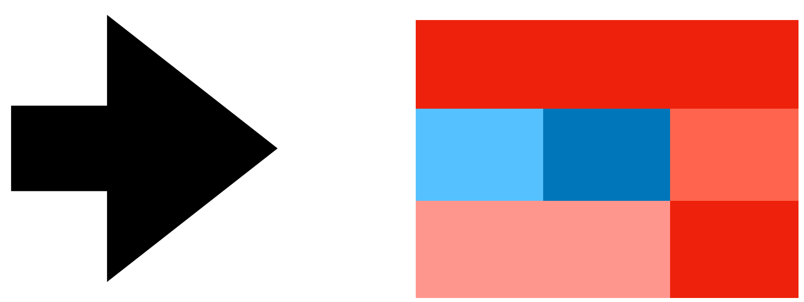
Segmenting those pixels with high saliency allows one to interpret NN decision

“Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”
Simonyan, Vedaldi, Zisserman, 2014

Interpreting neural networks with occlusion sensitivity



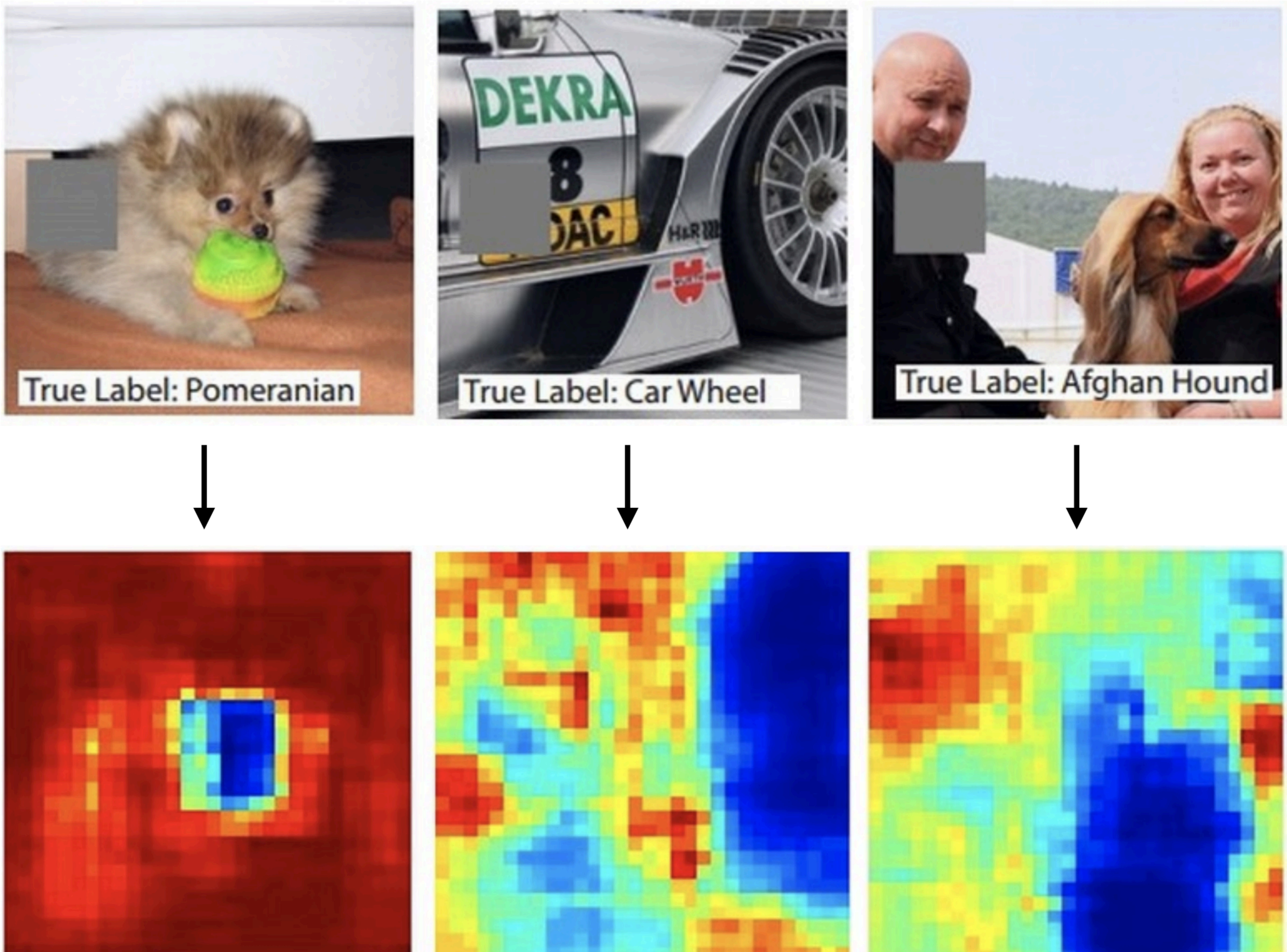




each occluded region is represented by $f_{\text{dog}}(x_{\text{occluded}})$ with a color map:
blue is low confidence and
red is high confidence

"Visualizing and Understanding Convolutional Networks", Zeiler, Fergus, 2013

Interpreting neural networks with occlusion sensitivity



Attacking Neural Networks with adversarial examples:

NNs are vulnerable

Attacking neural network with adversarial examples

- as an adversary, we want to generate an image that looks like a **cat**, but is classified as **iguana** (for a specific given NN classifier)

$\rightarrow$ CNN $\rightarrow$

$$\begin{array}{cl} \left[\begin{array}{c} 0.92 \\ 0.02 \\ 0.01 \\ \vdots \\ \vdots \end{array}\right] & \begin{array}{l} \textbf{“cat”} \\ \textbf{“dog”} \\ \textbf{“iguana”} \\ \\ \\ \end{array} \end{array}$$

Attacking neural network with adversarial examples

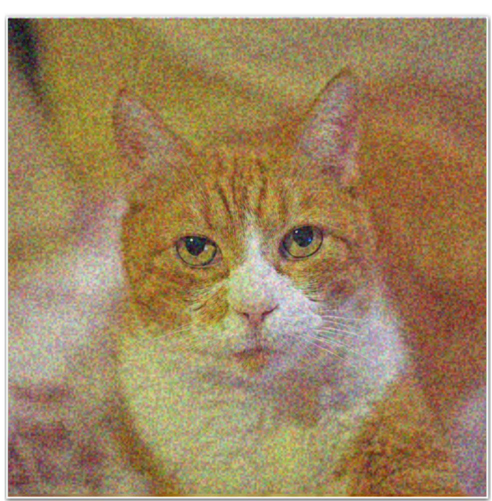
- as an adversary, we want to generate an image that looks like a **cat**, but is classified as **iguana** (for a specific given NN classifier)

x

→ CNN →

$$f_w(x) \quad \begin{bmatrix} 0.92 \\ 0.02 \\ 0.01 \\ \vdots \\ \vdots \end{bmatrix} \begin{array}{l} \text{“cat”} \\ \text{“dog”} \\ \text{“iguana”} \\ \\ \\ \end{array}$$

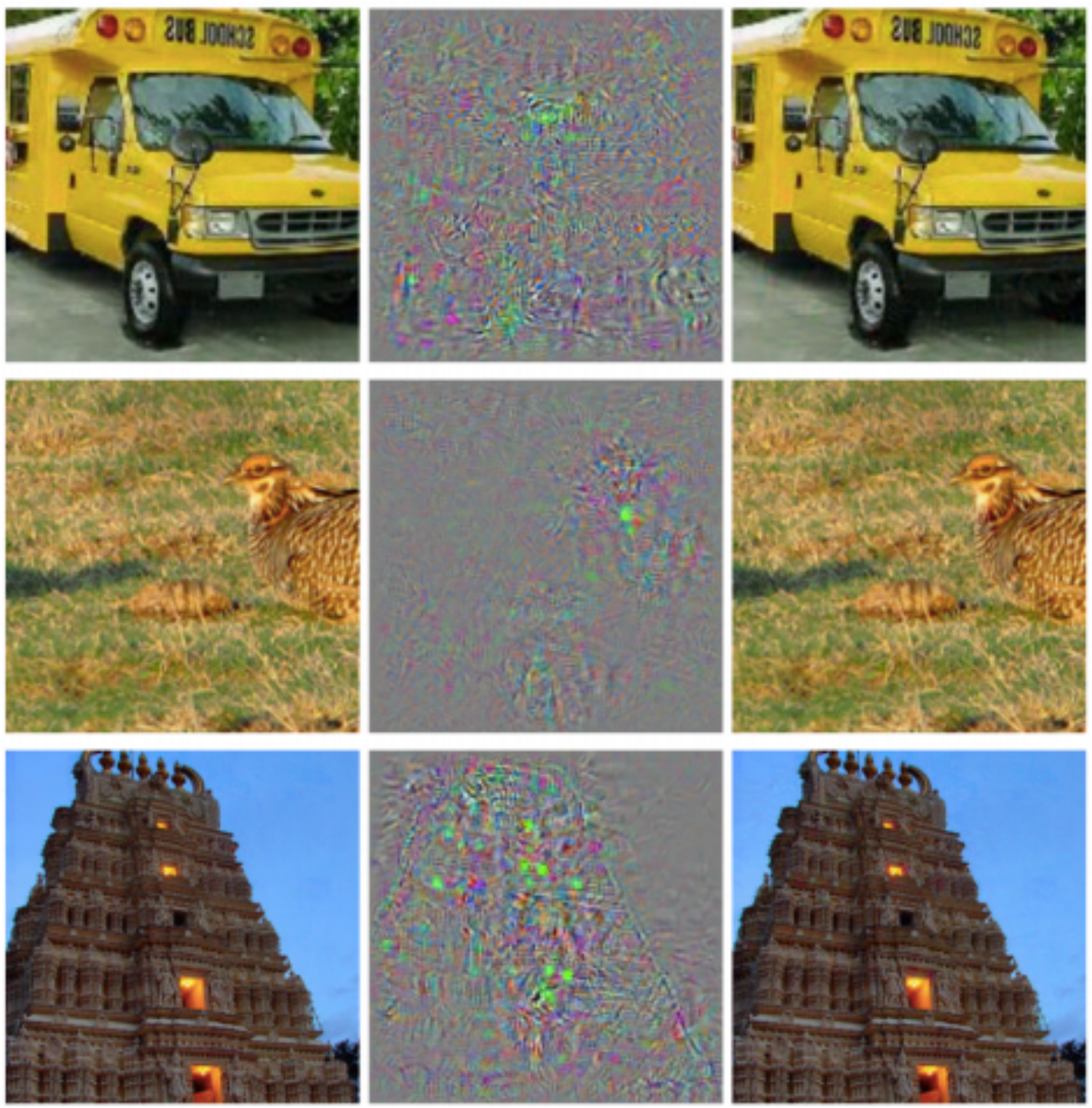
$x - \epsilon \,\text{sign}(\, \nabla_x \ell(f_w(x), y_{\text{iguana}})\,)$



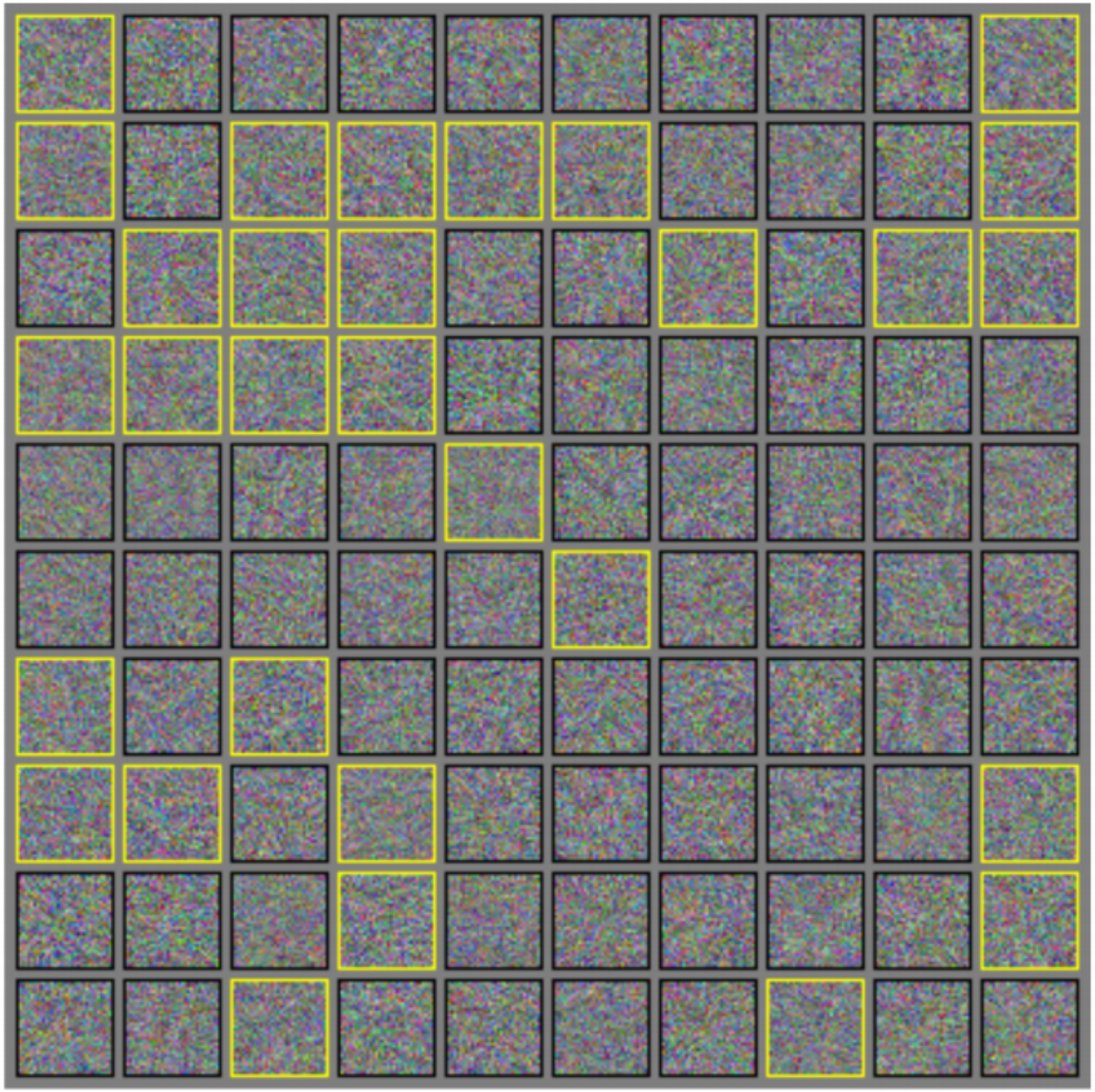
→ CNN →

$$f_w(x) \quad \begin{bmatrix} 0.02 \\ 0.02 \\ 0.94 \\ \vdots \\ \vdots \end{bmatrix} \begin{array}{l} \text{“cat”} \\ \text{“dog”} \\ \text{“iguana”} \\ \\ \\ \end{array}$$

- the adversarial examples are misclassified as ostriches, and in the middle we show the perturbation times ten.

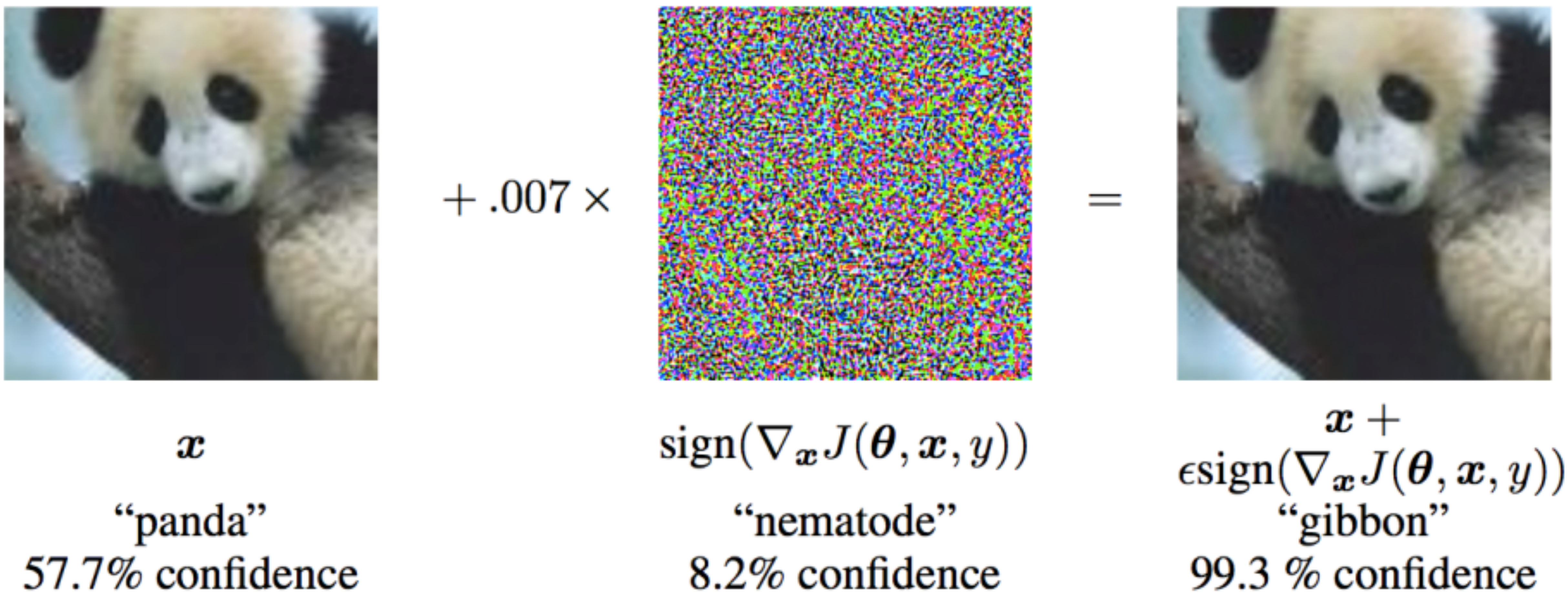


- In another experiment, you can start with a random noise and take **one** gradient step
- this often produces a confident classification
- the images outlined by yellow are classified as "airplane" with >50% confidence



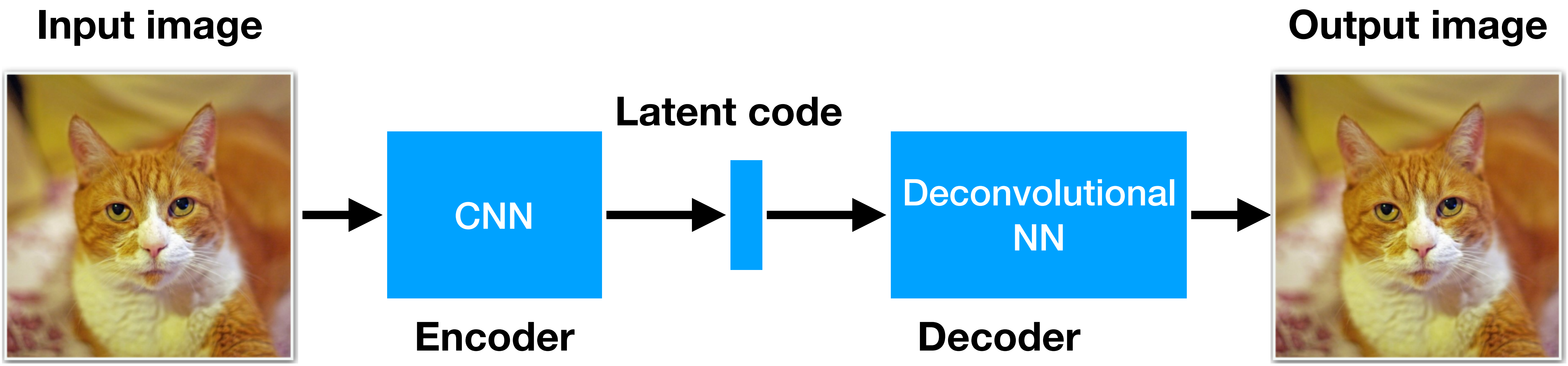
Attacking neural network with adversarial examples

- as an adversary, we want an image to be misclassified (to anything but Panda)



Attacking autoencoders

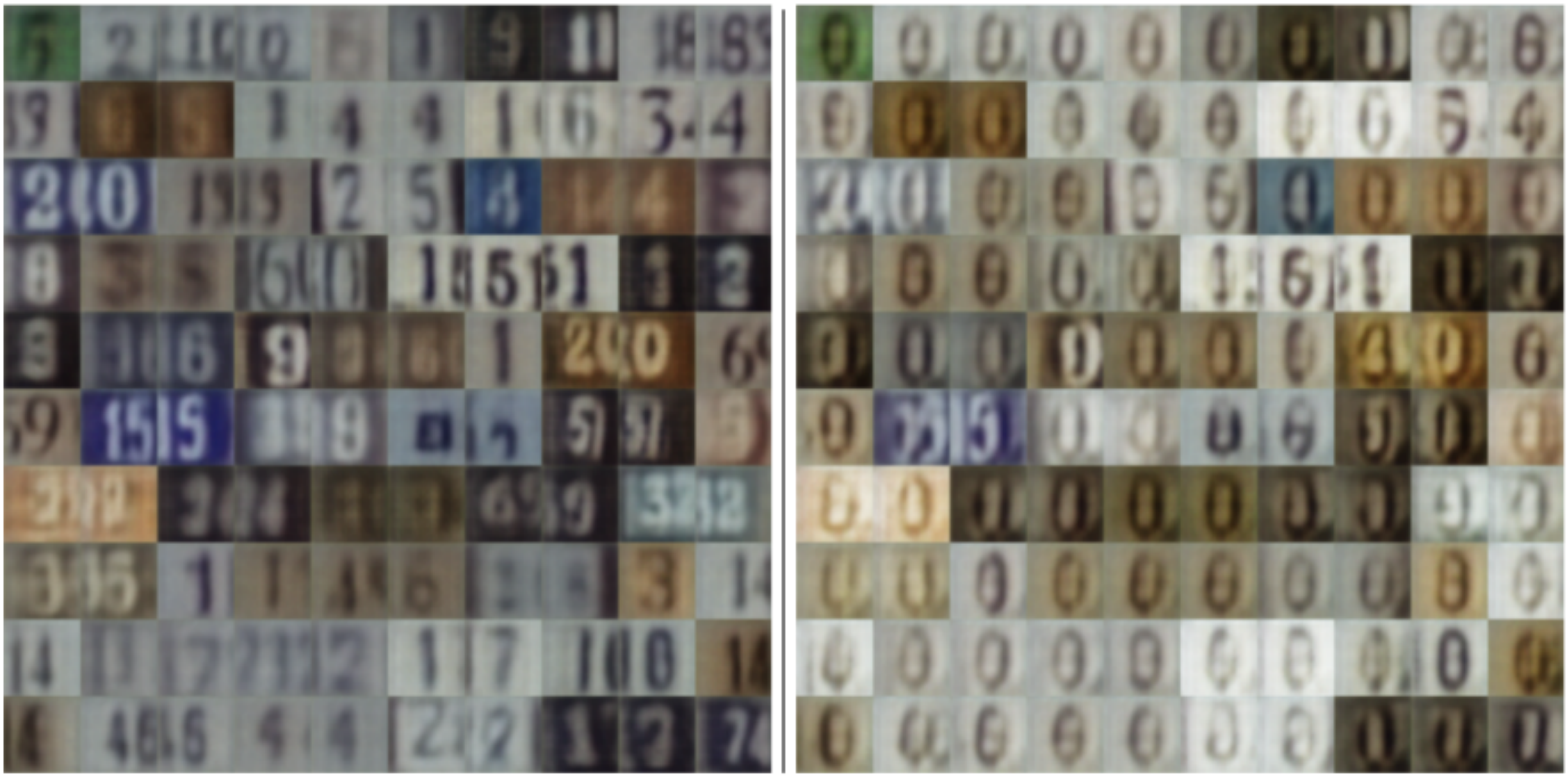
- Autoencoder: neural network that compresses the input, and recovers an example that is close to the input



- encoder and decoder are neural networks, jointly trained to minimize the squared loss between the input and output images

Adversarial examples

- one can create adversarial images that is reconstructed (after compression) as an entirely different image



Adversarial testing examples

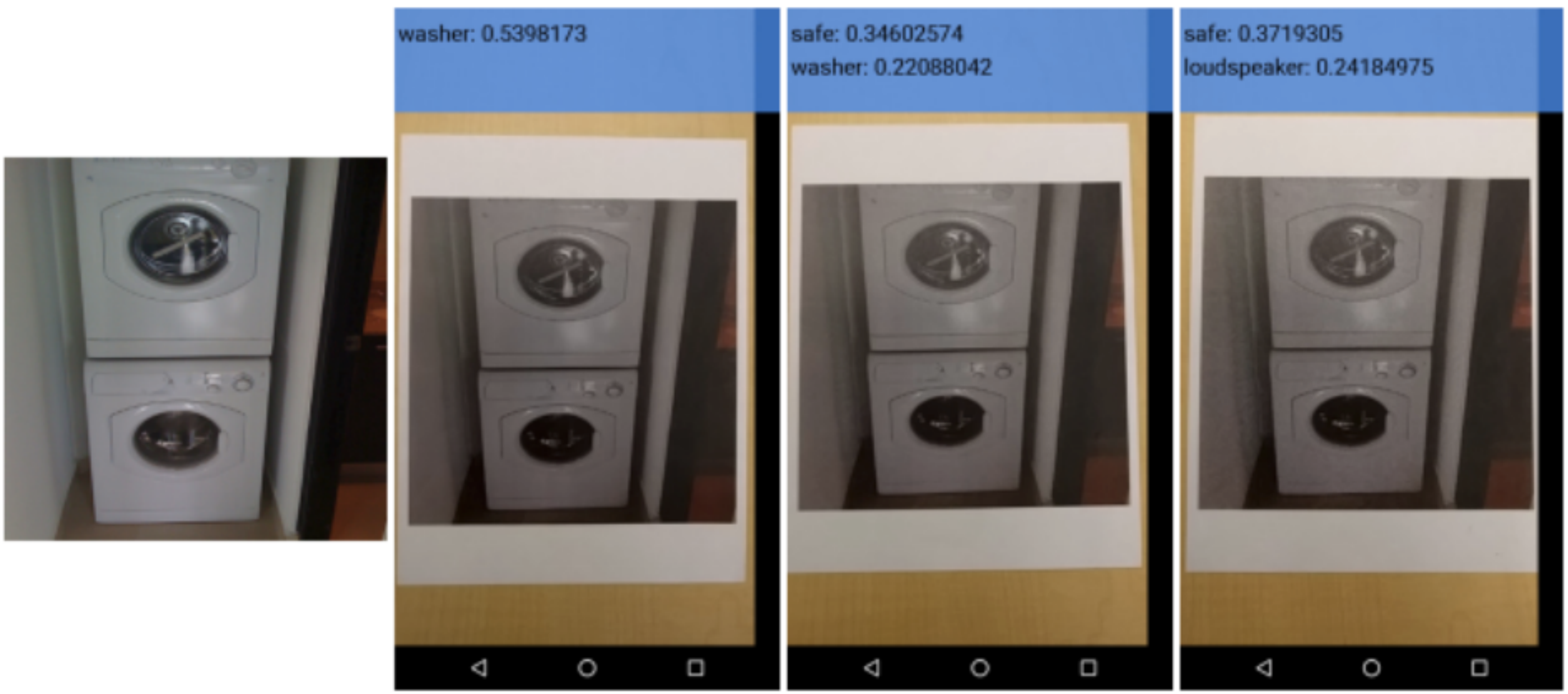
- First reported in ["Intriguing properties of neural networks", 2013, by Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus]
- Led to serious concerns for security as, for example,
  - one can create road signs that fools a self-driving car to act in a certain way
- this is serious as
  - there is no reliable defense against adversarial examples
  - adversarial examples transfer to different networks, trained on disjoint subset of training data
  - you do not need the access to the model parameters; you can train your own model and create adversarial examples
  - you only need a black-box access via APIs (MetaMind, Amazon, Google)

Adversarial examples with bloack-box access to NN

- ["Practical Black-Box Attacks against Machine Learning", 2016, Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami]
- no access to the gradient of the NN classifier, but only allowed black-box access to the output

Physical-world adversarial examples

- ["Adversarial examples in the physical world", 2016, Alexey Kurakin, Ian Goodfellow, Samy Bengio]
- You can fool a classifier by taking picture of a print-out.
- one can potentially print over a stop sign to fool a self-driving car



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

This 3-dimensional turtle is designed to be classified as “rifle”

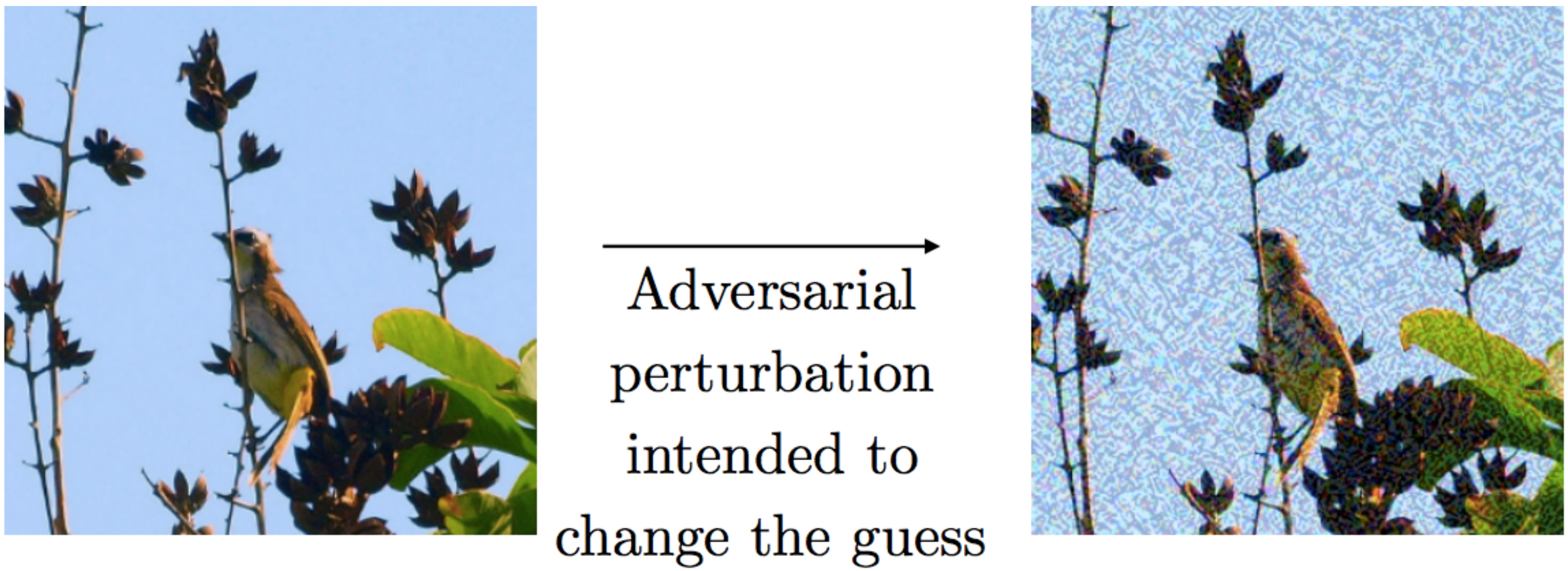
Classified as rifle

Classified as other

Defense mechanism

Defense mechanism 1

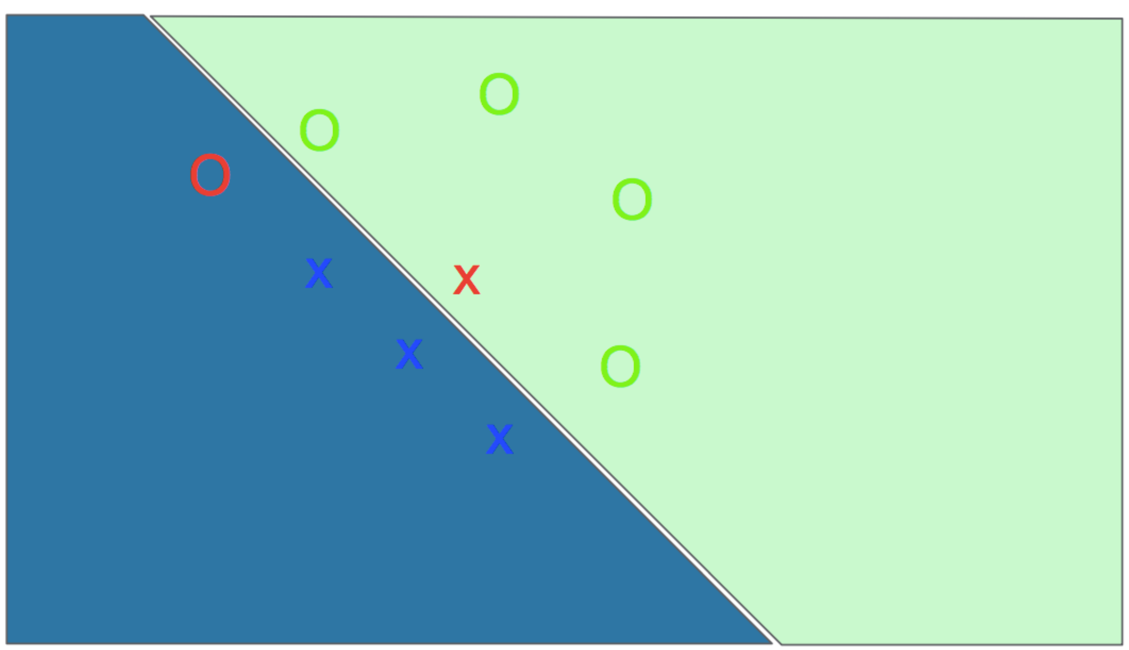
- include adversarial testing examples (but with the correct classes) in the training data.

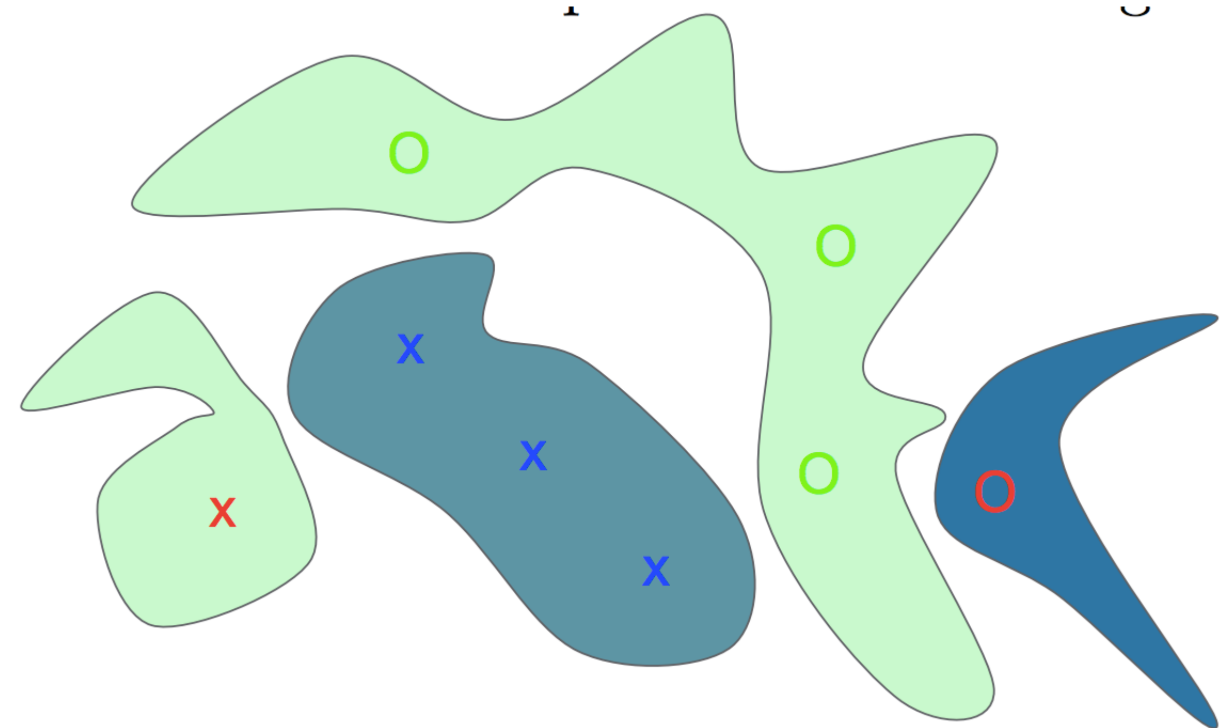


label: bird **label: bird**

Why are modern classifiers vulnerable

- **small margin** due to overfitting / high representation power
- there exists a direction from any example that can reach a boundary in a short distance





Defense mechanism 2

- **Defensive distillation:**
  - Two models are trained
  - model 1: trained on the training data in as standard manner
  - model 2 (the robust model) : is trained on the same training data, but uses **soft classes** which is the probability provided by the first model
  - This creates a model whose surface is smoothed in the directions an adversary will typically try to exploit, making it difficult for them to discover adversarial input tweaks that lead to incorrect categorization
  - [Distilling the Knowledge in a Neural Network, 2015, Geoffrey Hinton, Oriol Vinyals, Jeff Dean]
  - original idea came from model compression

- both are vulnerable against high-power adversary

Unsupervised Learning with Neural Networks

Deep generative model

- traditional parametric generative model
  - Gaussian:

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

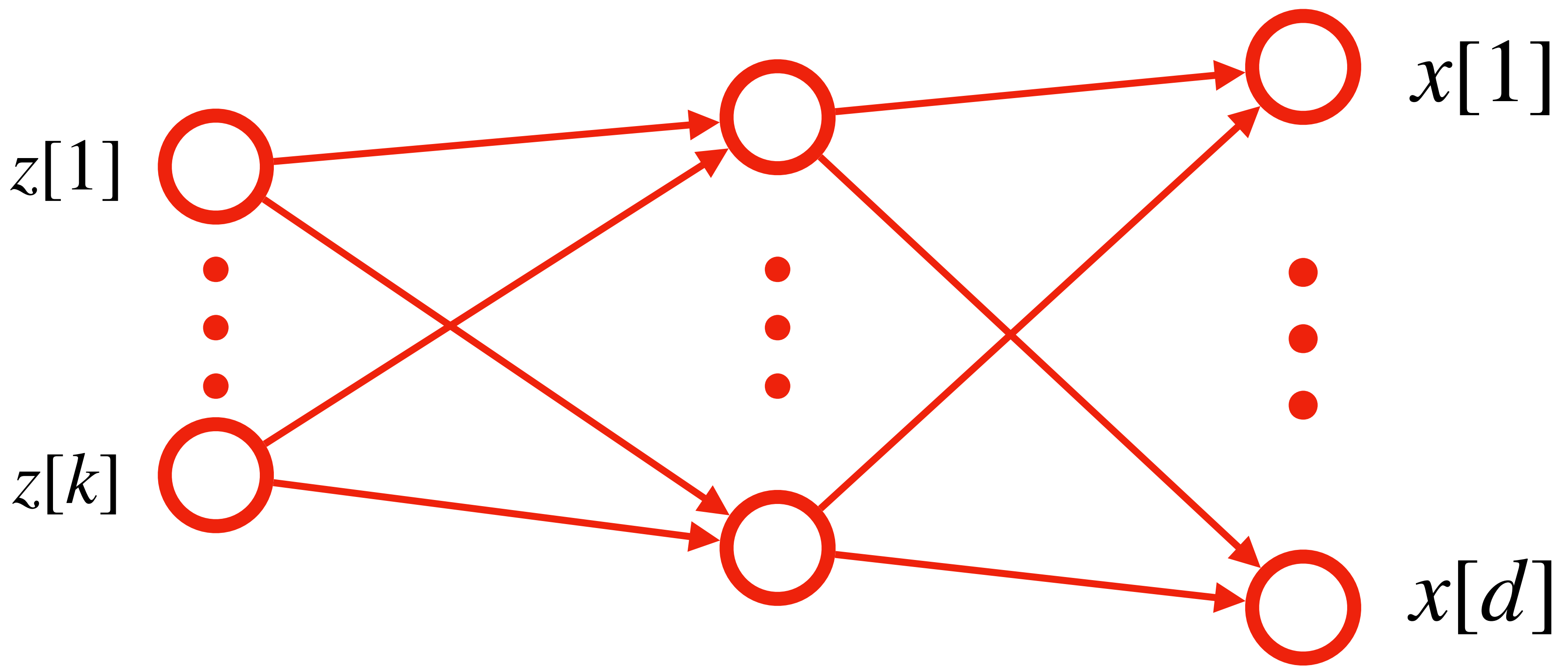
  - Gaussian Mixture Models (GMM)

$$f_{\{\mu_i\},\{\sigma_i\},\{\pi_i\}}(x) = \sum_{i=1}^{k} \pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

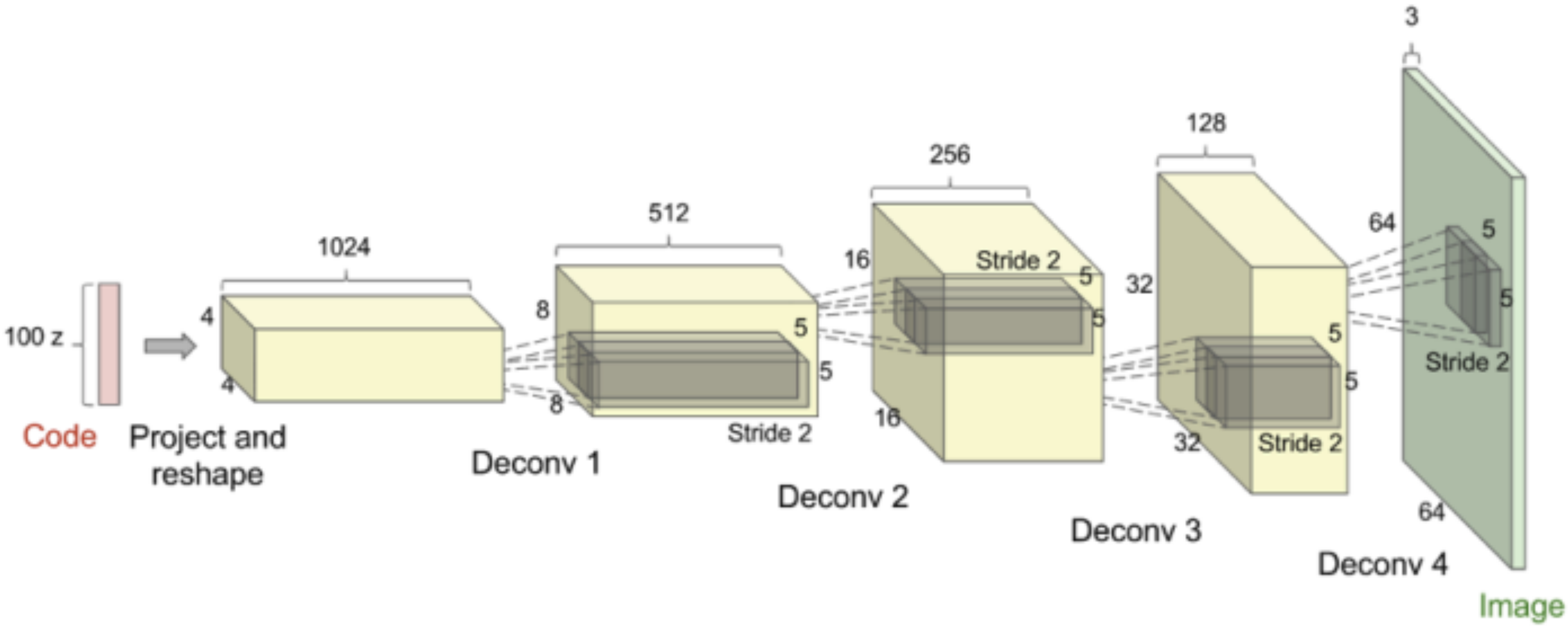
- deep generative model
  - easy to sample
  - high representation power
  - but no tractable evaluation of the density (i.e. p.d.f.)

Deep generative model

- sampling from a deep generative model, parametrized by w
  - first sample a **latent code** $z \in \mathbb{R}^k$ of small dimension $k \ll d$, from a simple distribution like standard Gaussian $N(0, \mathbf{I}_{k \times k})$
  - pass the code through a neural network of your choice, with parameter w
  - the output sample $x \in \mathbb{R}^d$ is the sample of this deep generative model



Deep generative model



Generative model

- a task of importance in unsupervised learning is fitting a generative model
- classically, if we fit a parametric model like mixture of Gaussians, we write the likelihood function explicitly in terms of the model parameters, and maximize it using some algorithms
- $$\text{maximize}_w \sum_{i=1}^{n} \log \left(\underbrace{P_w(x_i)}_{\text{p.d.f.}} \right)$$
- deep generative models use neural networks, but the likelihood of deep generative models cannot be evaluated easily, so we use alternative methods

Goal

- Given examples $\{x_i\}_{i=1}^n$ coming i.i.d from an unknown distribution $P(x)$, train a generative model that can generate samples from a distribution close to $P(x)$

These are computer generated images from the "bigGAN".

Adversarial training

- Classification
  - Consider the example of SPAM detection
  - Each sample x_i is an email
  - Distribution of **true email** is $P(x)$
  - Suppose spammers generate **spams** with distribution $Q(x)$
  - Spam detection: Typical classification task
    - Generate samples from true emails and label them $y_i = 1$
    - Generate samples from spams and label them $y_i = 0$
    - Using these as training data, train a classifier that outputs

$$\mathbb{P}(y_i = 1 \mid x_i) \simeq \frac{1}{1 + e^{-f_\theta(x)}}$$

for some neural network $f_\theta(\,\cdot\,)$ with parameter θ (this is the **logistic model** for binary classification)

Adversarial training

- Applying logistic regression, we want to solve

$$\max_{\theta} \sum_{i:y_i=1} \log\left(\frac{1}{1+e^{-f_\theta(x_i)}}\right) + \sum_{i:y_i=0} \log\left(1 - \frac{1}{1+e^{-f_\theta(x_i)}}\right)$$

- in **adversarial training**, it is customary to write

$$\boxed{D_\theta(x)} = \frac{1}{1+e^{-f_\theta(x)}}$$

  which is called a **discriminator**

- and find the "best" discriminator by solving for

$$\max_{\theta} \mathscr{L}(\theta) = \sum_{x_i \sim P(\cdot)} \log D_\theta(x_i) + \sum_{x_i \sim Q(\cdot)} \log(1 - D_\theta(x_i))$$

  as 1 labelled examples come from real distribution $P(\,\cdot\,)$
  and 0 labelled examples come from spam distribution $Q(\,\cdot\,)$

Adversarial training

- Suppose now that the **spam detector (i.e. the discriminator)** is fixed, then the spammer's job is to generate spams that can fool the detector by making the likelihood of the spams being classified as spams **small**:

$$\min_{Q(\cdot)} \mathscr{L}(\theta) \quad = \quad \underbrace{\sum_{x_i \sim P(\cdot)} \log D_\theta(x_i)}_{\text{does not depend on } Q(\cdot)} \quad + \sum_{x_i \sim Q(\cdot)} \log(1 - D_\theta(x_i))$$

- where 0 labelled examples are coming from the distribution $Q(\cdot)$, which is modeled by a **deep neural network generative model,** i.e. $x_i = G_w(z_i)$ where $z_i \sim N(0, \mathbf{I}_{k \times k})$.
- The minimization can be solved by finding. The "best" generative model that can fool the discriminator

$$\min_{w} \mathscr{L}(w, \theta) \quad = \quad \underbrace{\sum_{x_i \sim P(\cdot)} \log D_\theta(x_i)}_{\text{does not depend on } Q(\cdot)} \quad + \boxed{\sum_{\substack{\cancel{x_i \sim Q(\cdot)} \\ z_i \sim N(0, \mathbb{I}_{k \times k})}} \log\Big(1 - D_\theta\big(G_w(z_i) \big) \Big)}$$

Adversarial training

- Now we have a game between the spammer and the spam detector:

$$\min_{w} \max_{\theta} \sum_{x_i \sim P(\cdot)} \log D_\theta(x_i) + \sum_{z_i \sim N(0,\mathbf{I})} \log(1 - D_\theta(G_W(z_i)))$$

- Where $P(\cdot)$ is the distribution of real data (true emails), and $Q(\cdot)$ is the distribution of the generated data (spams) that we want to train with a **deep generative model**
- jointly training the discriminator and the generator is called **adversarial training**
- Alternating method is used to find the solution

Alternating gradient descent for adversarial training

- Gradient update for the **discriminator** (for fixed w)

$$\max_{\theta} \sum_{x_i \sim P(\cdot)} \log D_\theta(x_i) + \sum_{x_i \sim Q(\cdot)} \log(1 - D_\theta(x_i))$$

- First sample n examples from real data (in the training set) and the generator data $x_i \sim G_w(z_i)$
  (for the current iterate of the generator weight w)
- compute the gradient for those $2n$ samples using back-propagation
- Update the discriminator weight θ by subtracting the gradient with a choice of a step size

Alternating gradient descent for adversarial training

- gradient update for the generator (for fixed θ)

$$\min_w \sum_{x_i \sim P(\cdot)} \log D_\theta(x_i) + \sum_{z_i \sim N(0,\mathbf{I})} \log(1 - D_\theta(G_w(z_i)))$$

- Consider the gradient update on a single sample

$$\min_w \mathscr{L}(w, z_i) \;=\; \log(1 - D_\theta(G_w(z_i)))$$

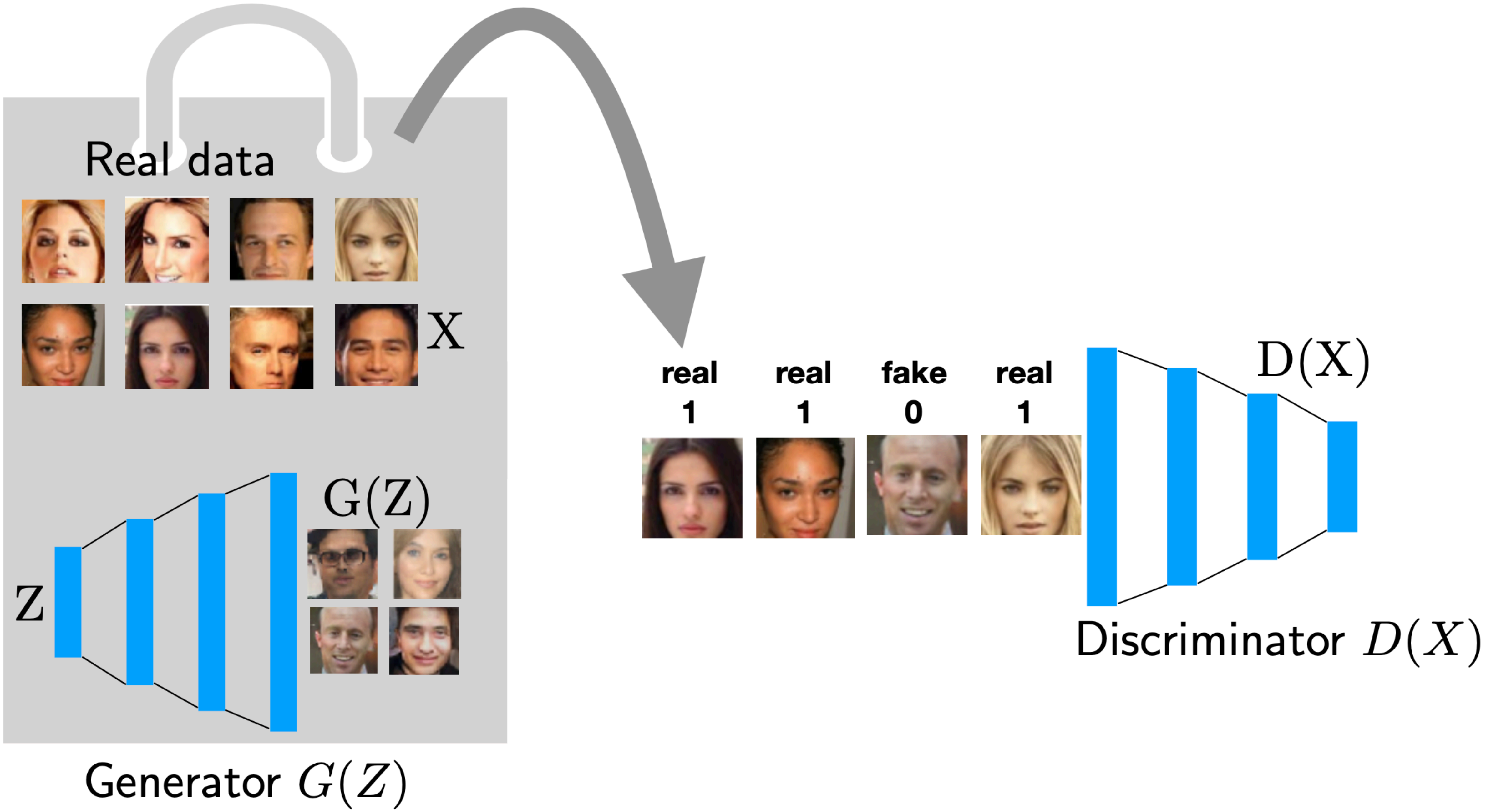
for a single $z_i \sim N(0,\mathbf{I})$ sampled from a Gaussian

- The gradient update is

$$\begin{aligned} w \;&=\; w - \eta \nabla_w \mathscr{L}(w, z_i) \\ &=\; w - \eta \; \nabla_w G_w(z_i) \; \nabla_x D_\theta(x) \; \frac{-1}{1 - D_\theta(x)} \end{aligned}$$

with $x = G_w(z_i)$

This gives a new way to train a deep generative model



$$\min_G \max_D V(G, D)$$

Not only is GAN amazing in generating realistic samples

http://whichfaceisreal.com

It opens new doors to exciting applications

- Cycle-GAN

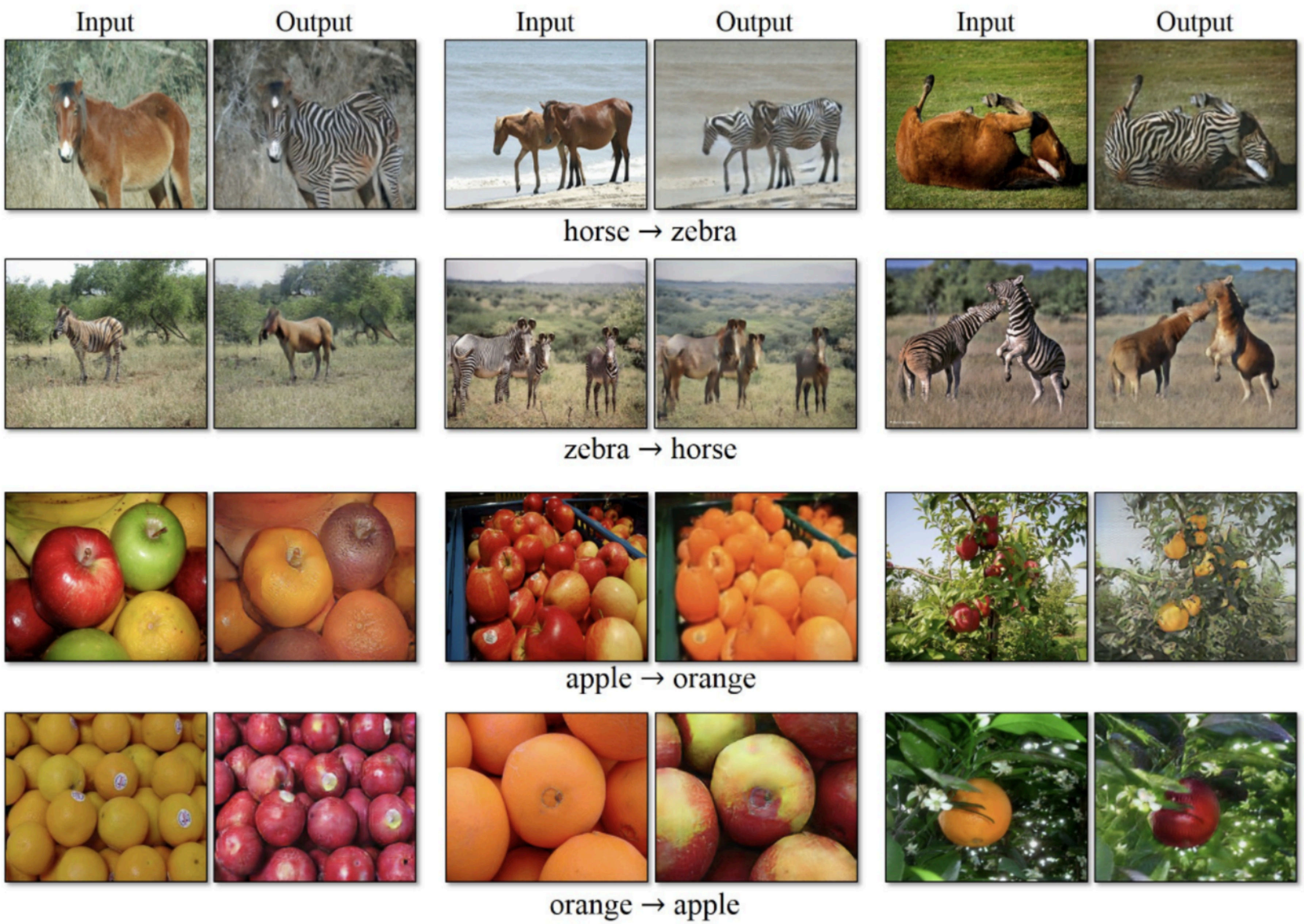


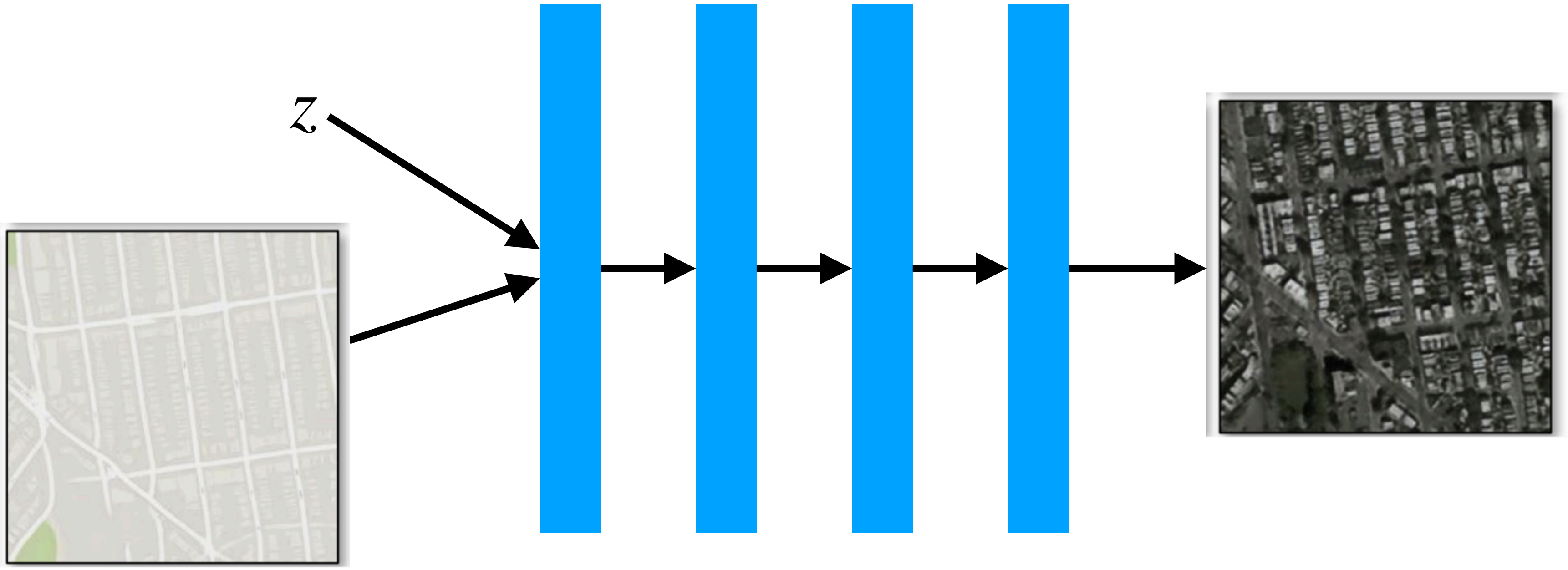
Figure 3: Street scene image translation results. For each pair, left is input and right is the translated image.

Style transfer with generative model

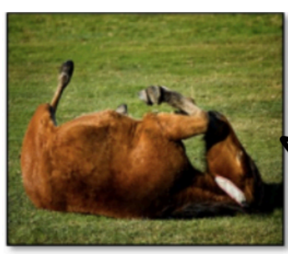
- If we have paired training data,

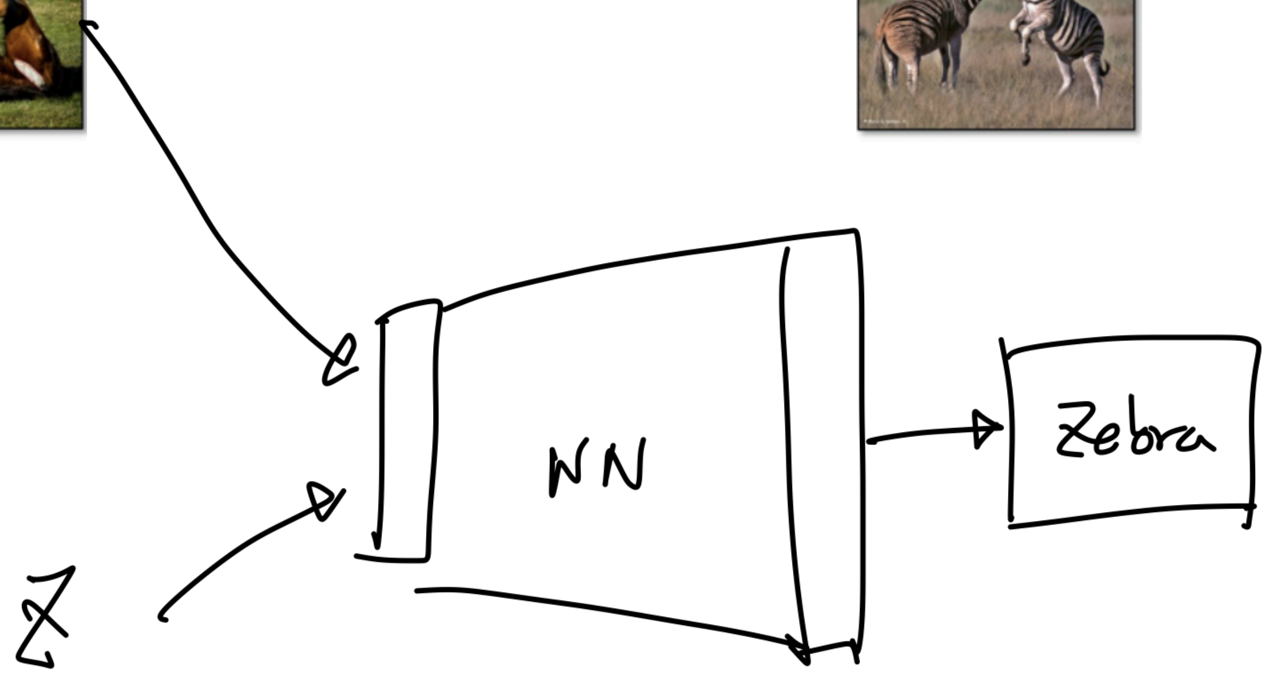
- And want to train a generative model G(x,z)=y,

- This can be posed as a regression problem

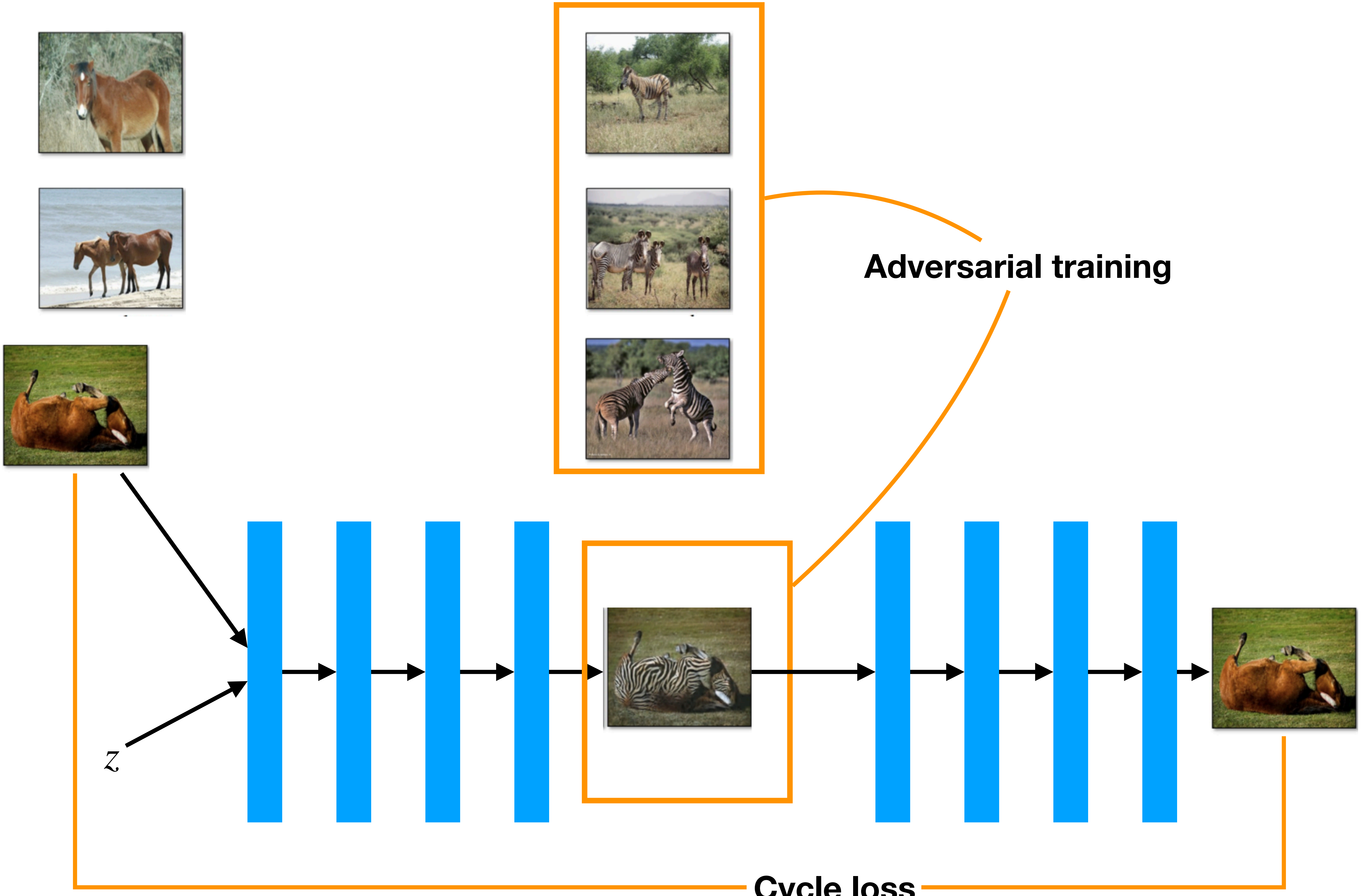


How do we do style transfer without paired data? Cycle-GAN

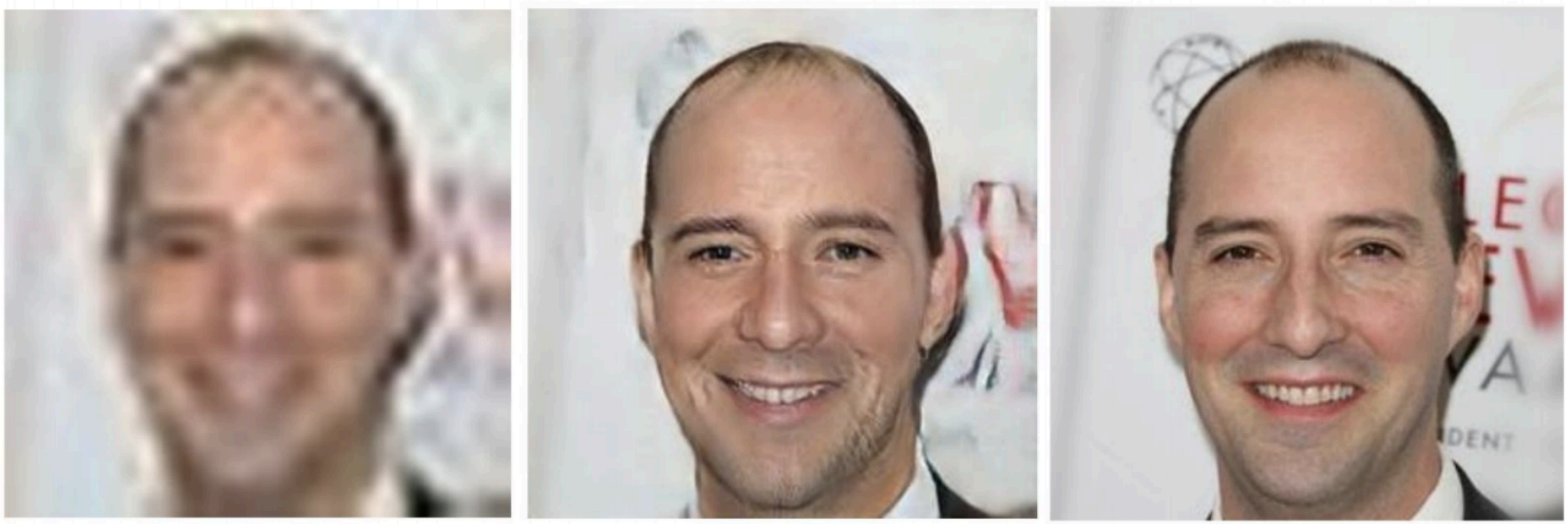




How do we do style transfer without paired data? Cycle-GAN

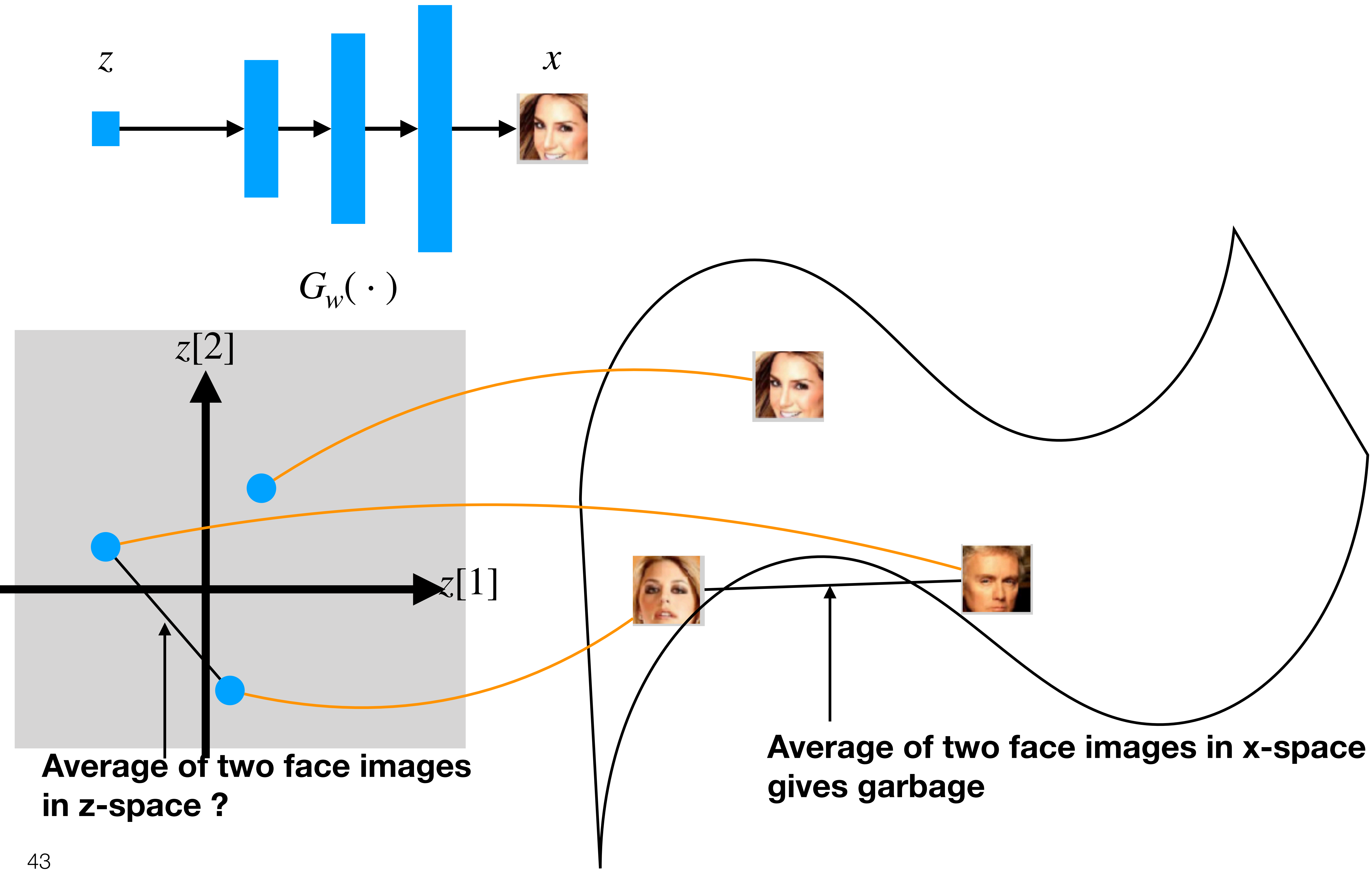


Super resolution



https://www.youtube.com/watch?v=PCBTZh41Ris

The learned latent space is important

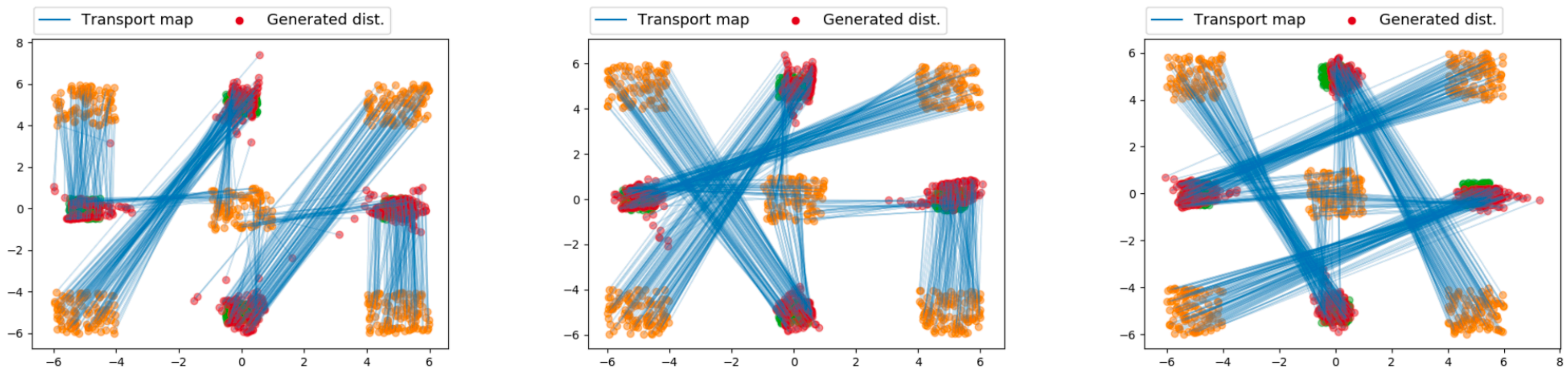


How do we check if we found the right manifold (of faces)?

- **latent traversal**

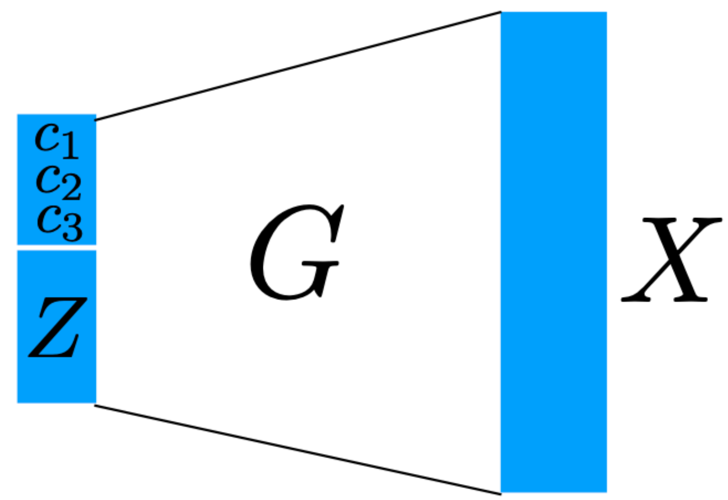
Can we make the relation between the latent space and the image space more meaningful?

- **Disentangling**
  - **GANs learn arbitrary mapping from z to x**
  - **As the loss only depends on the marginal distribution of x and not the conditional distribution of x given z (how z is mapped to x)**

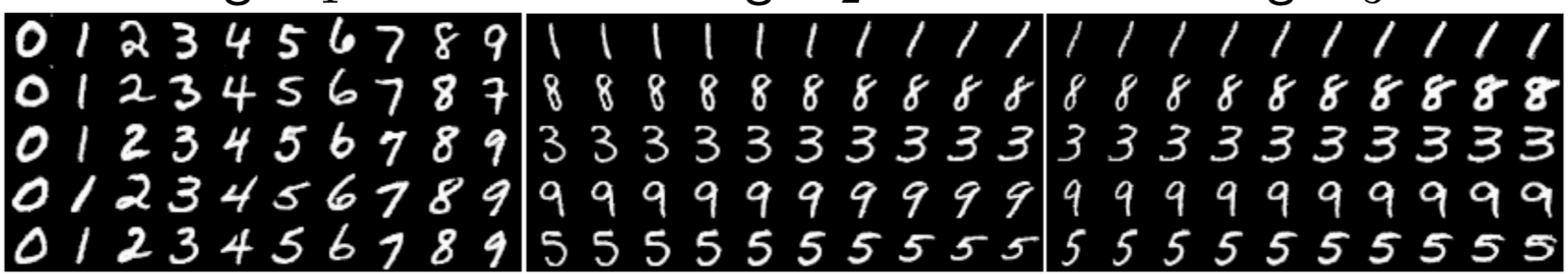


Disentangling seeks meaningful mapping from z to x

- there is no formal (mathematical) universally agreed upon definition of disentangling

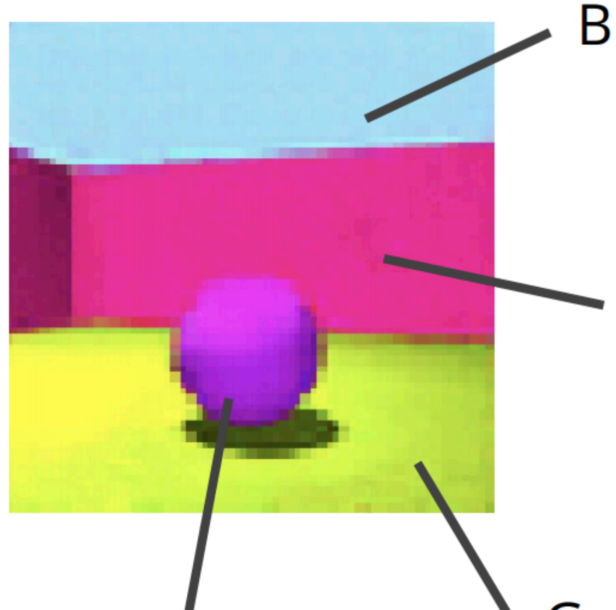


$\Rightarrow$ change c_1 $\Rightarrow$ change c_2 $\Rightarrow$ change c_3



- informally, we seek latent codes that
  - are "informative" or make "noticeable" changes
  - are "uncorrelated" or make "distinct" changes

Decompose data into a set of underlying **human-interpretable** factors of variation



Explainable models

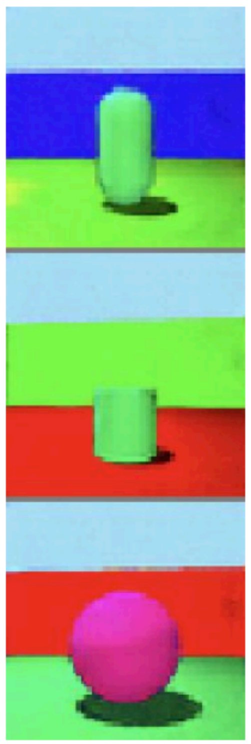
What is in the scene?

Controllable generation

Generate a red ball instead

Fully-supervised case

Strategy: Label everything



c_1 c_2 c_3

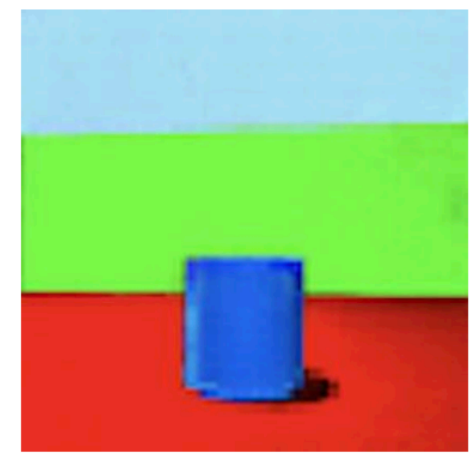
{dark blue wall, green floor, green oval}

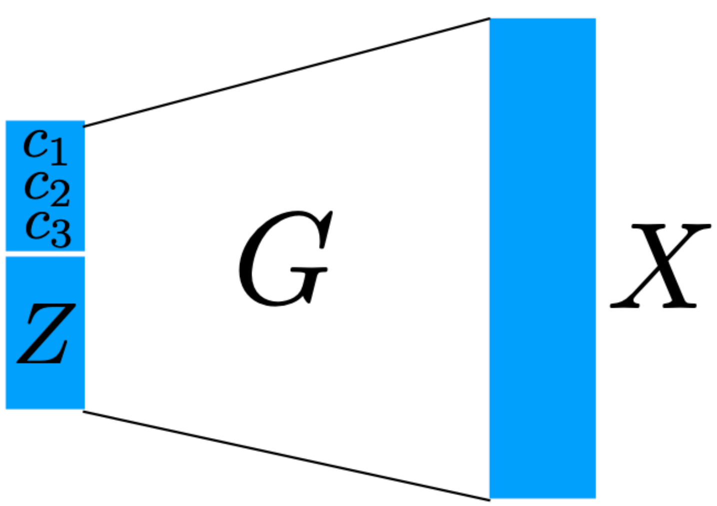
{green wall, red floor, green cylinder}

{red wall, green floor, pink ball}

Controllable generation as **label-conditional generative modeling**

green wall, red floor, blue cylinder





Train a **conditional GAN**, where (c_1, c_2, c_3) is a numerical representation of the **labels** given in the training data, and z is drawn from Gaussian

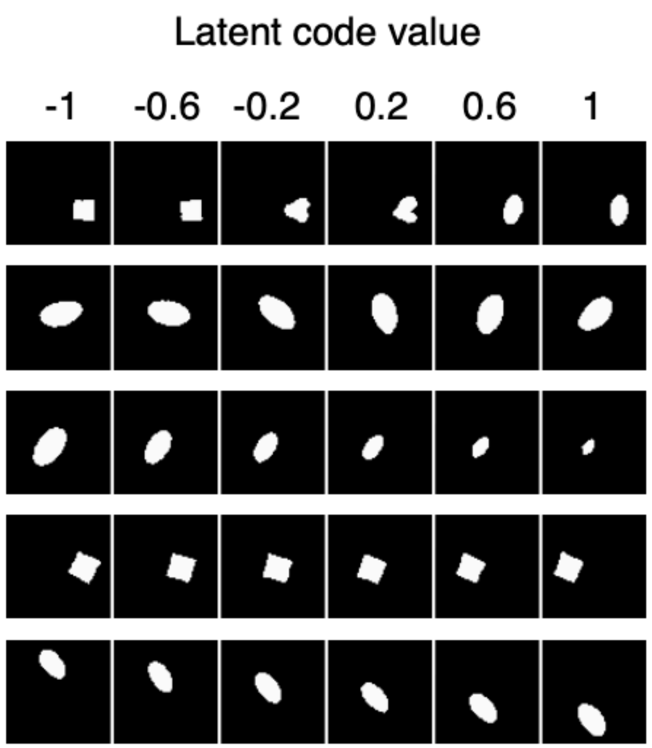
However, some properties are hard to represent numerically

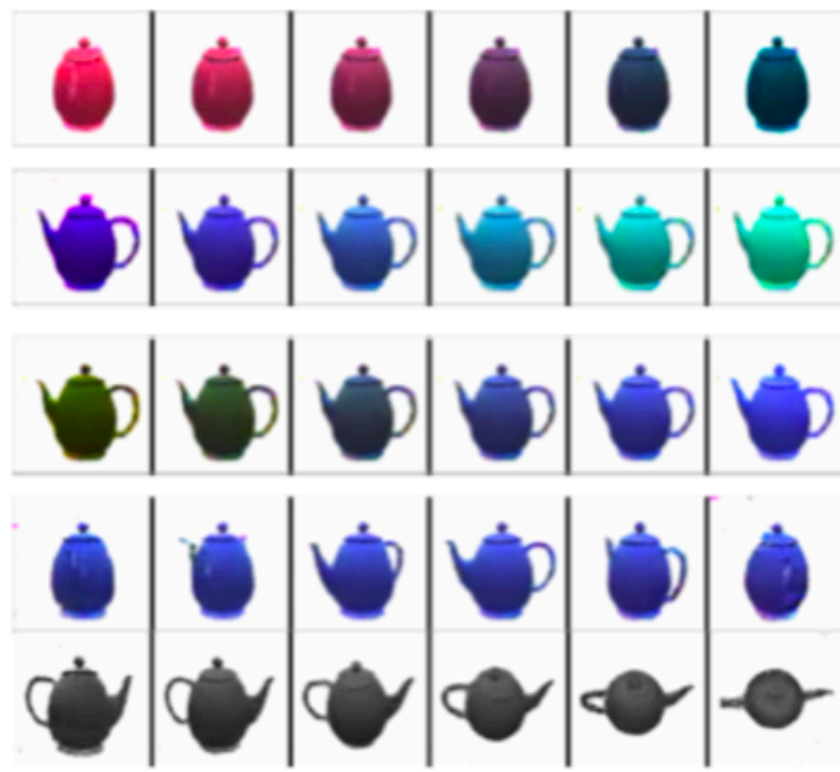
What kind of hairstyle?

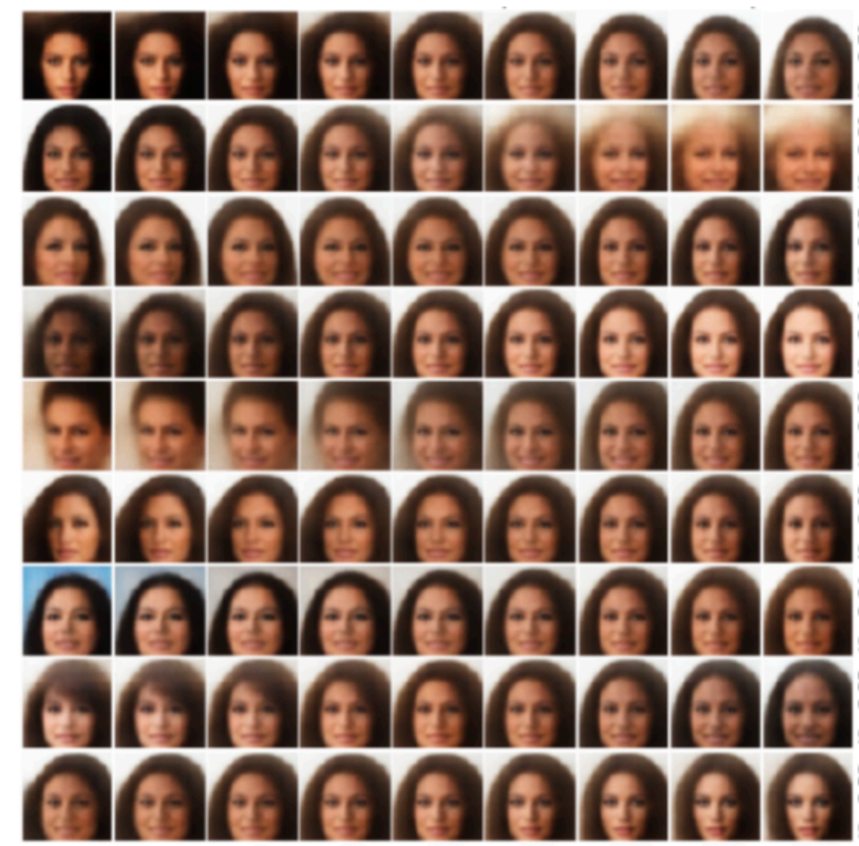
What kind of glasses?

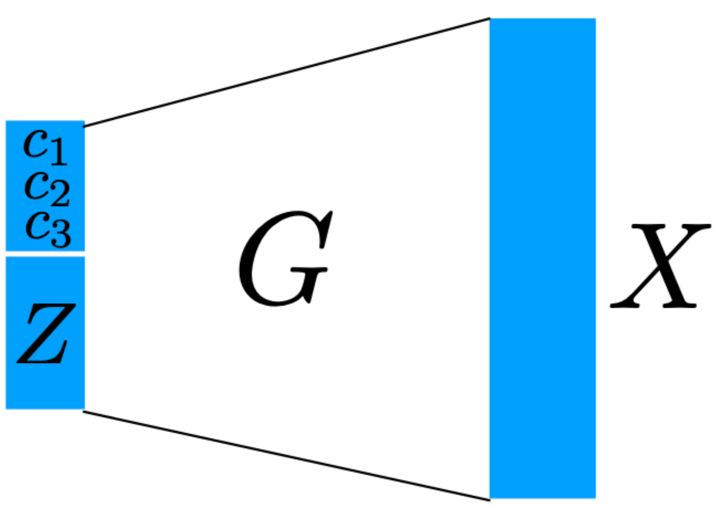
Generate this guy with this hair

Unsupervised training of Disentangled GAN



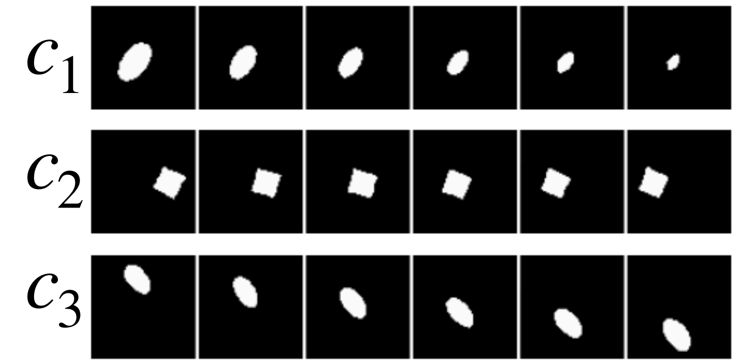






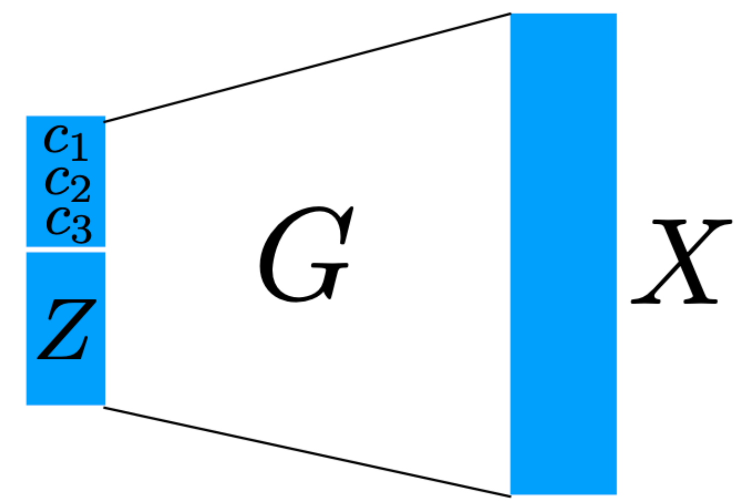
Disentangled GAN training: InfoGAN-CR, 2019

- 1. As in standard GAN training, we want $G_w(z)$ to look like training data (which is achieved by adversarial loss provided by a discriminator)



D( **) = {real,fake}**

- 2. We also want the controllable latent code c to be predictable from the image
  - add a NN regressor that predicts $\hat{c}(x)$, and train the generator that makes the prediction accuracy high (note that both this predictor and the generator works to make the prediction accurate, unlike adversarial loss)



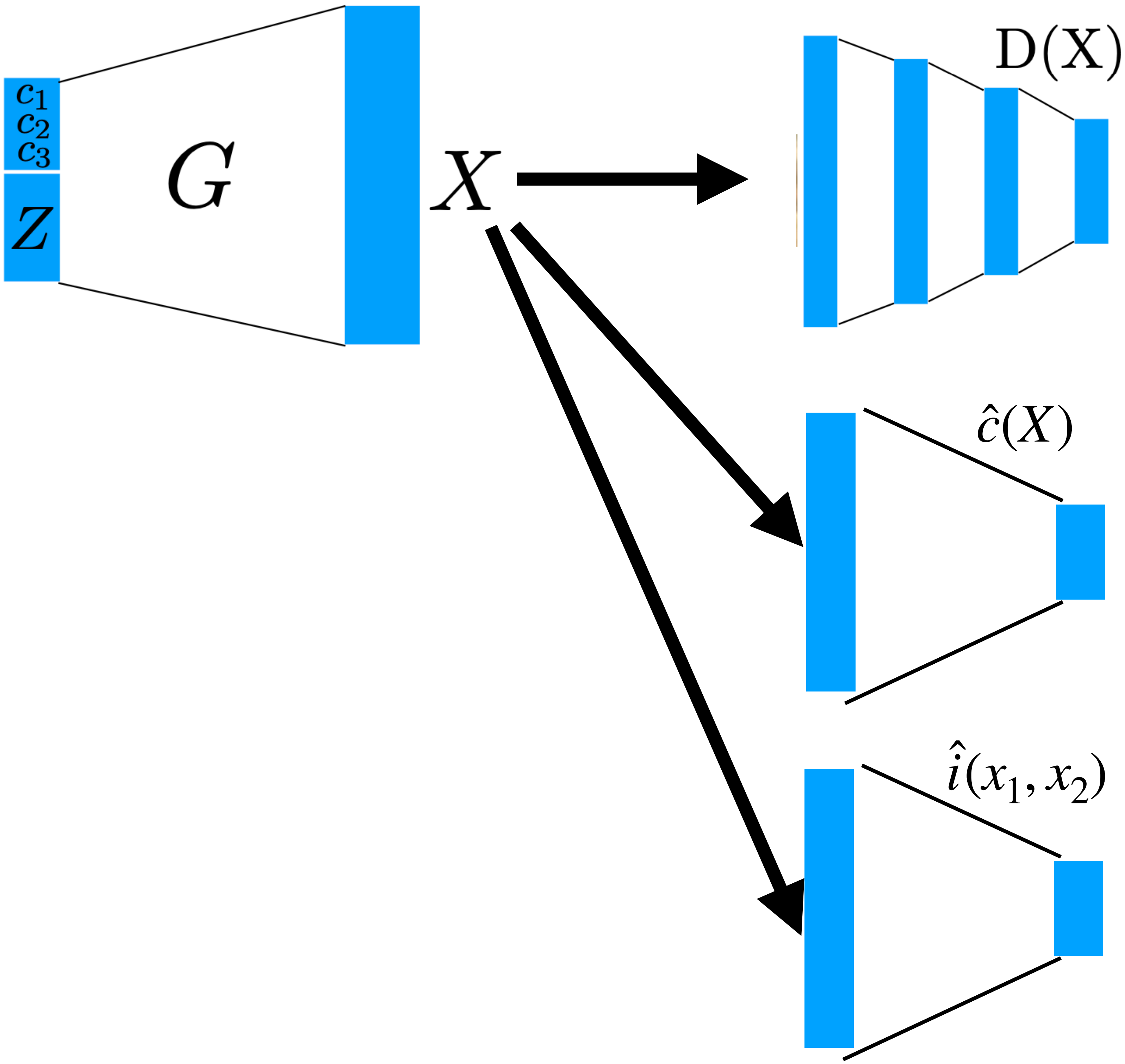
$\text{minimize} \quad \| \hat{c}($  $) - c\|^2$

- 3. We also want each code to control distinct properties
  - add a NN that predicts which code was changed

$\hat{i}($  $) \simeq i$

Disentangling with contrastive regularizer

- To train a disentangled GAN, we use contrastive regularizer



Discriminator encourages output x to be realistic

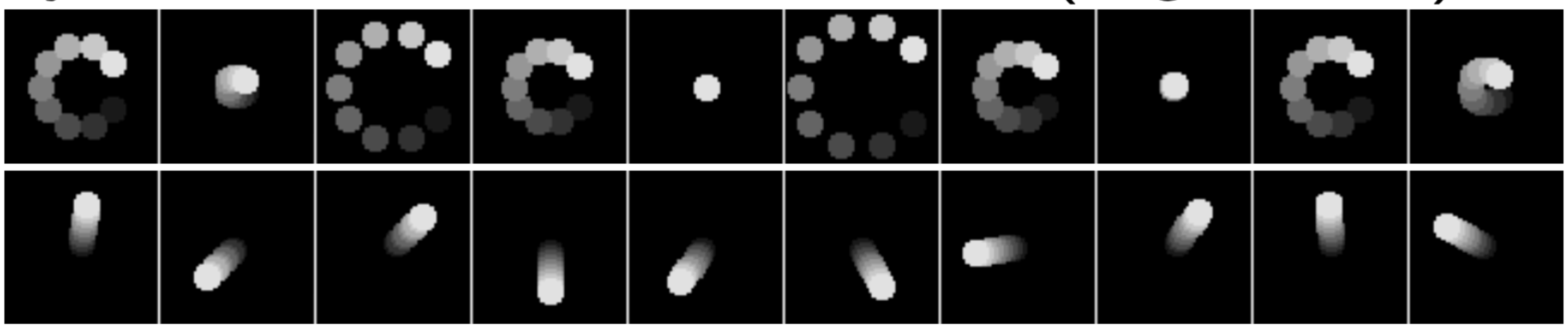
Predictor makes sure that the changes in c make noticeable changes in x

Contrastive regularizer detects which latent code c_i was the same in a paired (x_1, x_2)

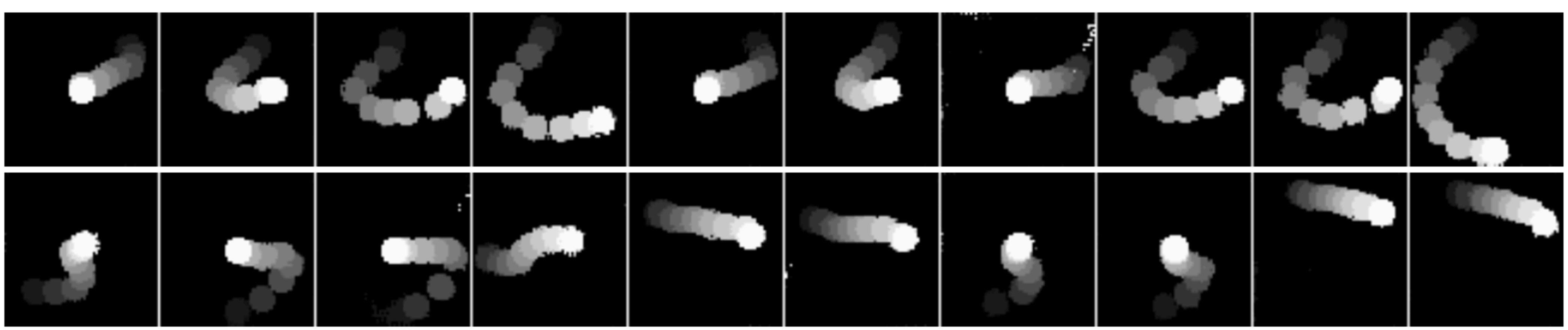
But is still challenging

- Synthetic training data (with planted disentangled representation)

Synthetic data with two attributes (angle,radius)

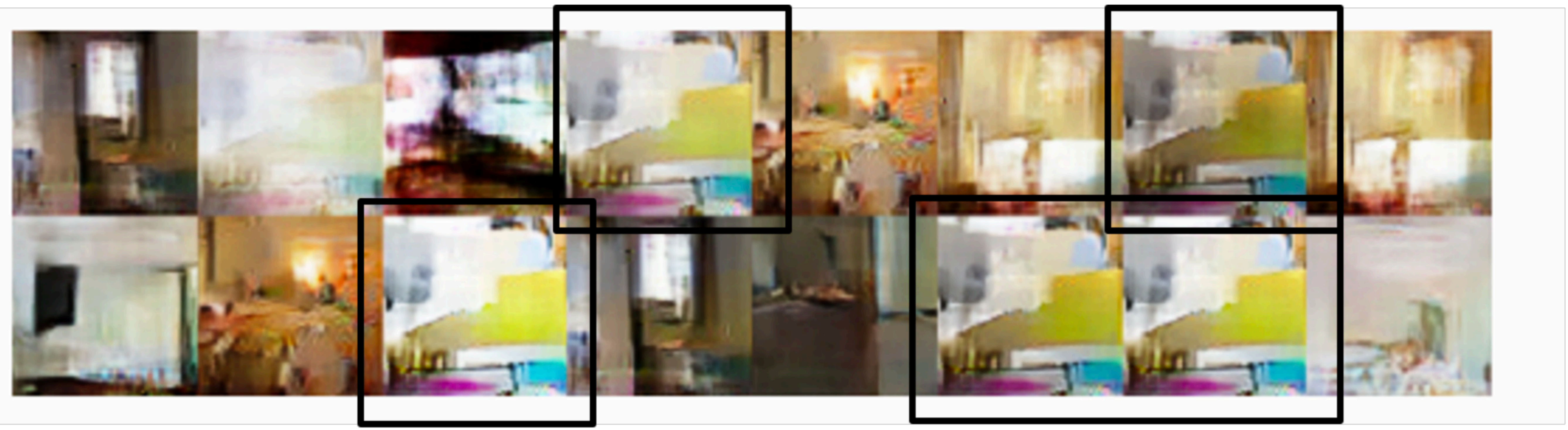


- Trained Disentangled GAN (latent traversal)



Challenges in training GANs

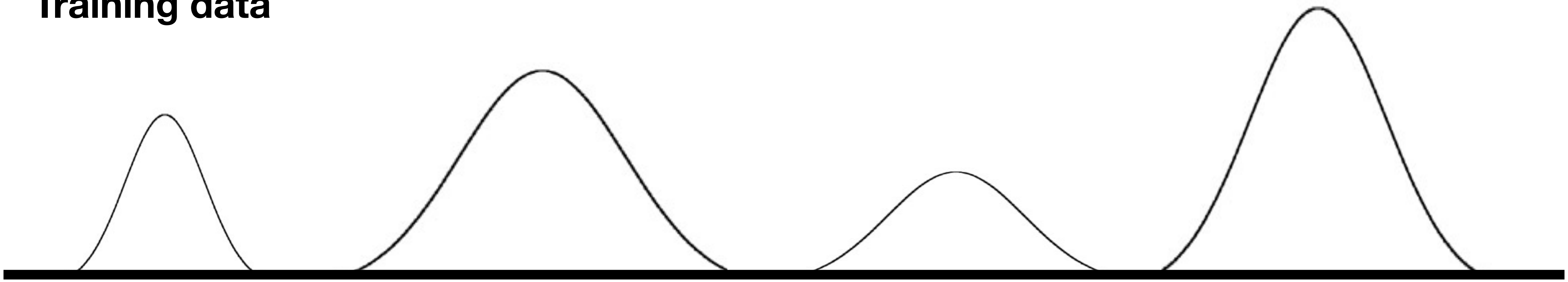
- GAN training suffers from **mode collapse**
- this refers to the phenomenon where the generated samples are not as diverse as the training samples



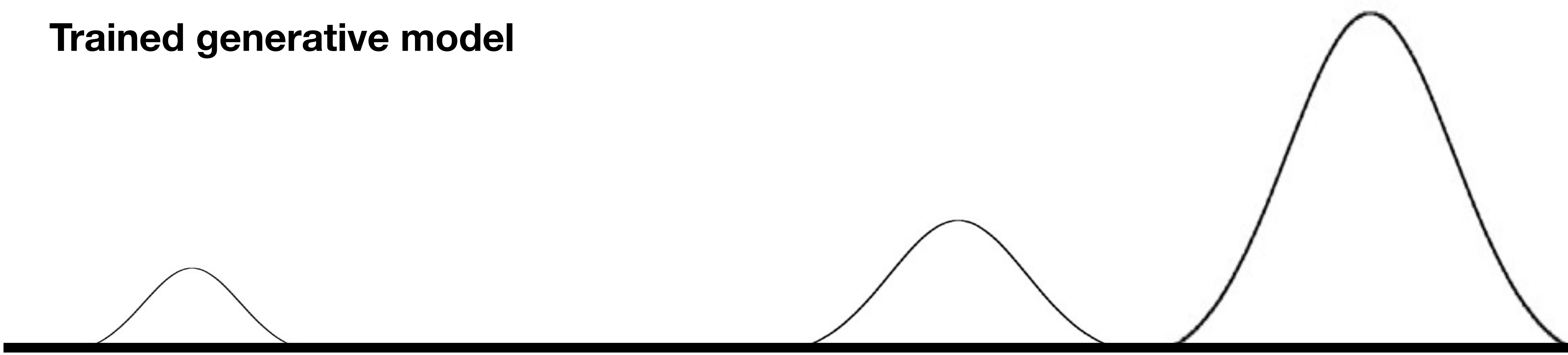
Arjovsky et al., 2017

Mode collapse

Training data

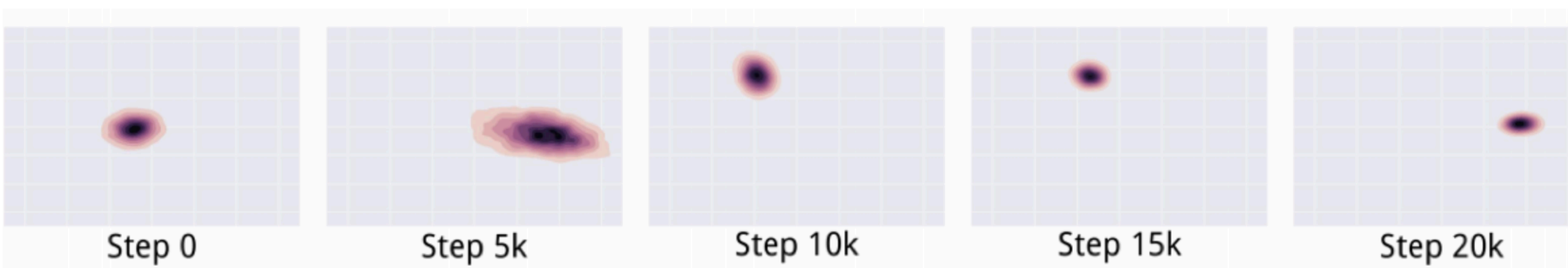


Trained generative model



Mode collapse

- True distribution is a mixture of Gaussians



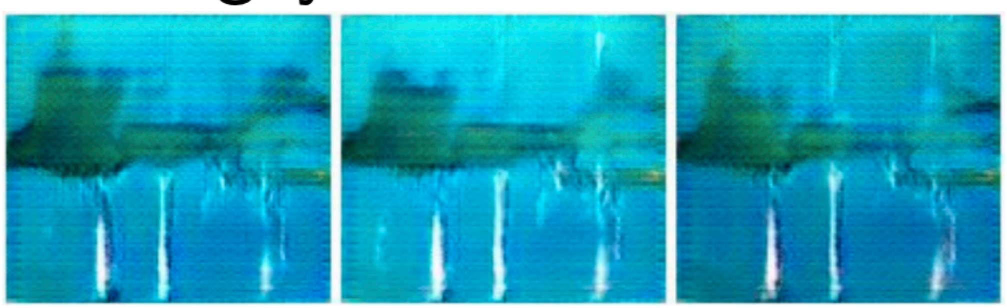
Source: Metz et al., 2017

- The generator distribution keeps oscillating between different modes

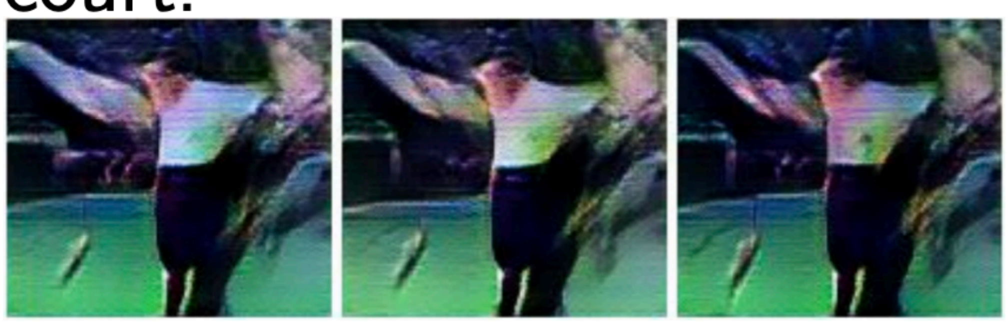
Mode collapse

- "A man in a orange jacket with sunglasses and a hat ski down a hill."

- "This guy is in black trunks and swimming underwater."



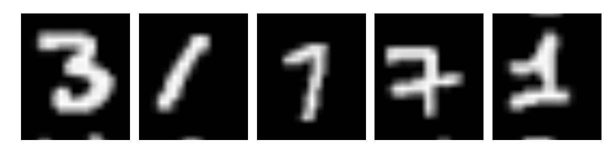
- "A tennis player in a blue polo shirt is looking down at the green court."



["Generating interpretable images with controllable structure", by Reed et al., 2016]

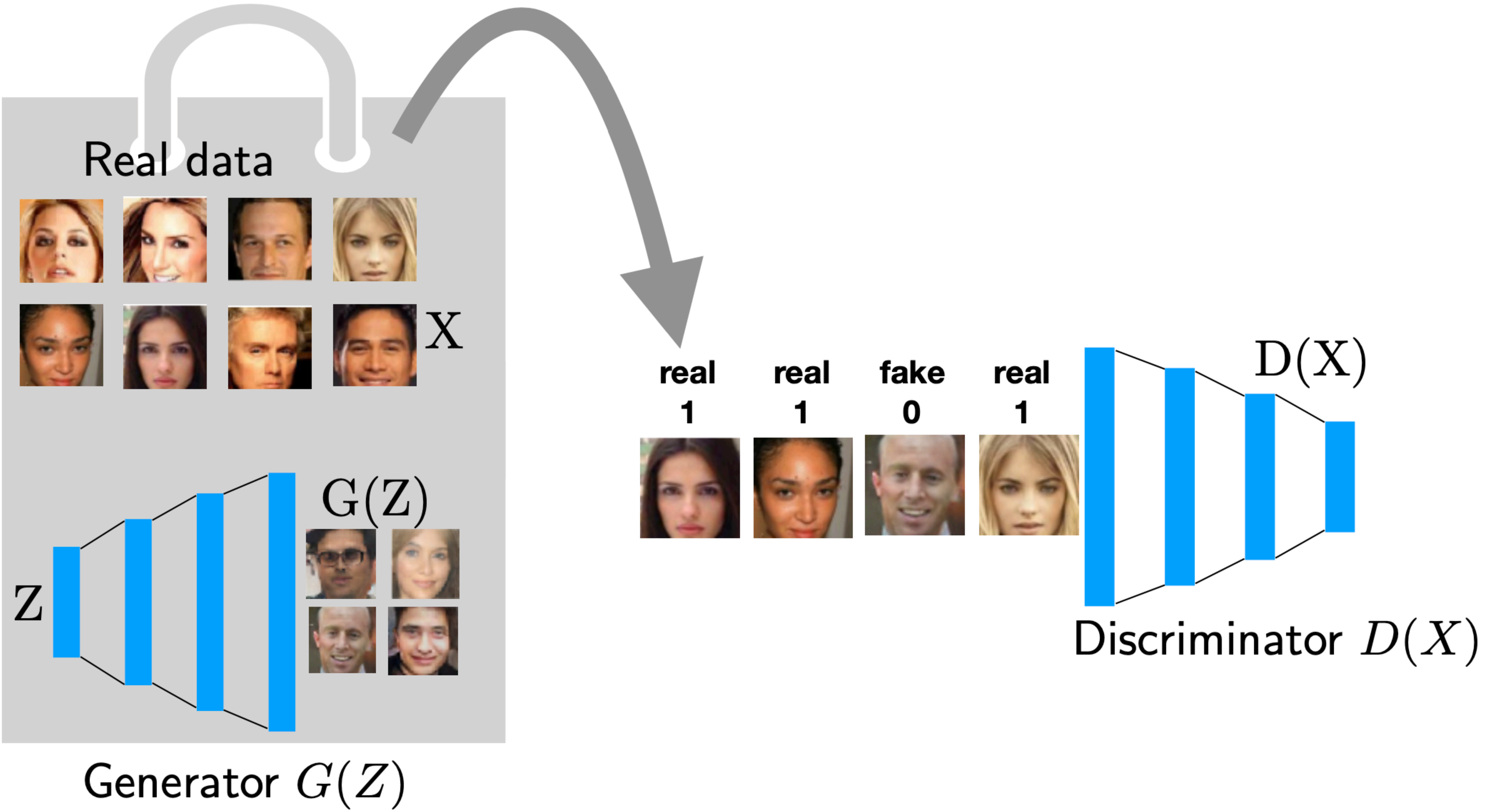
Principled approach to mode collapse

- Lack of diversity is easier to detect if we see multiple samples
- Consider MNIST hand-written digits
  - If we have a generator that generates 1,3,5,7 perfectly, it is hard to tell from a single sample that mode collapse has happened
  - But easier to tell from a collection of, say, 5 samples all from wither training data or all from generated data



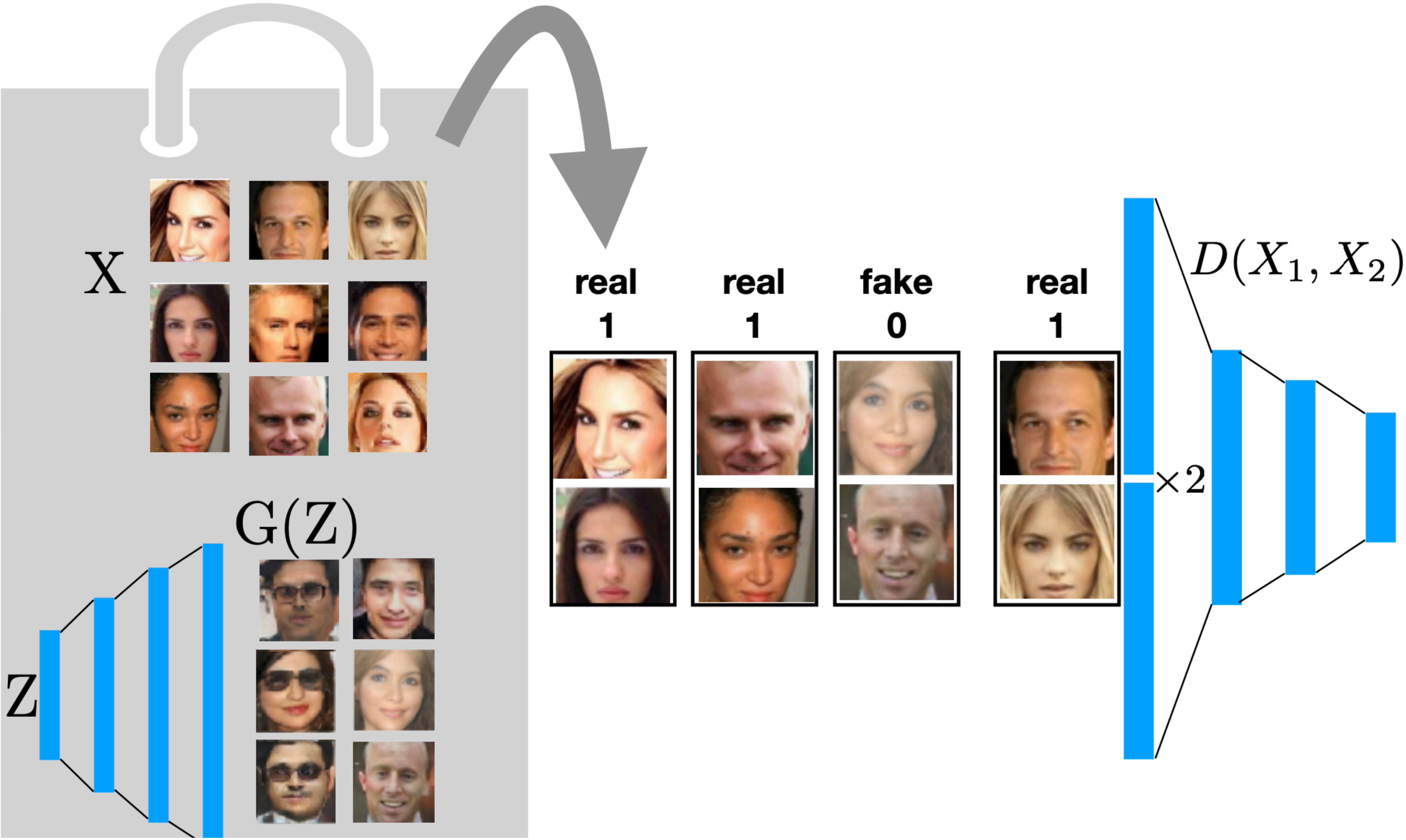
Principled approach to mode collapse

- Turning this intuition into a training algorithm:

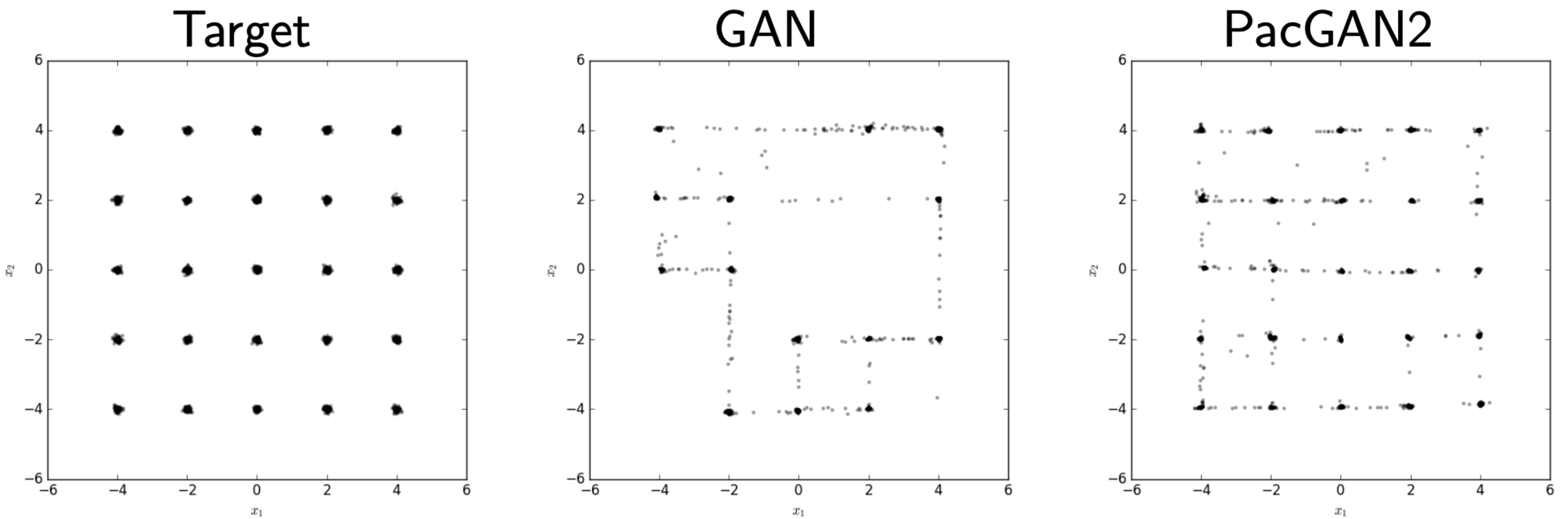


Principled approach to mode collapse: PacGAN, 2018

- Turning this intuition into a training algorithm:

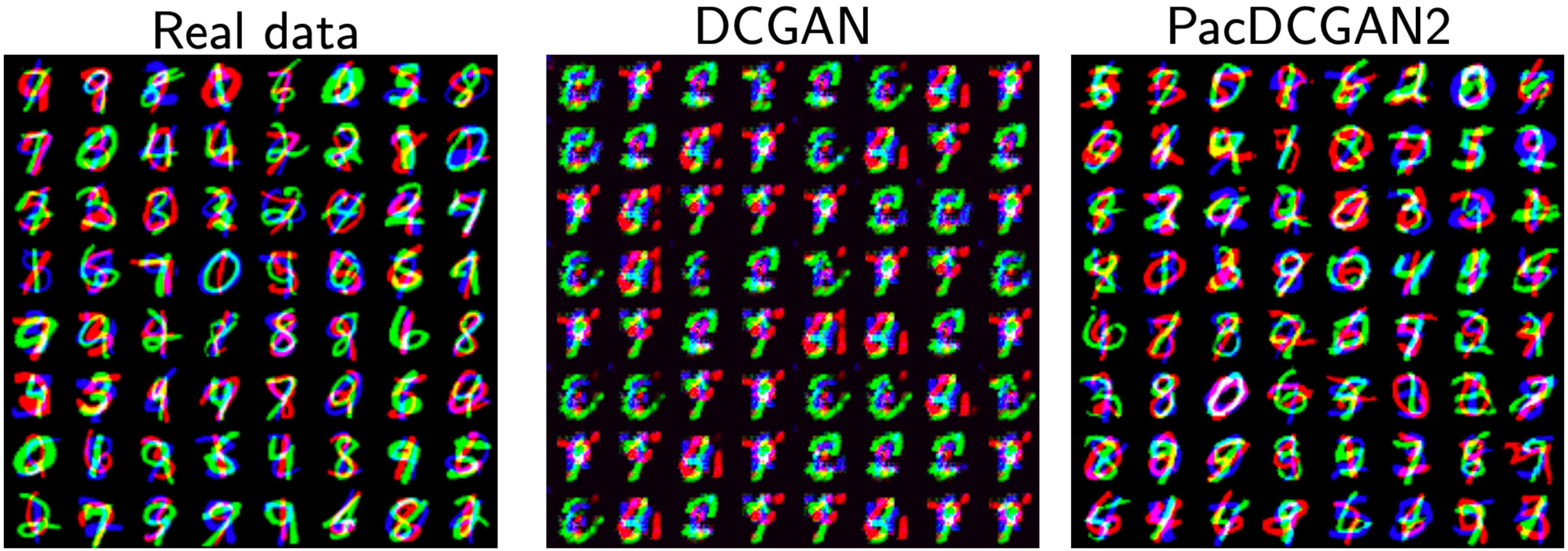


Principled approach to mode collapse



	Modes (Max 25)
GAN	17.3
PacGAN2	23.8
PacGAN3	24.6
PacGAN4	24.8

Principled approach to mode collapse

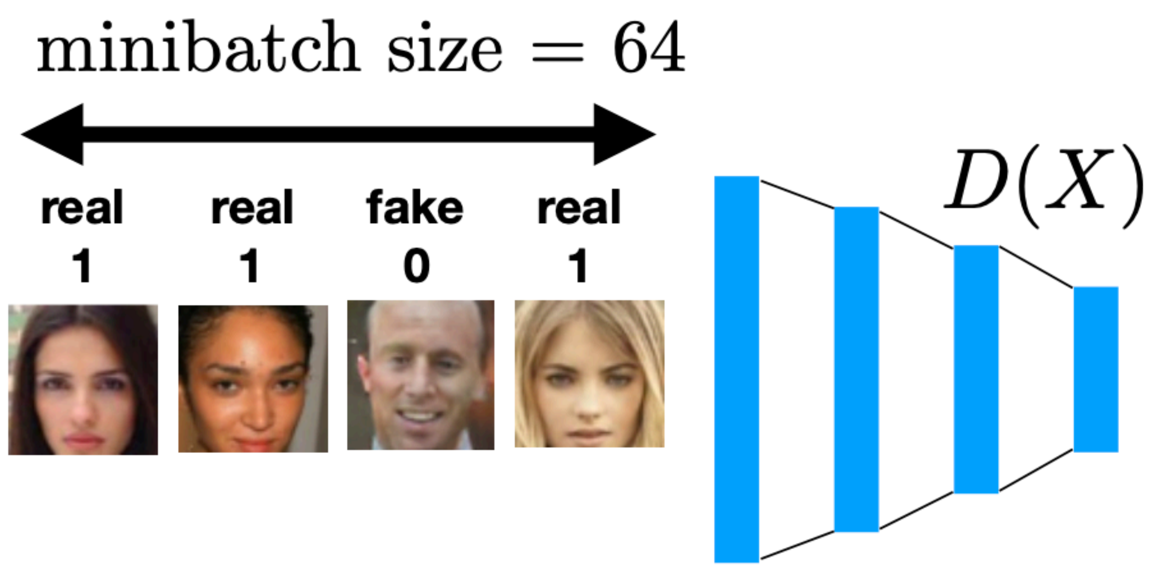


	Modes (Max 1000)
DCGAN	99.0
ALI	16.0
Unrolled GAN	48.7
VEEGAN	150.0
PacDCGAN2	1000.0
PacDCGAN3	1000.0
PacDCGAN4	1000.0

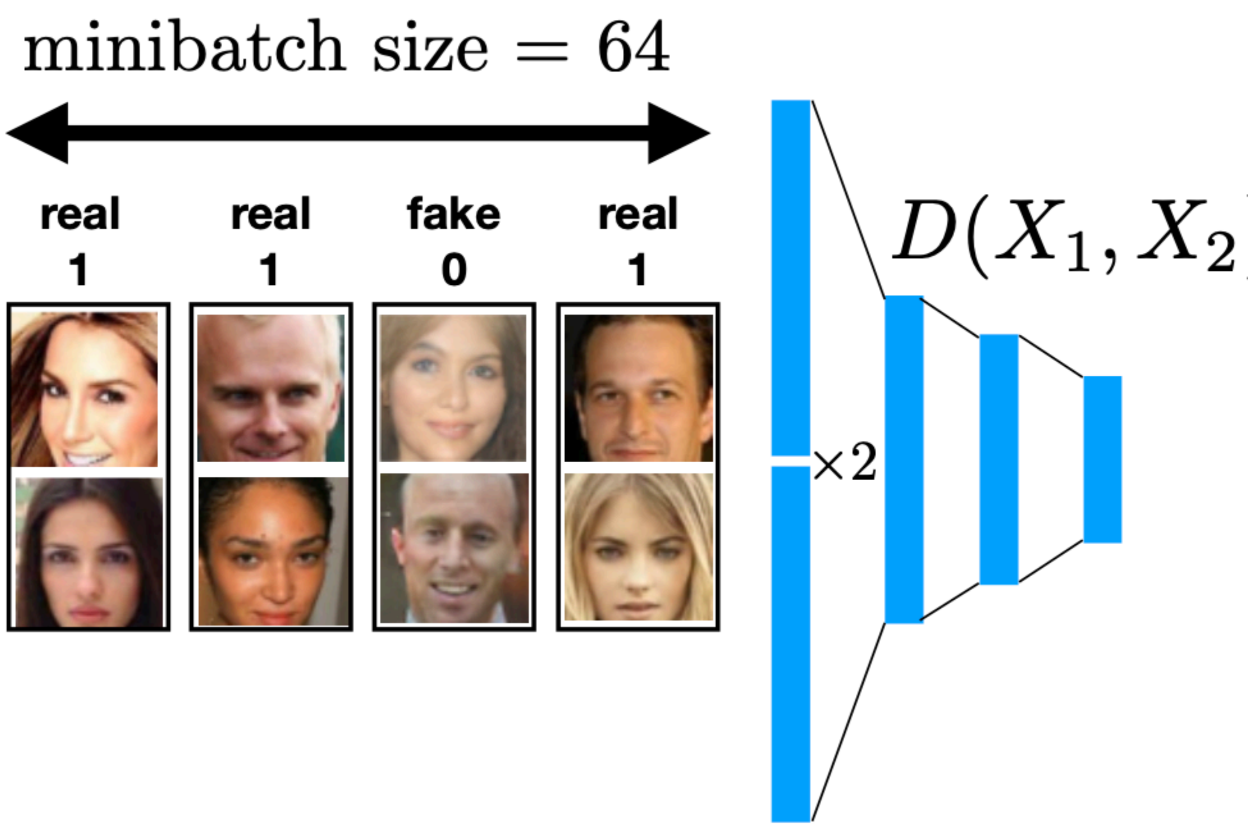
Principled approach to mode collapse

- Could PacGAN be cheating, as it is a larger discriminator network?

1. Discriminator size



GAN

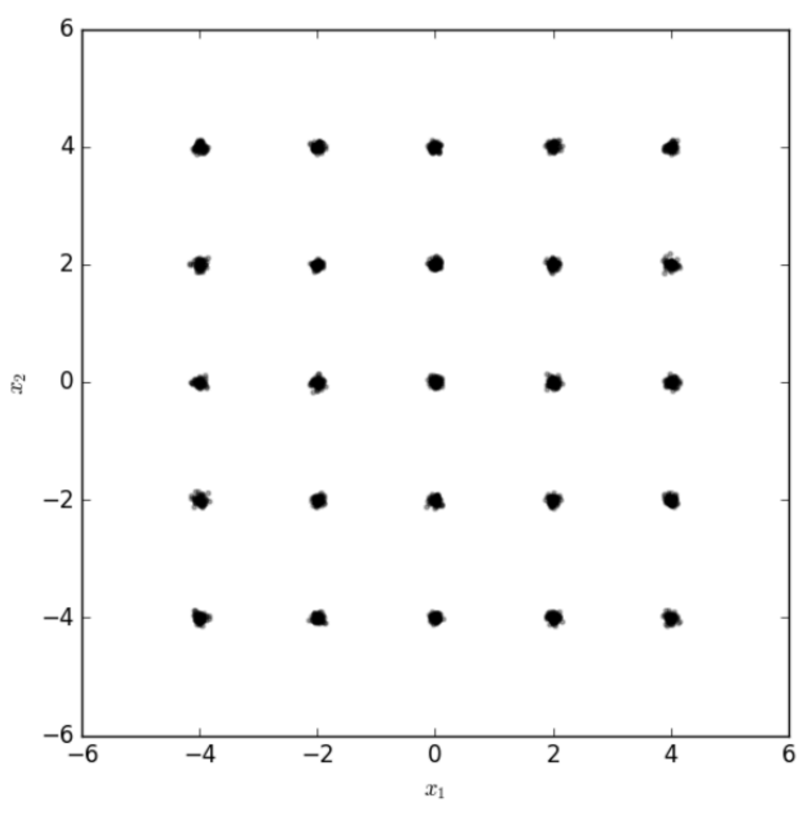


PacGAN2

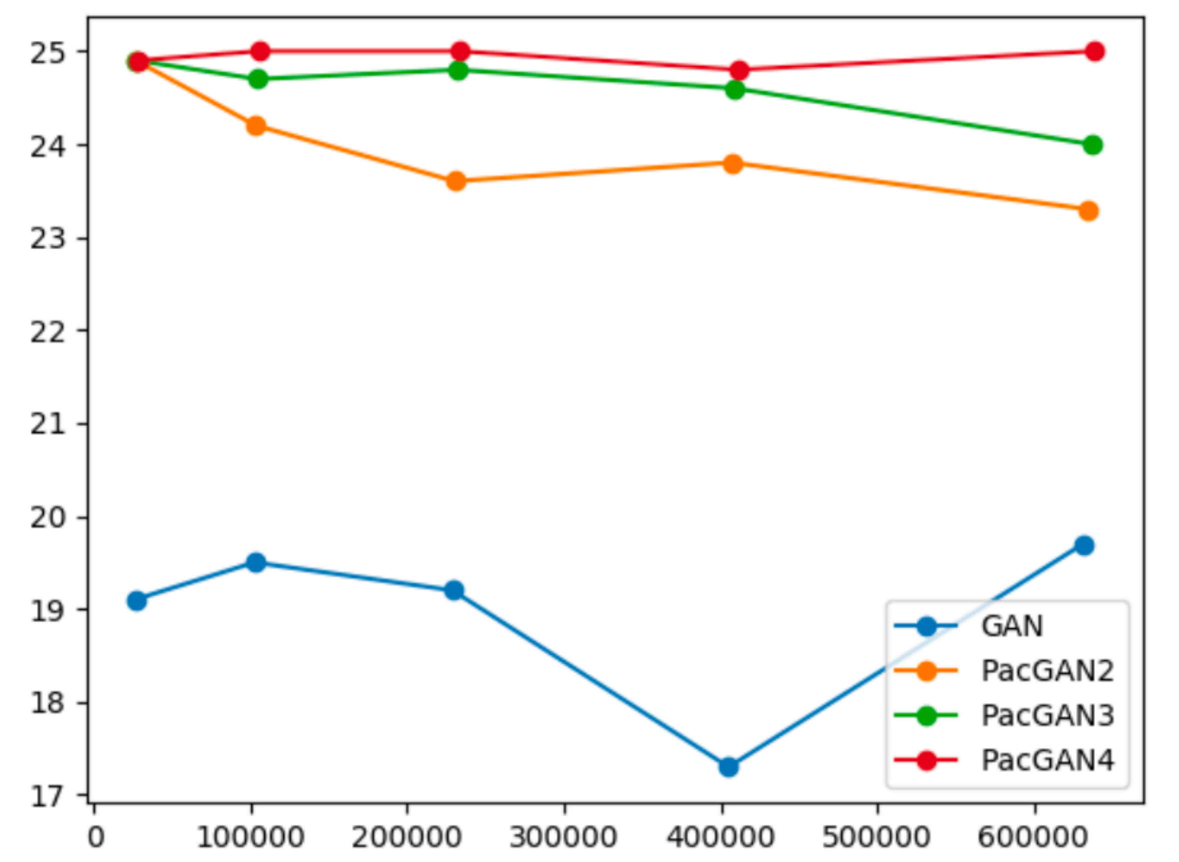
Principled approach to mode collapse

- Could PacGAN be cheating, as it is a larger discriminator network?

1. Discriminator size



modes captured

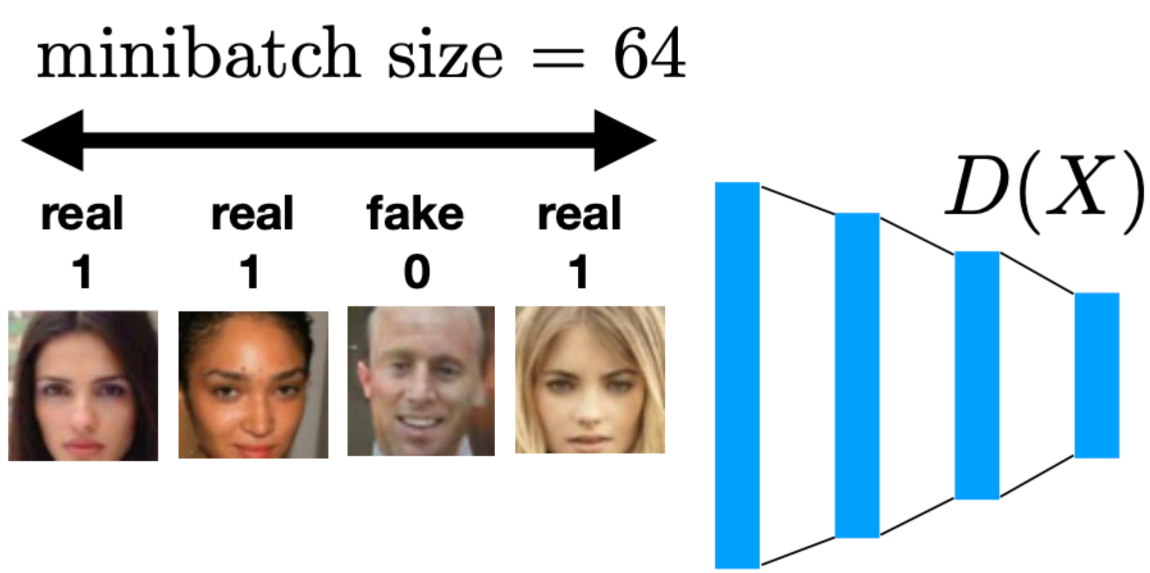


# of parameters in $D(\cdot)$

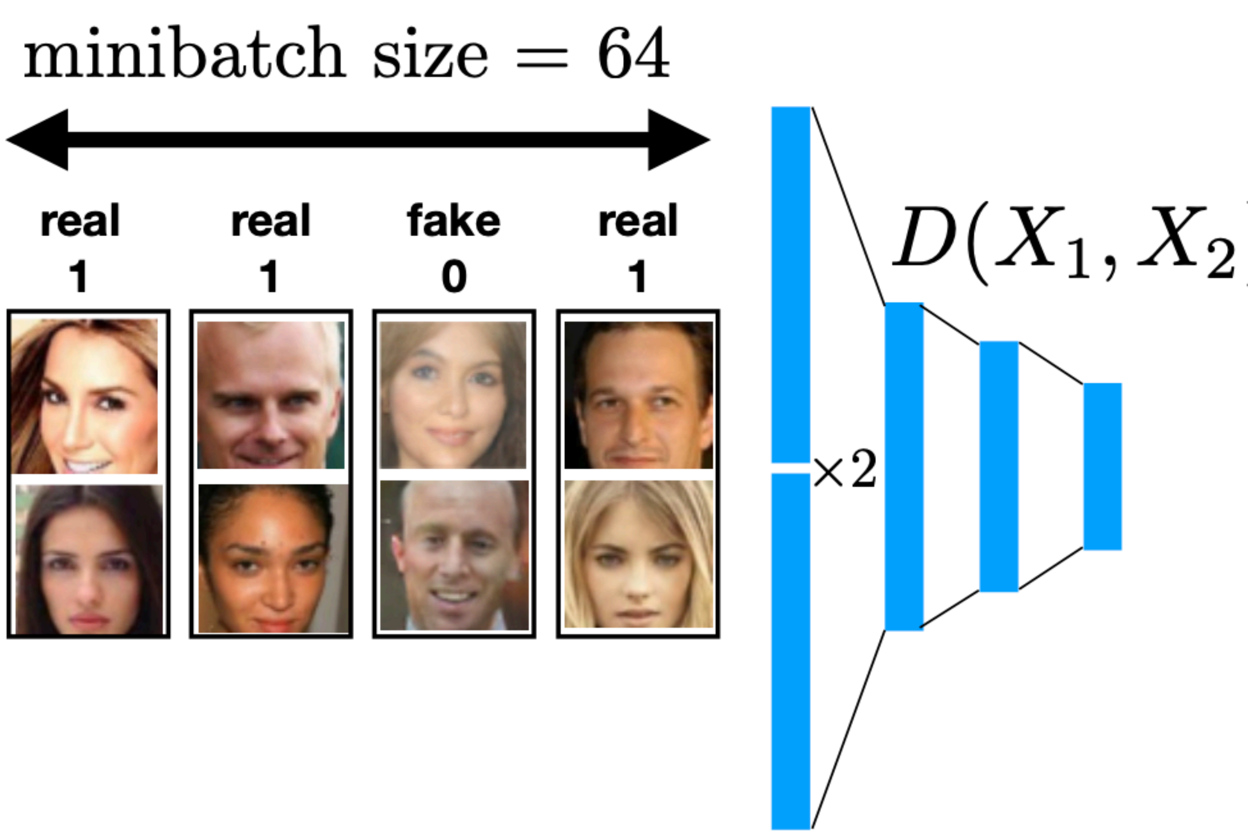
Principled approach to mode collapse

- Could PacGAN be cheating, as it uses more samples at each mini-batch?

1. Discriminator size



GAN

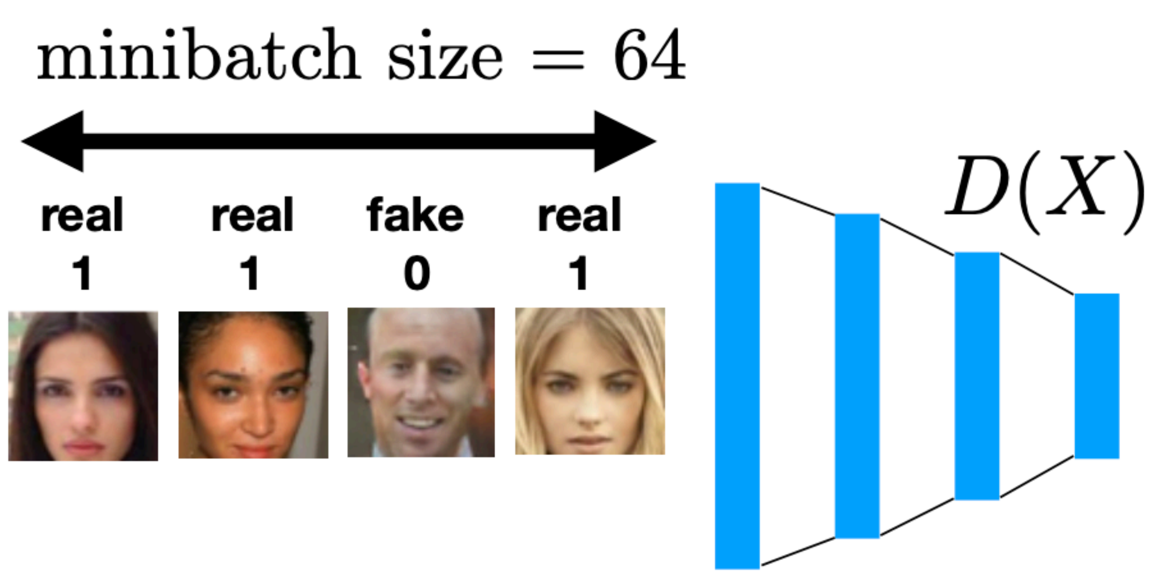


PacGAN2

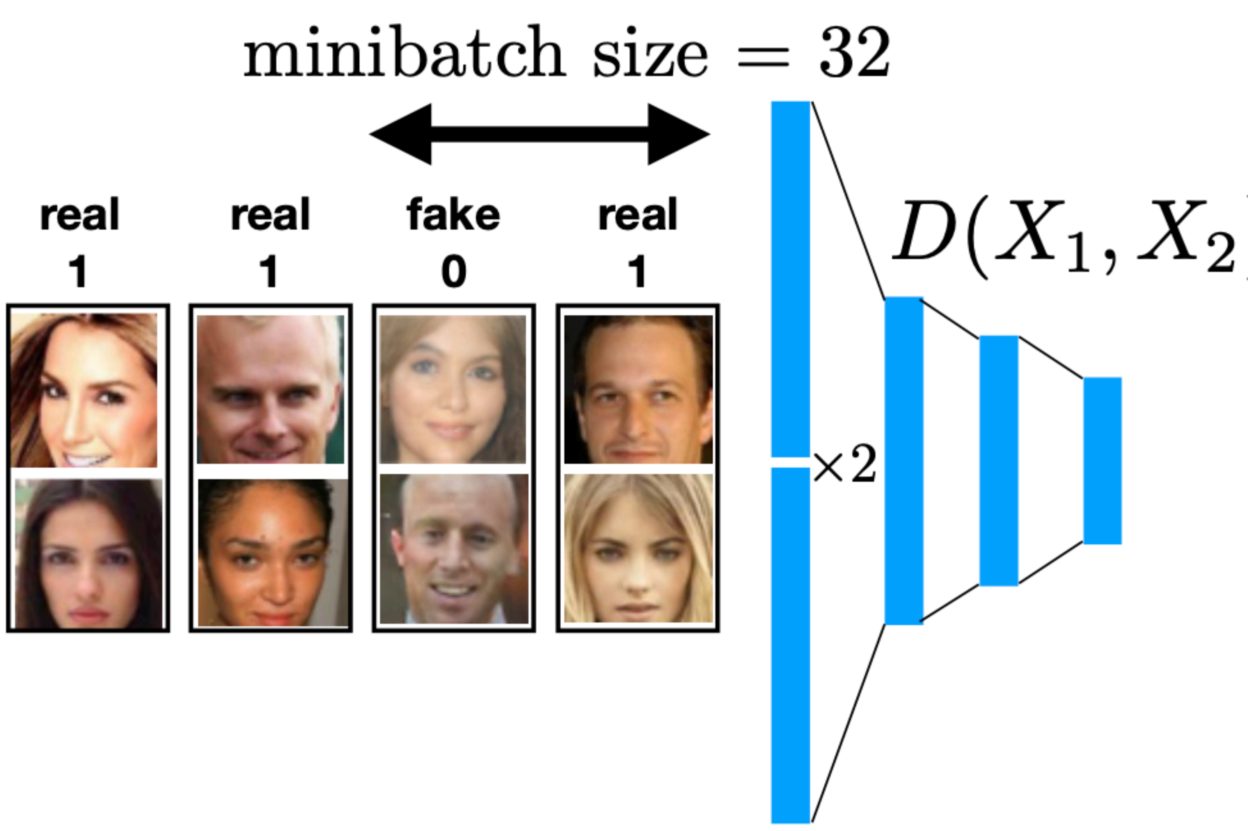
Principled approach to mode collapse

- Could PacGAN be cheating, as it uses more samples at each mini-batch?

2. Minibatch size



GAN

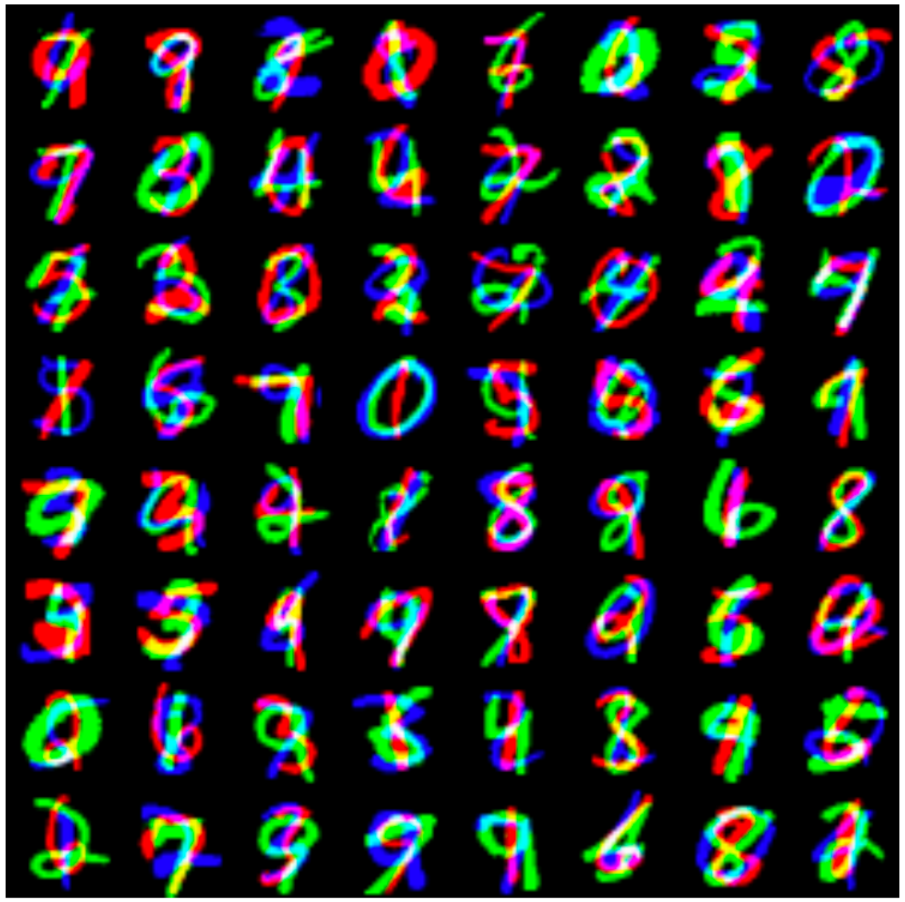


PacGAN2

Principled approach to mode collapse

- Could PacGAN be cheating, as it uses more samples at each mini-batch?

2. Minibatch size



	Modes
DCGAN	99.0
PacDCGAN2	1000.0

Theoretical intuition behind PacGAN

- Typical Gan training loss is

$$\min_{w} \max_{\theta} \sum_{x_i \sim P(\cdot)} \log D_\theta(x_i) + \sum_{z_i \sim N(0,\mathbf{I})} \log(1 - D_\theta(G_W(z_i)))$$

- We will consider

$$\min_{w} \max_{\theta} \sum_{x_i \sim P(\cdot)} D_\theta(x_i) + \sum_{z_i \sim N(0,\mathbf{I})} (1 - D_\theta(G_W(z_i)))$$

$$\text{subject to} \quad |D_\theta(x)| \leq 1 , \qquad \text{for all } x$$

Theoretical intuition behind PacGAN

- We will consider

$$\min_{w}\max_{\theta} \sum_{x_i \sim P(\cdot)} D_\theta(x_i) + \sum_{z_i \sim N(0,\mathbf{I})} (1 - D_\theta(G_W(z_i)))$$

$$\text{subject to} \quad |D_\theta(x)| \le 1\,, \qquad \text{for all } x$$

- this is a finite sample approximation of the following expectation

$$\min_{w}\max_{\theta} \ \mathbb{E}_{x\sim P(\cdot)}\big[D_\theta(x)\big] + \mathbb{E}_{z\sim N(0,\mathbf{I})}\big[1 - D_\theta(G_W(z))\big]$$

- let $Q(\cdot)$ denote the distribution of the generator $G_w(z_i)$

$$\min_{Q(\cdot)}\max_{\theta} \ \mathbb{E}_{x\sim P(\cdot)}\big[D_\theta(x)\big] + \mathbb{E}_{x\sim Q(\cdot)}\big[1 - D_\theta(x)\big]$$

$$\text{subject to} \quad |D_\theta(x)| \le 1\,, \qquad \text{for all } x$$

- at this point, we can solve the maximization w.r.t. D_θ assuming it can represent any functions (for the purpose of theoretical analysis)
    - the optimal solution is

$$D_\theta(x) = \begin{cases} +1 & \text{if } P(x) \ge Q(x) \\ -1 & \text{if } P(x) < Q(x) \end{cases}$$

Theoretical intuition behind PacGAN

$$\min_{Q(\cdot)} \max_{\theta} \quad \mathbb{E}_{x\sim P(\cdot)}\big[\, D_\theta(x) \,\big] + \mathbb{E}_{x\sim Q(\cdot)}\big[\, 1 - D_\theta(x) \,\big]$$

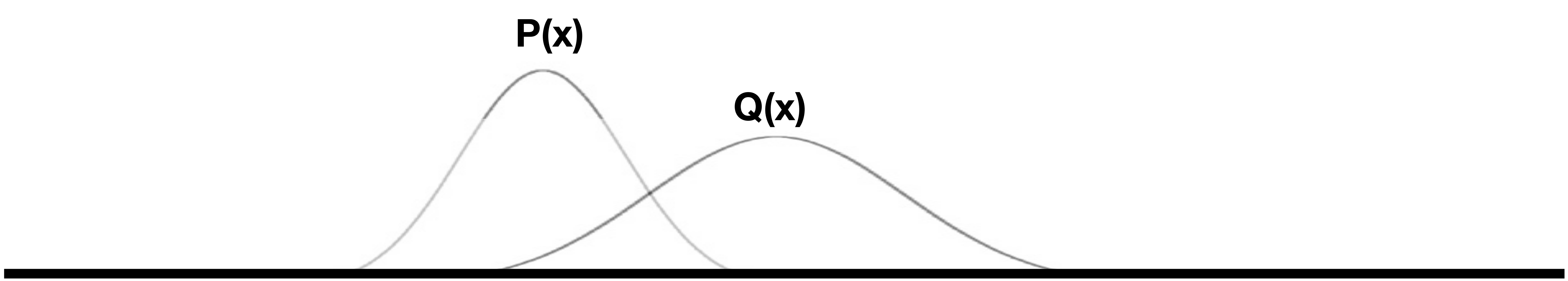
$$\text{subject to} \quad |\, D_\theta(x) \,| \leq 1 \,, \qquad \text{for all } x$$

- at this point, we can solve the maximization w.r.t. D_θ assuming it can represent any functions (for the purpose of theoretical analysis)
  - the optimal solution is

$$D_\theta(x) \;=\; \begin{cases} +1 & \text{if } P(x) \geq Q(x) \\ -1 & \text{if } P(x) < Q(x) \end{cases}$$

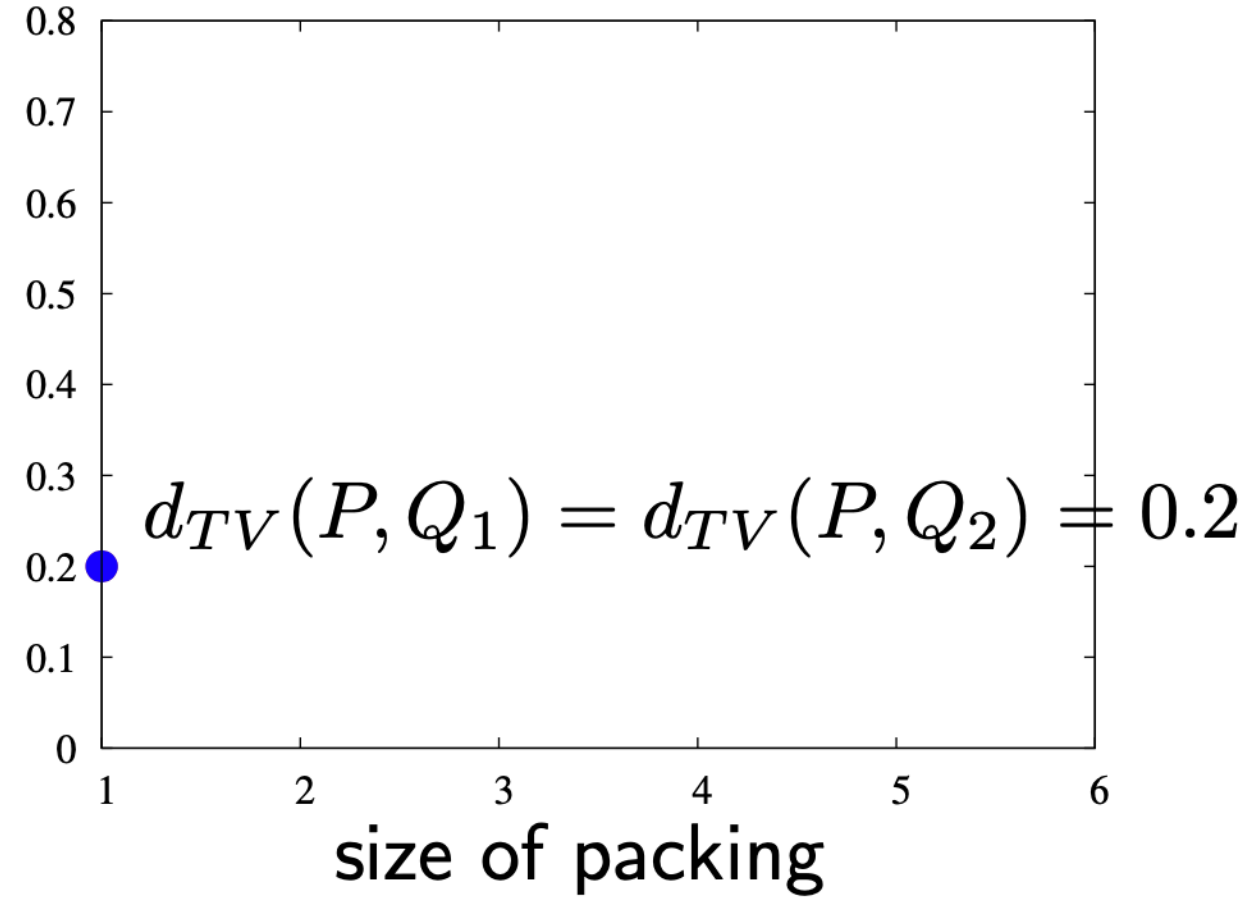
- Plugging this back in to the loss, we get

$$\min_{Q(\cdot)} \; D_{\mathrm{TV}}(P, Q) \;=\; \mathbb{E}_{x\sim P(\cdot)}\left[\, \left| 1 - \frac{Q(x)}{P(x)} \right| \,\right]$$

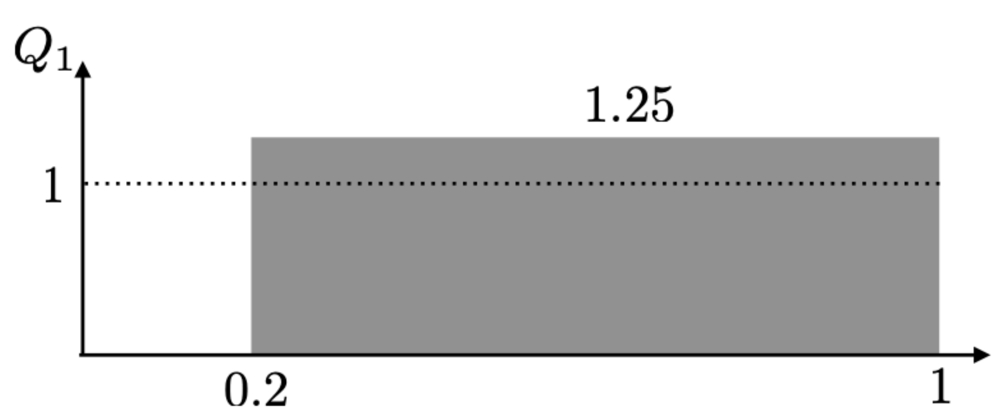


Theoretical intuition behind PacGAN

Target distribution P

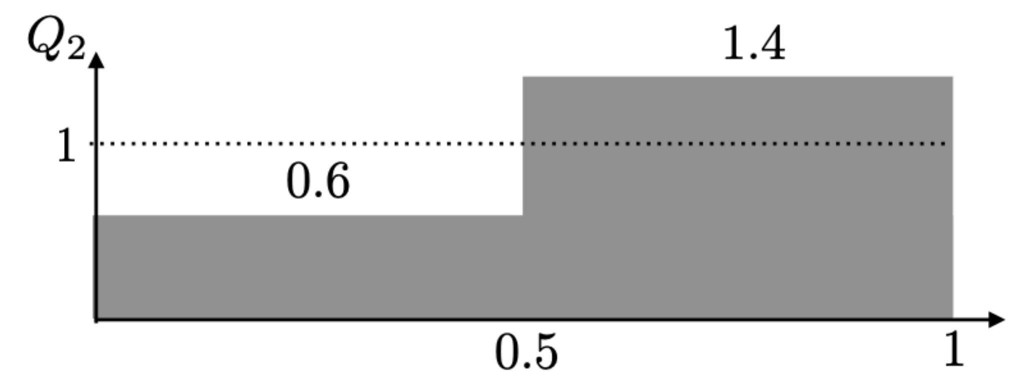


Generator Q_1
with mode collapse



$d_{\mathrm{TV}}(P, Q_1) = 0.2$

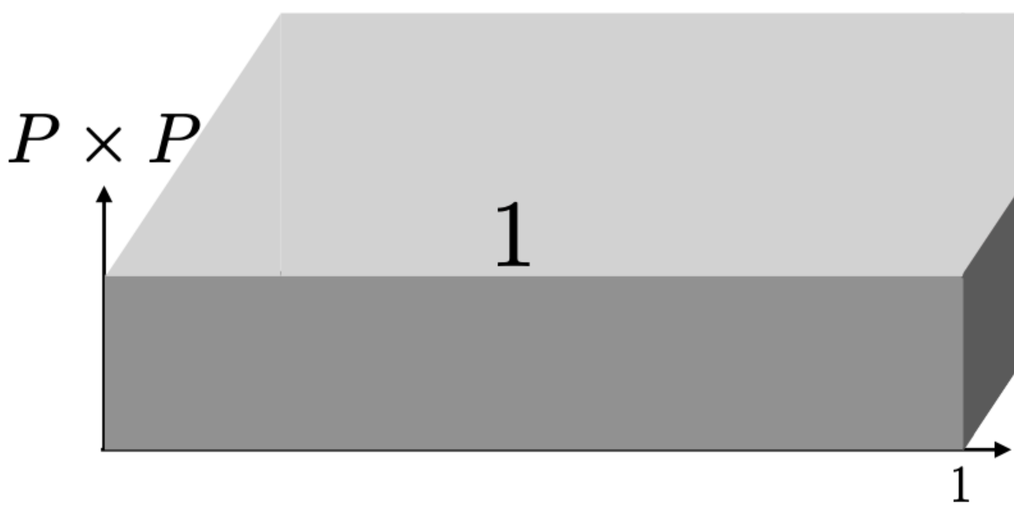
Generator Q_2
without mode collapse

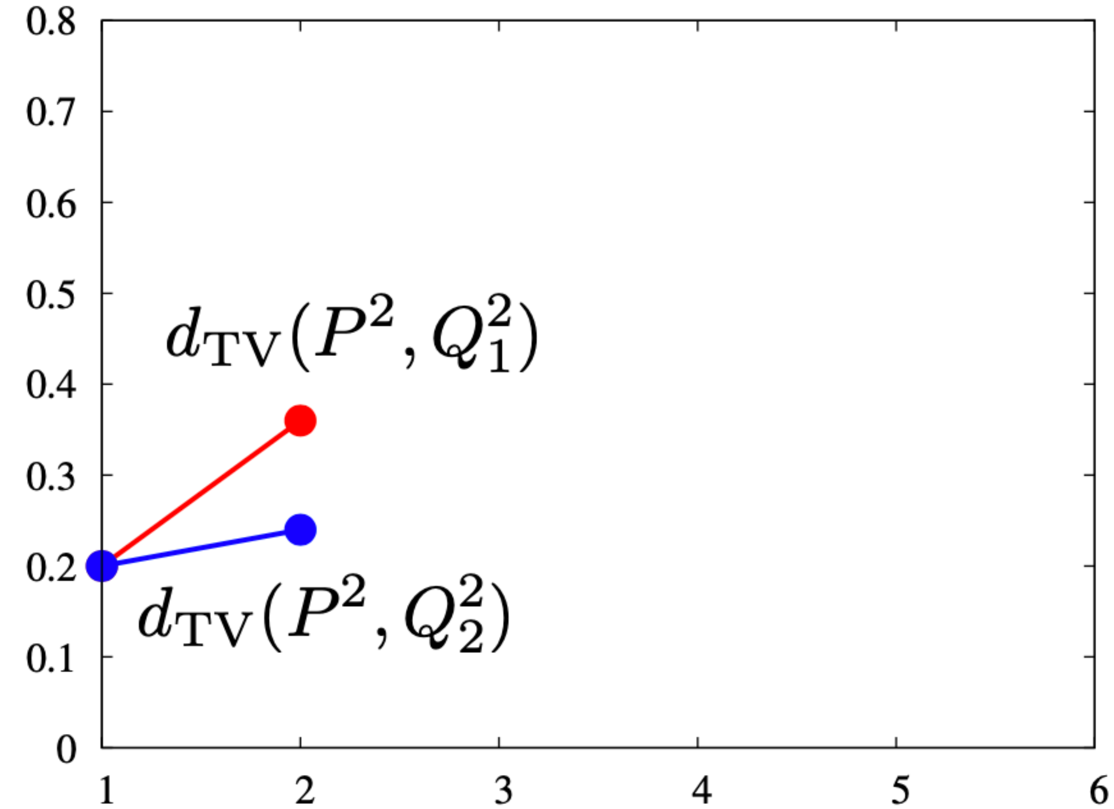


$d_{\mathrm{TV}}(P, Q_2) = 0.2$

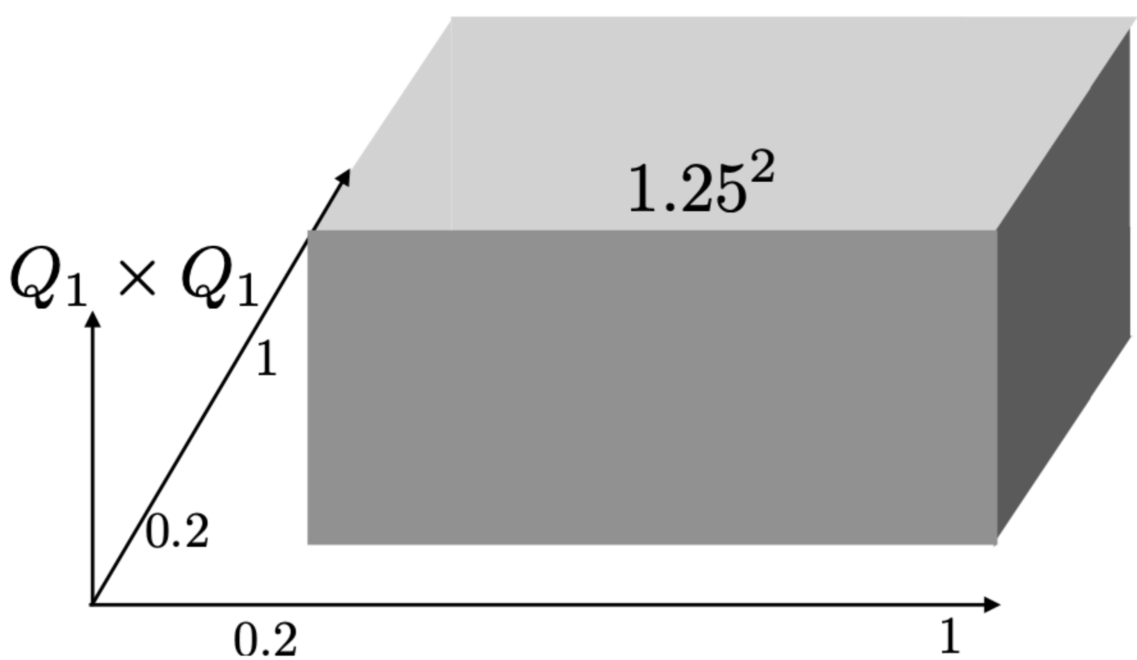
Theoretical intuition behind PacGAN

Target distribution P



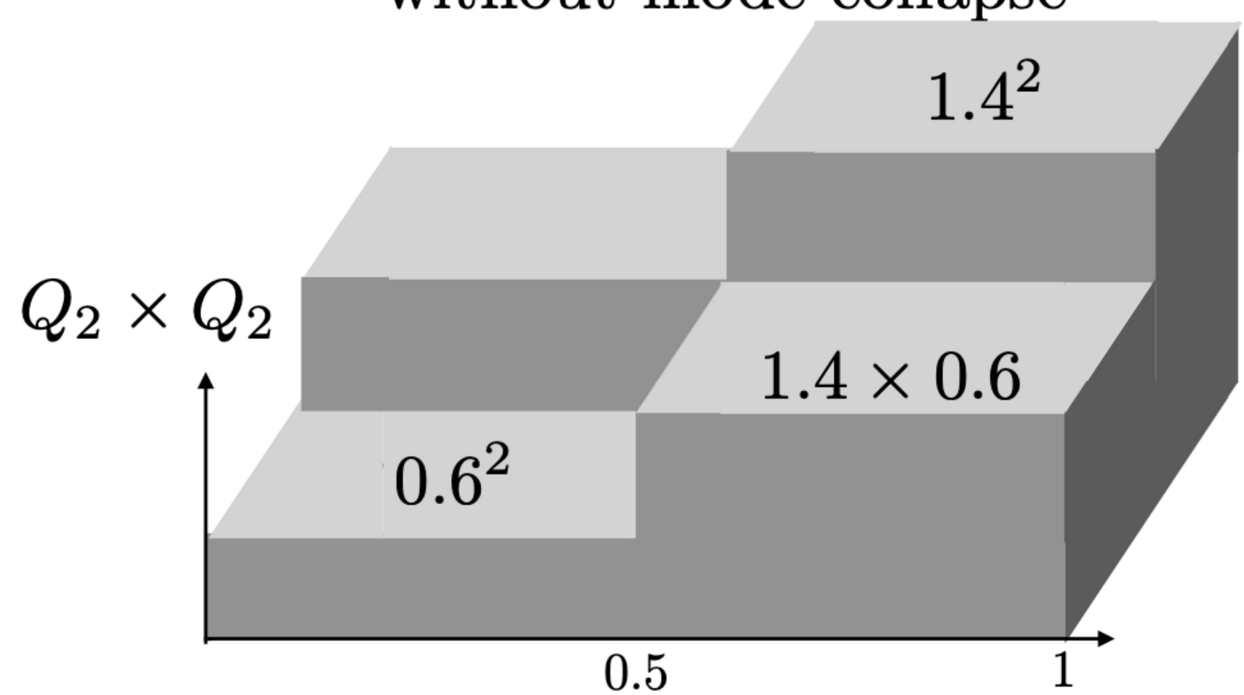


Generator Q_1
with mode collapse



$d_{\mathrm{TV}}(P \times P, Q_1 \times Q_1) = 0.36$

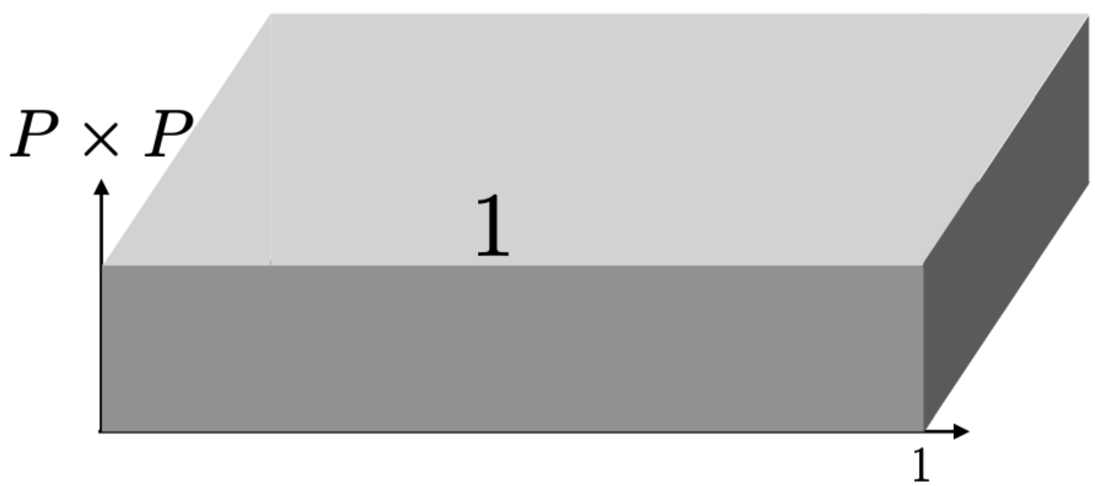
Generator Q_2
without mode collapse

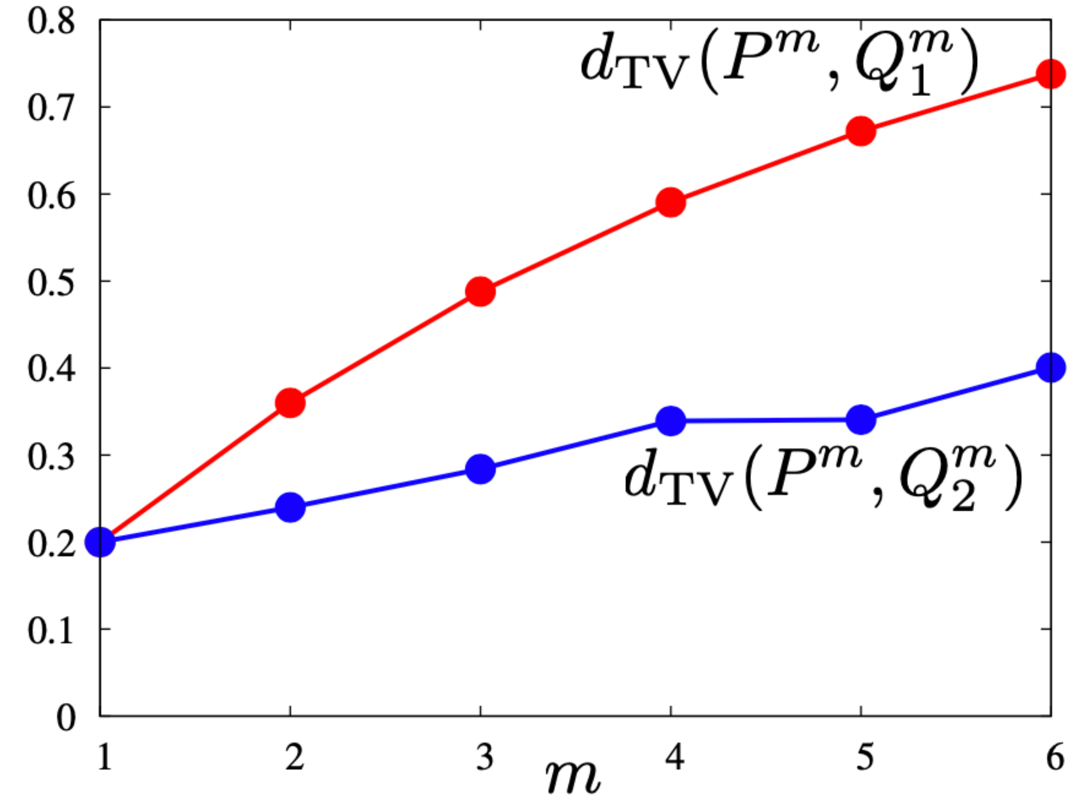


$d_{\mathrm{TV}}(P \times P, Q_2 \times Q_2) = 0.24$

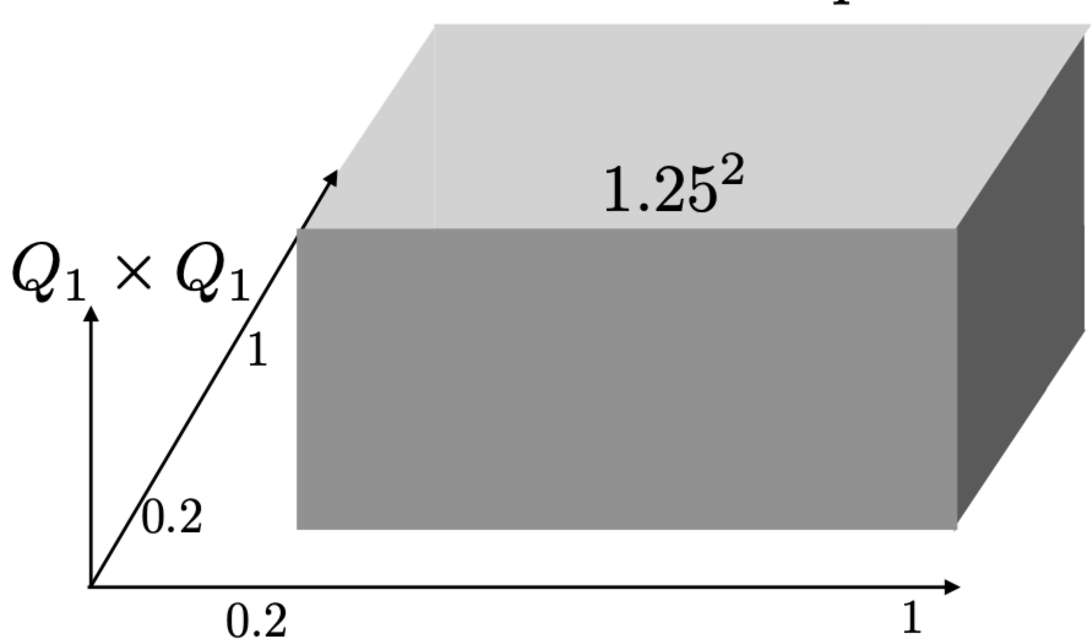
Theoretical intuition behind PacGAN

Target distribution P



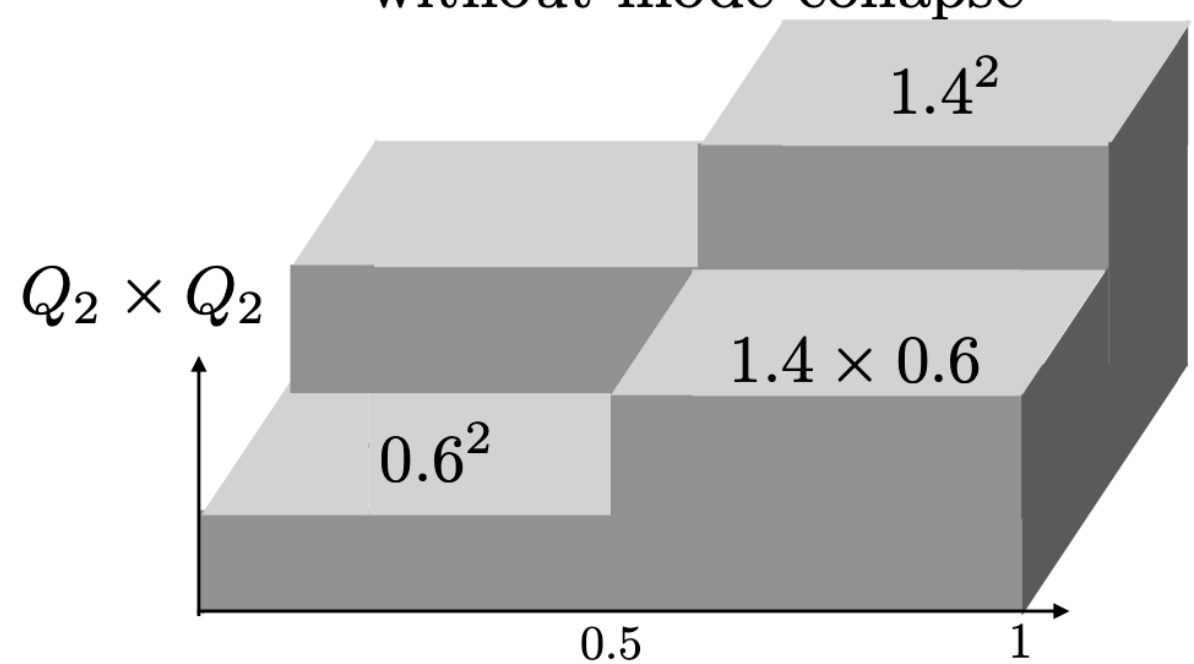


Generator Q_1
with mode collapse



$d_{\mathrm{TV}}(P \times P, Q_1 \times Q_1) = 0.36$

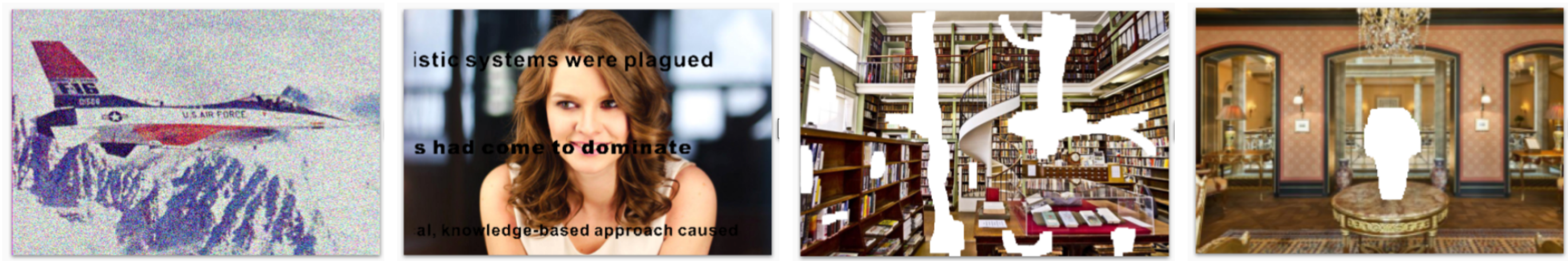
Generator Q_2
without mode collapse



$d_{\mathrm{TV}}(P \times P, Q_2 \times Q_2) = 0.24$

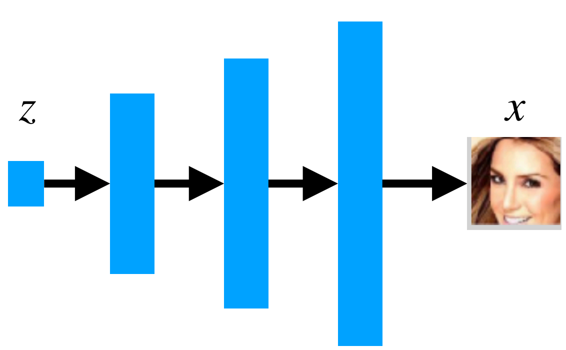
Deep Image prior

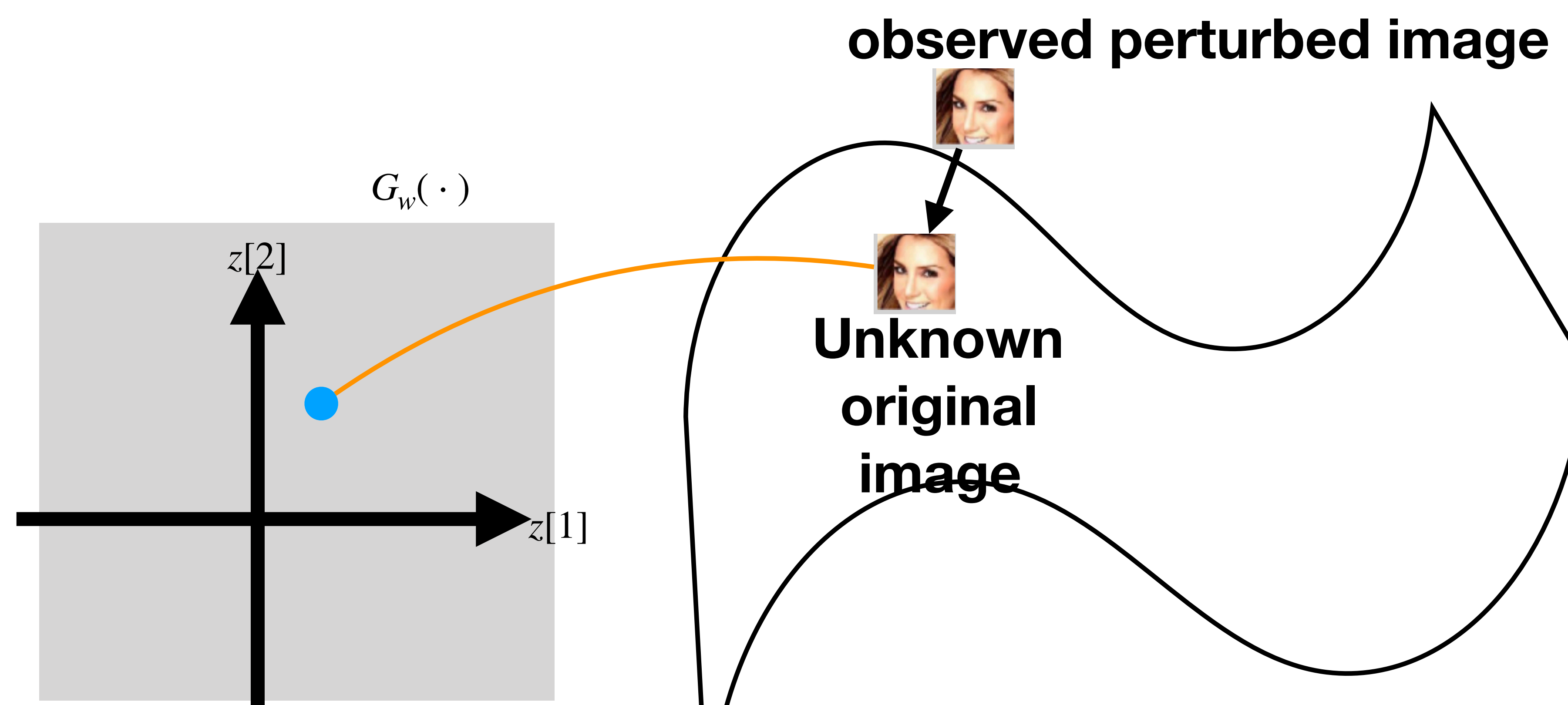
- in standard de-noising/inpainting with **trained** GAN
  we want to recover original image from some distortion



Corrupted Corrupted Corrupted Corrupted

- if we have a GAN trained on similar class of images, then we can use the latent space and the manifold of natural images to recover the image as follows



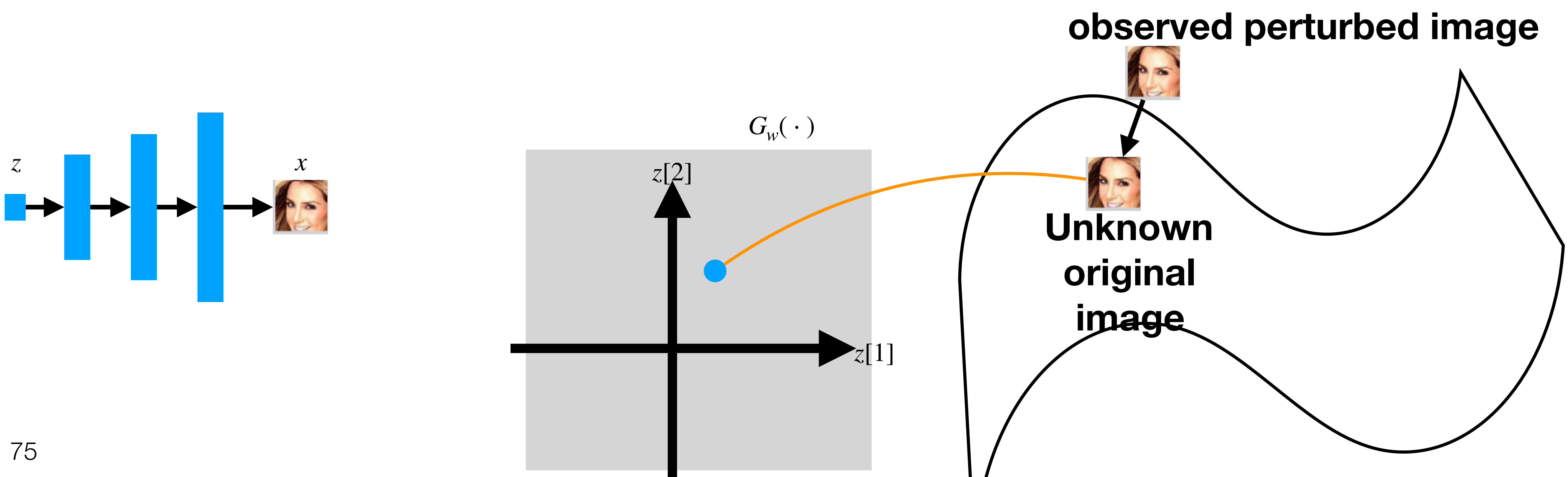


Deep Image prior

- Given a trained generator w that knows the manifold of natural images, find the latent vector z that

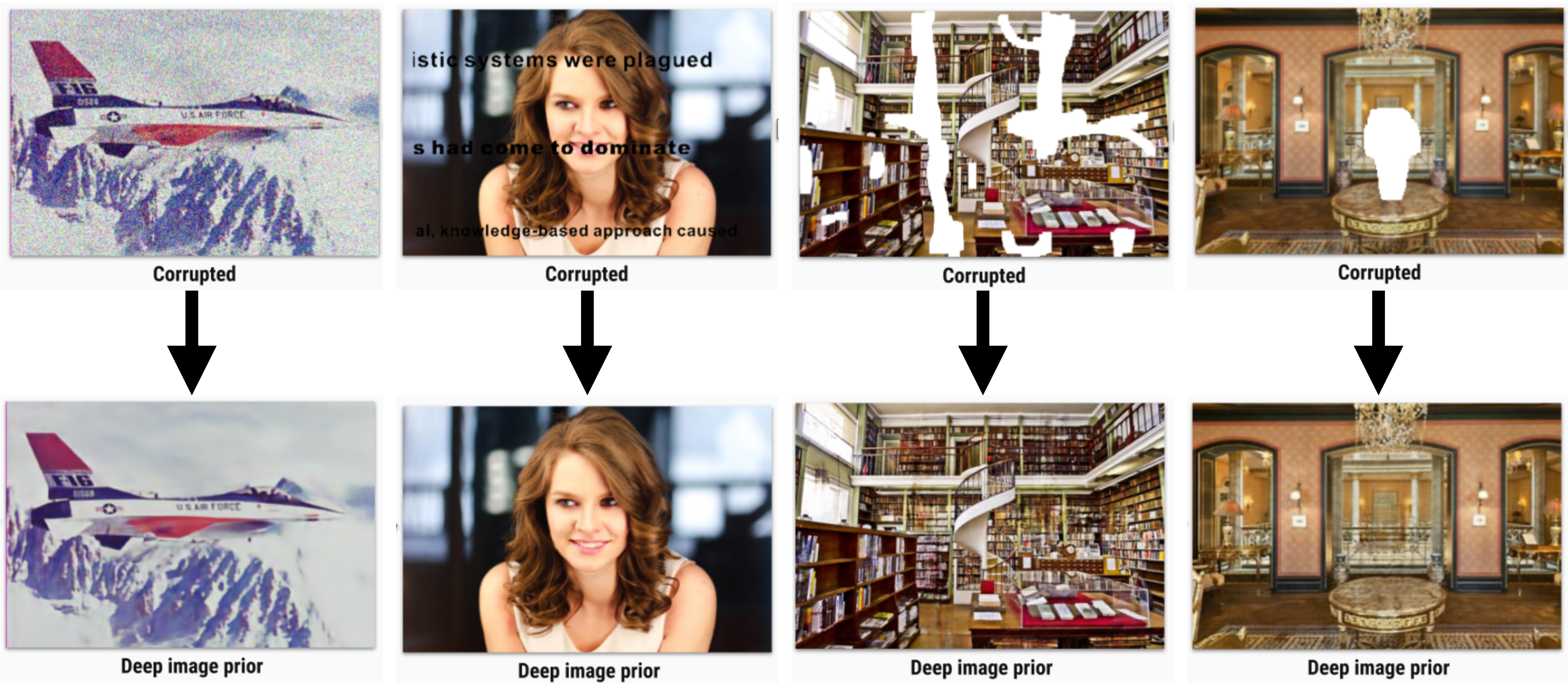
$$\text{minimize}_z \quad \ell\Big(G_w(z), \quad \Big)$$

- let $G_w(z)$ be the recovered image



Deep image prior

- **deep image prior** does amazing recovery, **without training**



Deep image prior

- fix z to be something random and find w that

  $$\text{minimize}_z \quad \ell\Big(G_w(z),$$

  Corrupted

  $$\Big)$$

  and let $G_w(z)$ be the recovered image

https://www.youtube.com/watch?v=kSLJriaOumA&feature=youtu.be