



Lecture 19: Deep Generative Models III: VAE & GAN

Lan Xu
SIST, ShanghaiTech
Fall, 2022

Outline

- VAEs
 - Inverse graphics network
 - Attribute2Image
- Generative Adversarial Networks
 - Implicit generative models
 - Adversarial learning

Acknowledgement: Feifei Li et al's cs231n notes

The Variational Autencoder: overview

■ Learning:

- Given a large dataset of observations $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$
- Estimating the parameters in Deep LVM

$$p(x) = \int p(x, z) dz \quad \text{where} \quad p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} | \mathbf{z})p(\mathbf{z})$$

- Based on Maximum Likelihood

$$\max \sum_{i=1}^N \log p(x^{(i)})$$

- Direct optimization is challenging: use EM learning strategy
- Jointly learning inference model with the deep latent variable model

VAE objective

- Recall lower bound of the data log likelihood

$$\begin{aligned}\log p_{\theta}(x) &= \log \int_z p_{\theta}(x, z) dz \\ &= \log \int_z q_{\phi}(z|x) \frac{p_{\theta}(x, z)}{q_{\phi}(z|x)} dz \\ &\geq \int_z q_{\phi}(z|x) \log \frac{p_{\theta}(x, z)}{q_{\phi}(z|x)} dz \quad (\text{Jensen's Inequality}) \\ &= \mathbf{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x, z) - \log q_{\phi}(z|x)] = \mathcal{L}(x; \theta, \phi)\end{aligned}$$

$$\log p_{\theta}(x) = \boxed{\mathcal{L}(x; \theta, \phi)} + D_{KL}(q_{\phi}(z|x) || p_{\theta}(z|x))$$

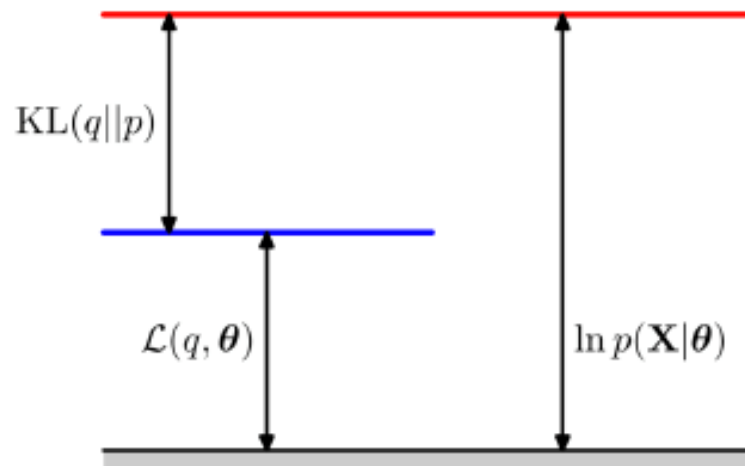
- Learning: maximize the lower bound of data likelihood
- The evidence lower bound (ELBO)

Review: VAEs

■ Main ideas:

- Introduce a parametric model $q_\phi(z | x)$ to approximate the true posterior $p_\theta(z | x)$
- Jointly learn approximate posterior with the deep latent variable model
- Variational EM: lower-bound of the Maximum Likelihood

$$\log p_\theta(x) = \mathcal{L}(x; \theta, \phi) + D_{KL}(q_\phi(z|x) || p_\theta(z|x))$$



Review: VAEs

■ Main ideas:

- Introduce a parametric model $q_\phi(z | x)$ to approximate the true posterior $p_\theta(z | x)$
- Jointly learn approximate posterior with the deep latent variable model
- Variational EM: lower-bound of the Maximum Likelihood

$$\begin{aligned}\mathcal{L}(\theta, \phi, x) &= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x, z) - \log q_\phi(z | x)] \\ &= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x | z) + \log p_\theta(z) - \log q_\phi(z | x)] \\ &= \underbrace{-D_{\text{KL}}(q_\phi(z | x) \| p_\theta(z))}_{\text{regularization term}} + \underbrace{\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x | z)]}_{\text{reconstruction term}}\end{aligned}$$

VAE learning

■ EM perspective

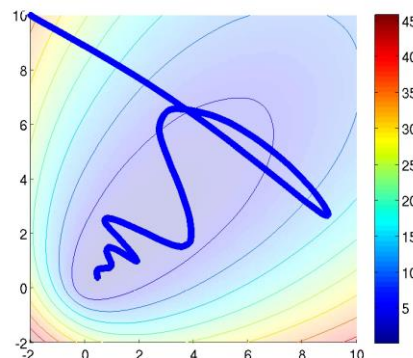
- Expectation Maximization alternately optimizes the ELBO, $\mathcal{L}(q, \theta)$, with respect to q (the E step) and θ (the M step)
- Initialize $\theta^{(0)}$
- At each iteration $t = 1, \dots$
 - **E step:** Hold $\theta^{(t-1)}$ fixed, find $q^{(t)}$ which maximizes $\mathcal{L}(q, \theta^{(t-1)})$
 - **M step:** Hold $q^{(t)}$ fixed, find $\theta^{(t)}$ which maximizes $\mathcal{L}(q^{(t)}, \theta)$

$$\mathcal{L}(x, \phi, \theta) = E_{q_{\phi}(z|x)}[\log p_{\theta}(x, z) - \log q_{\phi}(z|x)]$$

Two views of Learning VAE

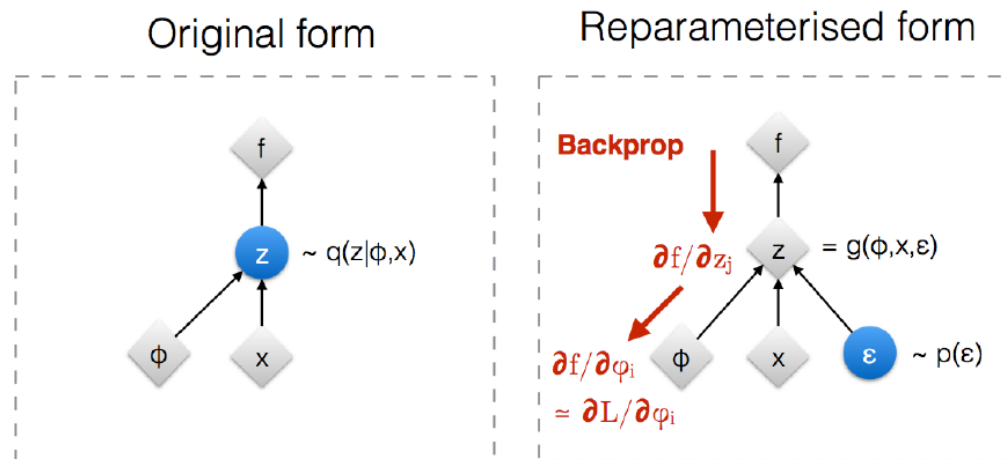
■ Optimization interpretation

- Stochastic gradient-based



■ Network interpretation

- Backpropagation



Optimization interpretation

- Recall VAE objective

$$\mathcal{L}(x, \phi, \theta) = E_{q_{\phi}(z|x)}[\log p_{\theta}(x, z) - \log q_{\phi}(z|x)]$$

- Or rewrite as $\mathcal{L}(x, \phi, \theta) = E_{q_{\phi}(z|x)}[f_{\phi, \theta}(x, z)]$

- Often no analytic solution to exact gradient

$$\nabla_{\phi, \theta} \mathcal{L}(x, \phi, \theta)$$

- Solution: stochastic gradient ascent
 - Requires unbiased estimates of gradient
 - Can use small minibatches or single point of data

$$\nabla_{\phi} \mathcal{L}(x, \phi, \theta) \approx \nabla_{\phi} f_{\phi, \theta}(x, z^{(i)}), \quad z^{(i)} \sim q_{\phi}(z|x)$$

High variance for gradient estimation

Reparameterization trick

- Reparameterize $\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z}|\mathbf{x})$ using a differentiable transformation of an auxiliary noise variable ϵ

$$\mathbf{z} = g_\phi(\epsilon, \mathbf{x}) \quad \text{with} \quad \epsilon \sim q(\epsilon)$$

- Then we can write the ELBO as

$$\mathcal{L}(x, \phi, \theta) = E_{q_\phi(z|x)}[f_{\phi, \theta}(x, z)] = E_{q(\epsilon)}[f_{\phi, \theta}(x, g_\phi(\epsilon, \mathbf{x}))]$$

- And its gradient estimation with L samples

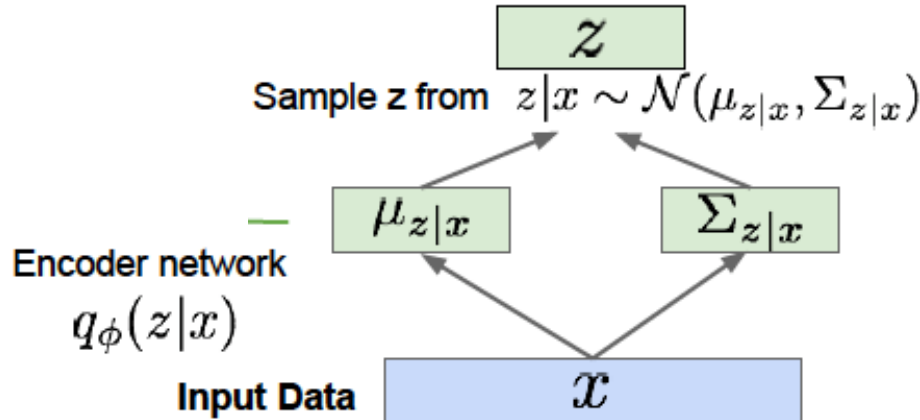
$$\nabla_\phi \mathcal{L}(x, \phi, \theta) = E_{q(\epsilon)}[\nabla_\phi f_{\phi, \theta}(x, z)] \approx \frac{1}{L} \sum_{i=1}^L \nabla_\phi f_{\phi, \theta}(x, g_\phi(\epsilon^{(i)}, x)), \quad \epsilon^{(i)} \sim q(\epsilon)$$

VAE Example

- Univariate Gaussian $z \sim p(z|x) = \mathcal{N}(\mu, \sigma^2)$

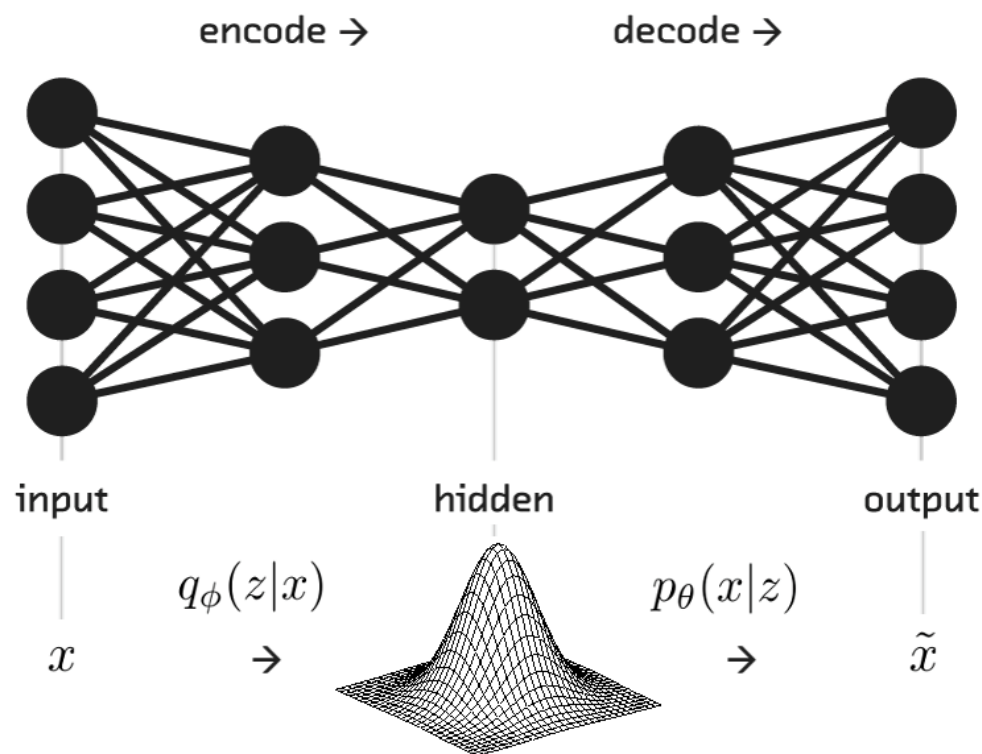
$$z = \mu + \sigma\epsilon \quad \epsilon \sim \mathcal{N}(0, 1)$$

$$\mathbb{E}_{\mathcal{N}(z;\mu,\sigma^2)} [f(z)] = \mathbb{E}_{\mathcal{N}(\epsilon;0,1)} [f(\mu + \sigma\epsilon)] \simeq \frac{1}{L} \sum_{l=1}^L f(\mu + \sigma\epsilon^{(l)})$$



Autoencoder Interpretation

- Objective $\mathcal{L}(x, \phi, \theta) = -D_{KL}(q_{\phi}(z|x)||p_{\theta}(z)) + E_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)]$
Regularization term Reconstruction term



VAE Example

■ Learning objective

Putting it all together: maximizing the likelihood lower bound

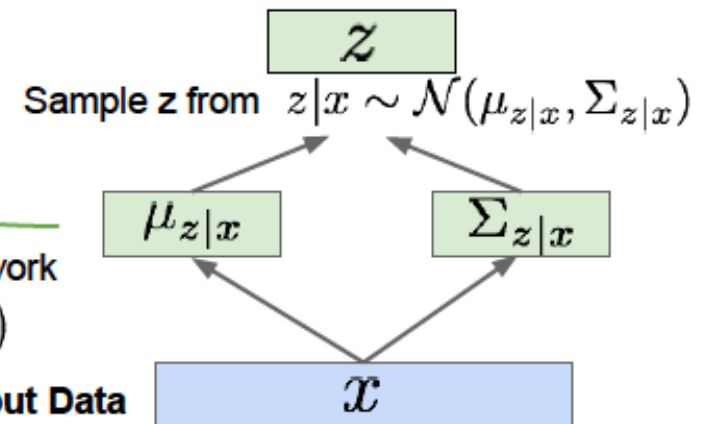
$$\underbrace{\mathbf{E}_z \left[\log p_\theta(x^{(i)} | z) \right] - D_{KL}(q_\phi(z | x^{(i)}) || p_\theta(z))}_{\mathcal{L}(x^{(i)}, \theta, \phi)}$$

Make approximate posterior distribution close to prior

Encoder network

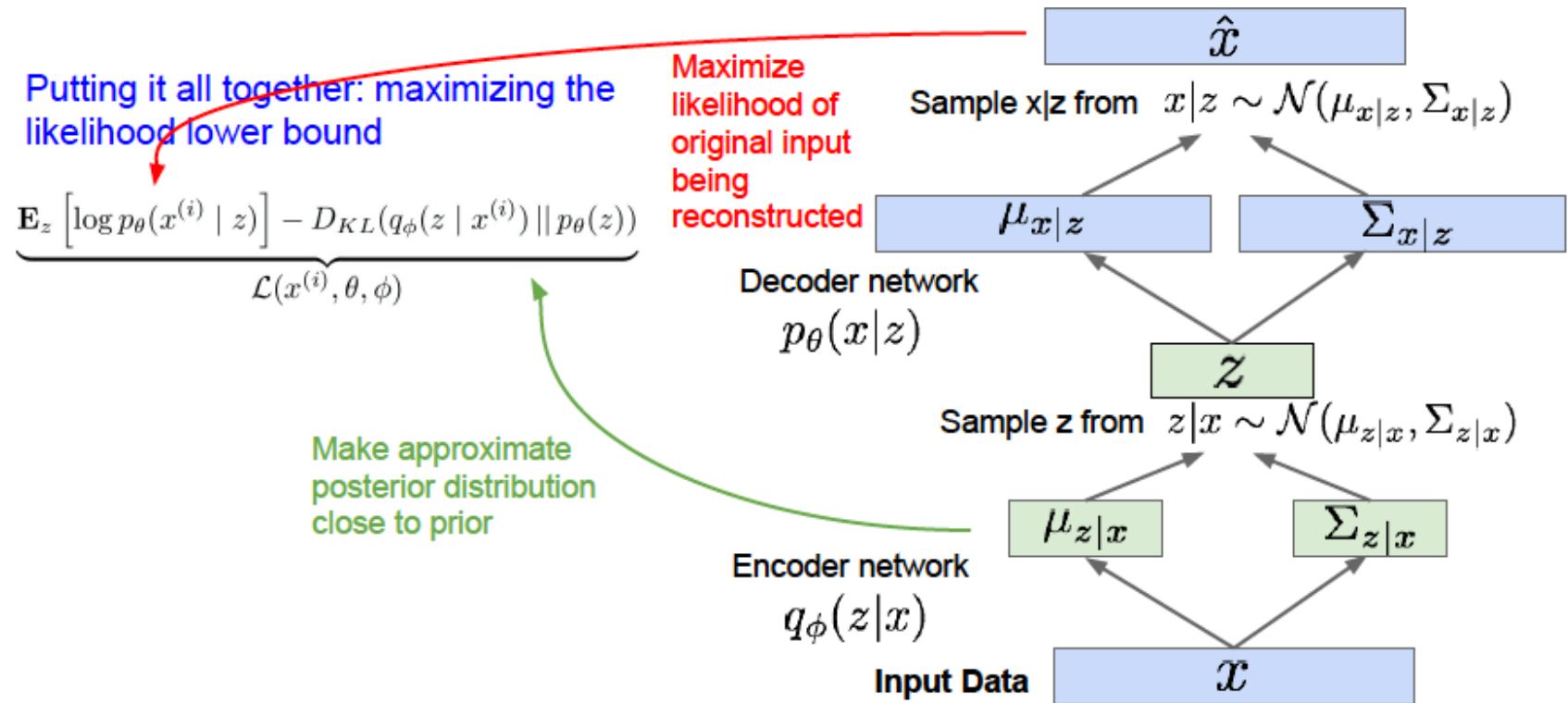
$$q_\phi(z|x)$$

Input Data



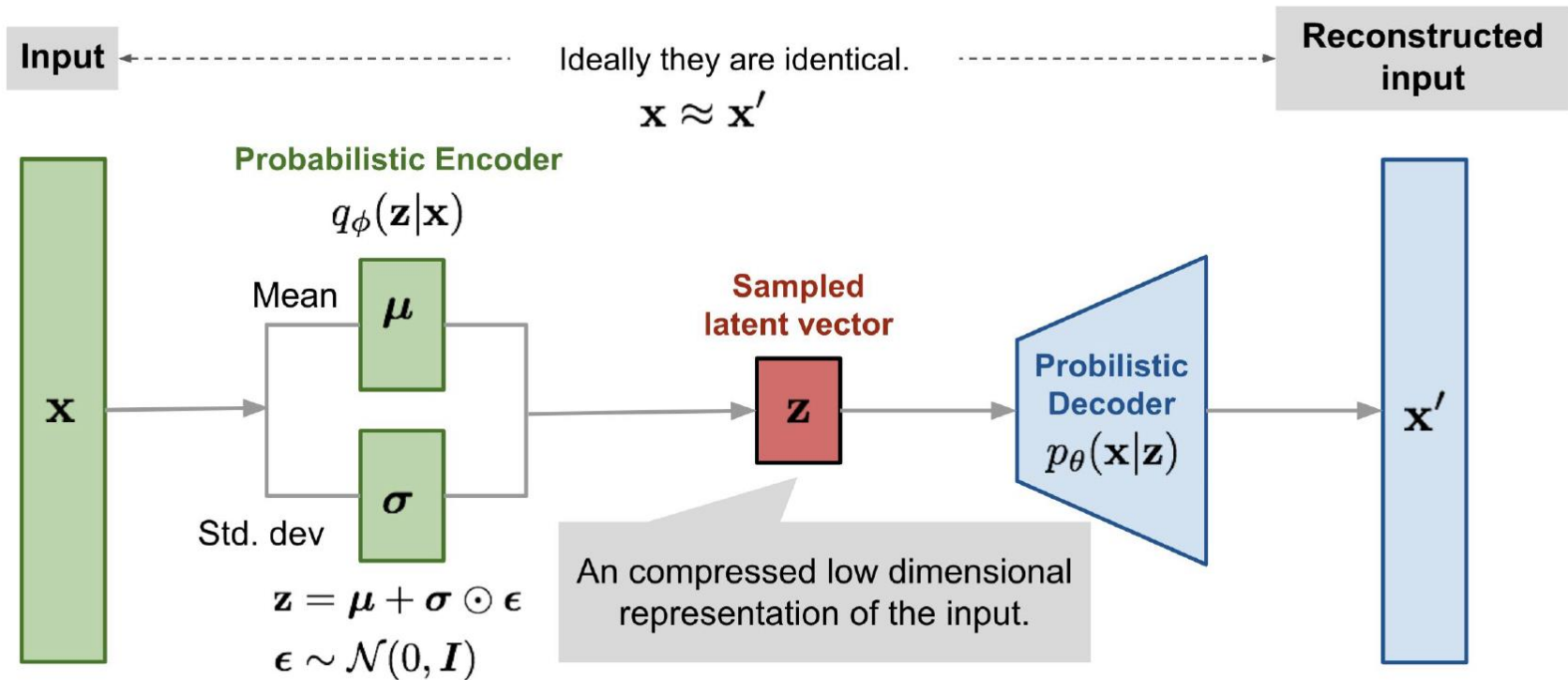
VAE Example

■ Learning objective



Autoencoder Interpretation

- Objective $\mathcal{L}(x, \phi, \theta) = \underbrace{-D_{KL}(q_{\phi}(z|x)||p_{\theta}(z))}_{\text{Regularization term}} + \underbrace{E_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)]}_{\text{Reconstruction term}}$



- The objective function can be represented as an Autoencoder-like **computation graph**.

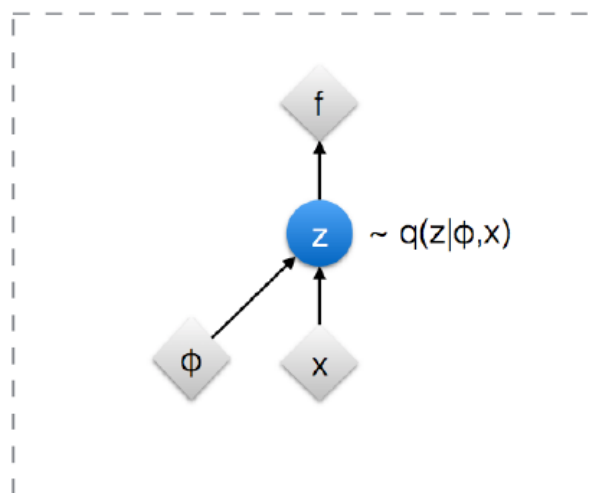
Network interpretation

$$\mathcal{L}(x, \phi, \theta) = E_{q_{\phi}(z|x)}[f_{\phi, \theta}(x, z)]$$

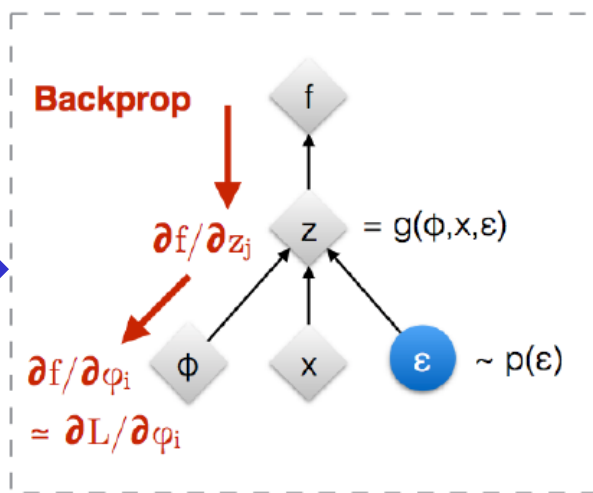


$$\mathcal{L}(x, \phi, \theta) = E_{q(\epsilon)}[f_{\phi, \theta}(x, z)] \approx \frac{1}{L} \sum_{i=1}^L f_{\phi, \theta}(x, g_{\phi}(\epsilon^{(i)}, x)), \quad \epsilon^{(i)} \sim q(\epsilon)$$

Original form



Reparameterised form



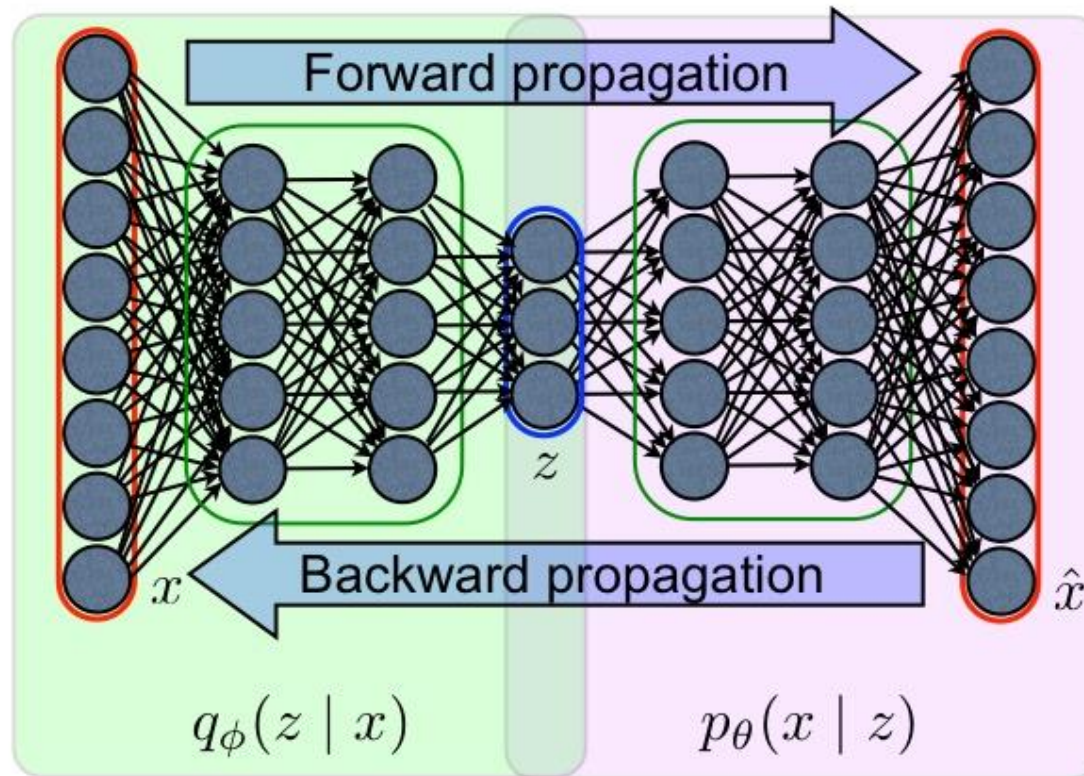
◊ : Deterministic node

● : Random node

[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

Training with Backpropagation

- Due to **reparametrization** trick, we can simultaneously train both the **generative model** and the **inference model** by optimizing the variational bound using the gradient **backpropagation**.



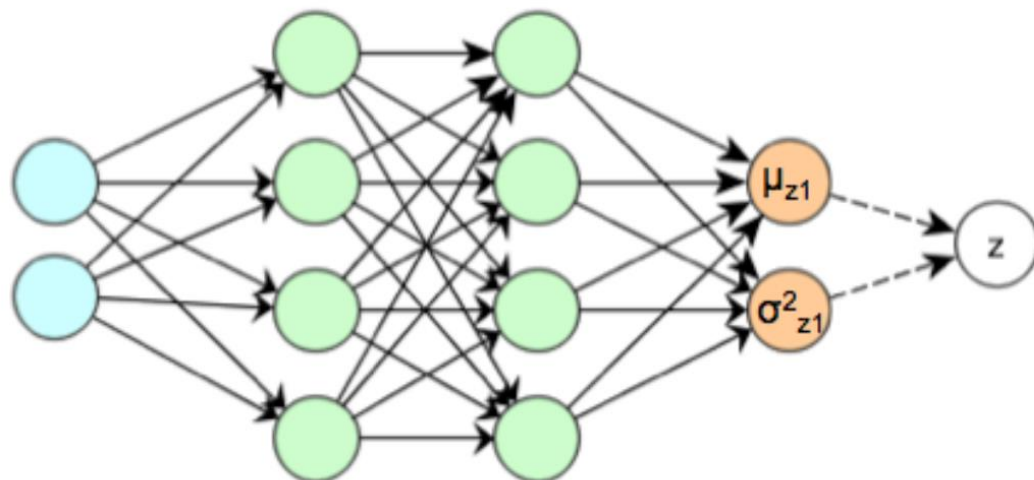
1D Gaussian Case

- We can compute the KL regularization in close form

Use $N(0,1)$ as prior for $p(z)$

$q(z|x^{(i)})$ is Gaussian with parameters $(\mu^{(i)}, \sigma^{(i)})$ determined by NN

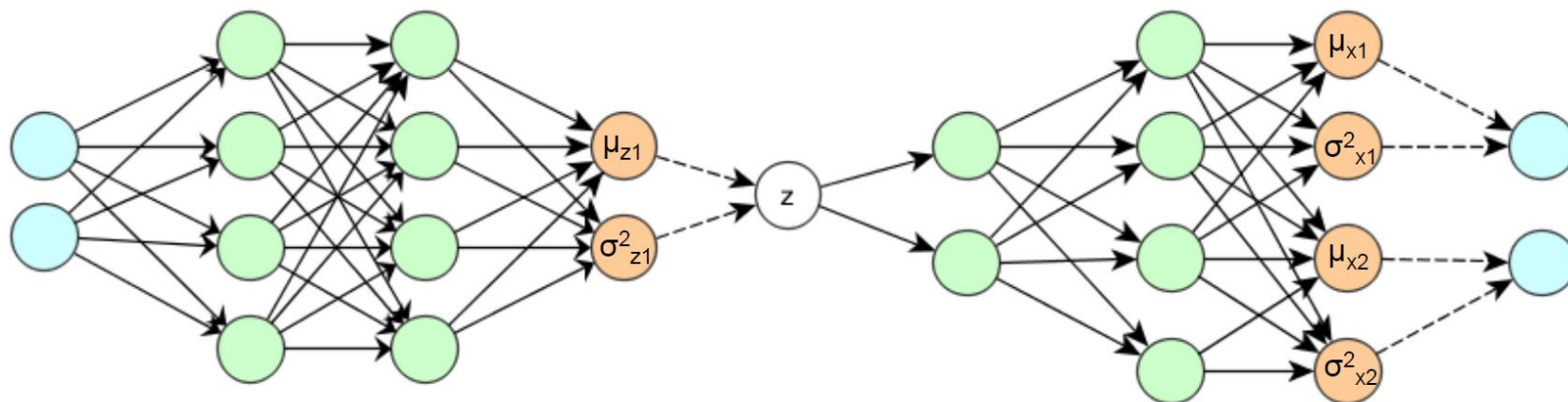
$$-D_{\text{KL}}(q(z|x^{(i)})||p(z)) = \frac{1}{2} \sum_{j=1}^J \left(1 + \log(\sigma_{z_j}^{(i)^2}) - \mu_{z_j}^{(i)^2} - \sigma_{z_j}^{(i)^2} \right)$$



1D Gaussian Case

Overall loss function for BP

Prior $p(z) \sim N(0,1)$ and p, q Gaussian, extension to $\dim(z) > 1$ trivial



Cost: Regularisation

$$-D_{\text{KL}}(q(z|x^{(i)})||p(z)) = \frac{1}{2} \sum_{j=1}^J \left(1 + \log(\sigma_{z_j}^{(i)^2}) - \mu_{z_j}^{(i)^2} - \sigma_{z_j}^{(i)^2} \right)$$

We use mini batch gradient decent to optimize the cost function over all $x^{(i)}$ in the mini batch

Cost: Reproduction

$$-\log(p(x^{(i)}|z^{(i)})) = \sum_{j=1}^D \frac{1}{2} \log(\sigma_{x_j}^2) + \frac{(x_j^{(i)} - \mu_{x_j})^2}{2\sigma_{x_j}^2}$$

Least Square for constant variance

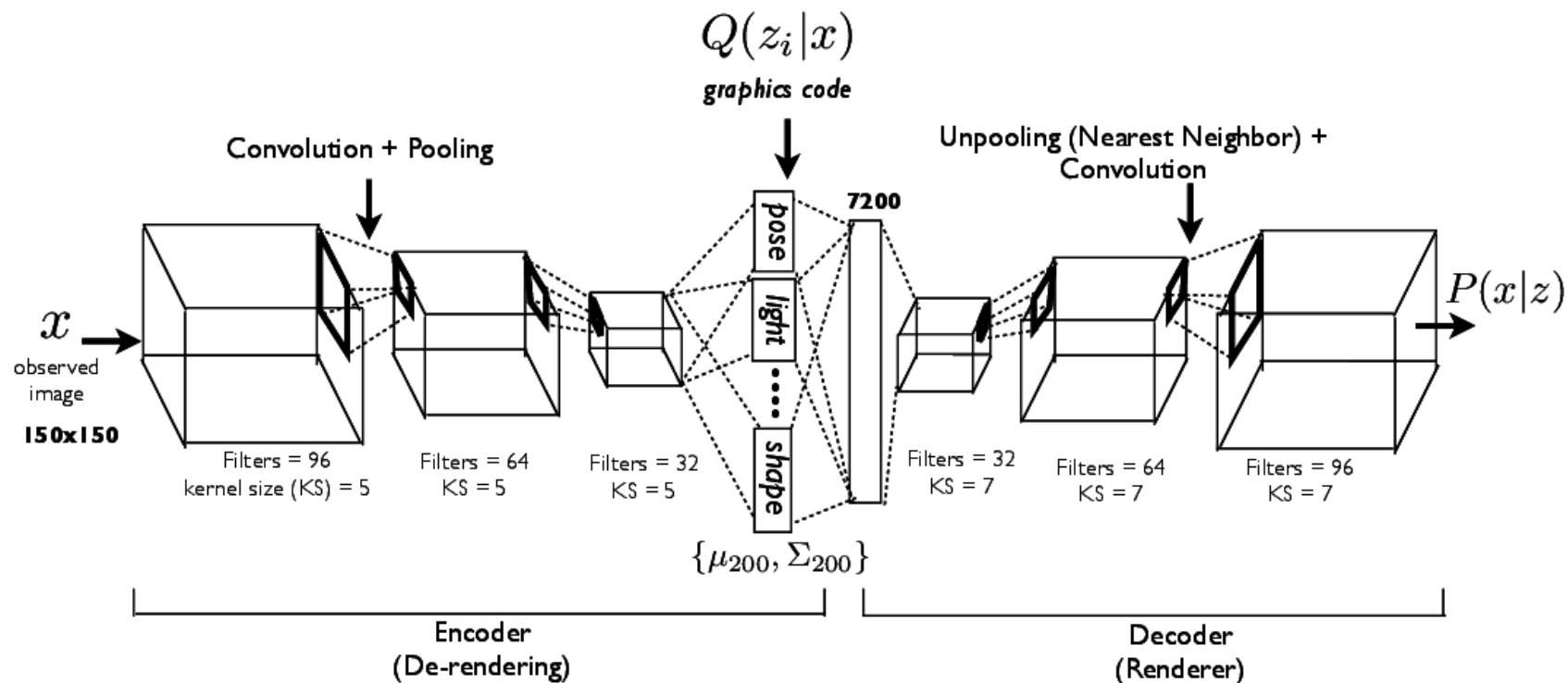
Interpreting the latent space



<https://arxiv.org/pdf/1610.00291.pdf>

Vision task I – Inverse graphics network

■ Main ideas



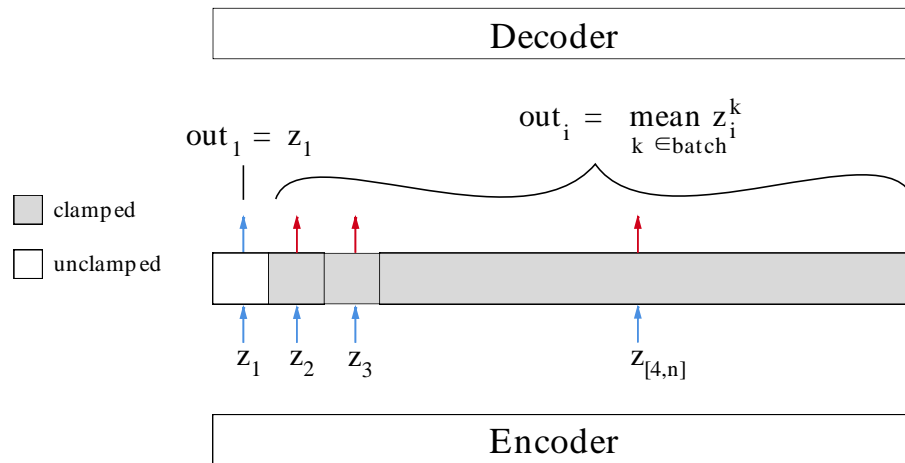
$$Z = \begin{bmatrix} z_1 & z_2 & z_3 & z_{[4,n]} \end{bmatrix}$$

corresponds to $\phi \quad \alpha \quad \phi_L$ intrinsic properties (shape, texture, etc)

Vision task I – Inverse graphics network

■ Mini-batch training

Forward



Backward

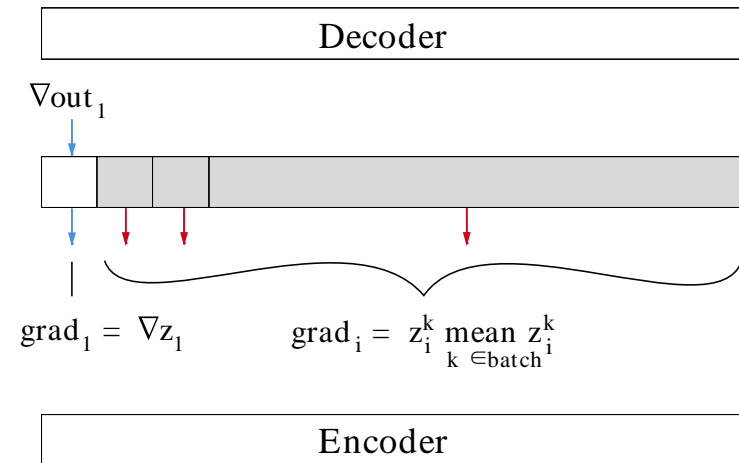
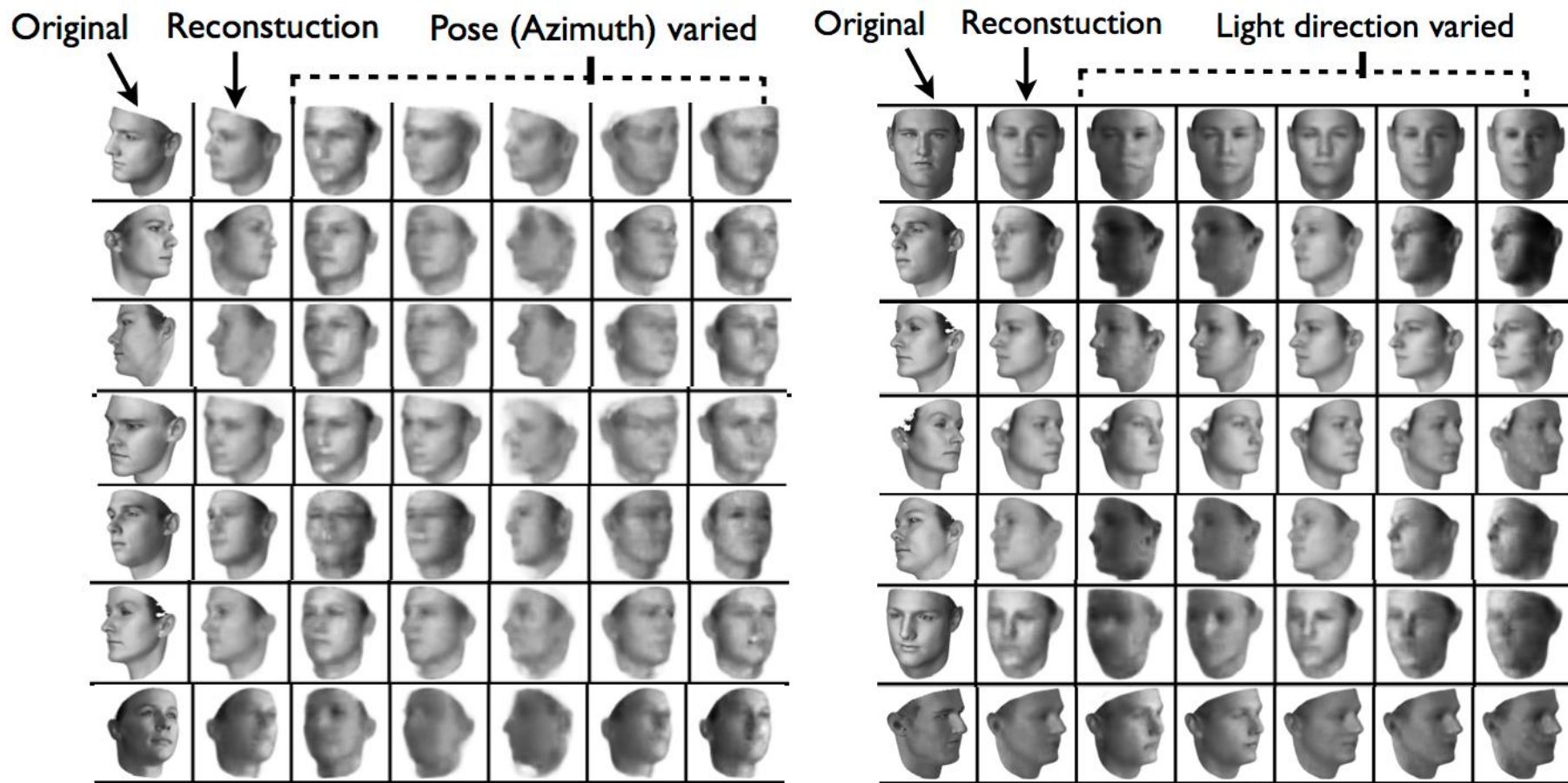


Figure 3: **Training on a minibatch in which only ϕ , the azimuth angle of the face, changes.** During the forward step, the output from each component $z_i \neq z_1$ of the encoder is altered to be the same for each sample in the batch. This reflects the fact that the generating variables of the image (e.g. the identity of the face) which correspond to the desired values of these latents are unchanged throughout the batch. By holding these outputs constant throughout the batch, the single neuron z_1 is forced to explain all the variance within the batch, i.e. the full range of changes to the image caused by changing ϕ . During the backward step z_1 is the only neuron which receives a gradient signal from the attempted reconstruction, and all $z_i \neq z_1$ receive a signal which nudges them to be closer to their respective averages over the batch. During the complete training process, after this batch, another batch is selected at random; it likewise contains variations of only one of $\phi, \alpha, \phi_L, \text{intrinsic}$; all neurons which do not correspond to the selected latent are clamped; and the training proceeds.

Vision task I – Inverse graphics network

■ Results



Vision task I – Inverse graphics network

- MoFA: Model-based Face Autoencoder (ICCV2017)

International Conference
on Computer Vision



Venice, Italy
October 22-29, 2017

MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction

(Oral Presentation)

[with voice-over]

Ayush Tewari¹, Michael Zollhöfer¹, Hyeongwoo Kim¹, Pablo Garrido¹,
Florian Bernard^{1,2}, Patrick Pérez³, Christian Theobalt¹

¹Max Planck Institute for Informatics

²LCSB, University of Luxembourg

³Technicolor



Vision task I – Inverse graphics network

- FML: Face model learning from videos (CVPR2019)

FML: Face Model Learning from Videos

(includes audio)



Ayush Tewari¹, Florian Bernard¹, Pablo Garrido², Gaurav Bharaj², Mohamed Elgharib¹,
Hans-Peter Seidel¹, Patrick Pérez³, Michael Zollhöfer⁴, Christian Theobalt¹

¹MPI Informatics, Saarland Informatics Campus ²Technicolor ³Valeo.ai ⁴Stanford University

Vision task I – Inverse graphics network

- DVP: Deep Video Portraits (SIGGRAPH 2018)

Deep Video Portraits

Hyeonwoo Kim¹ Pablo Garrido² Ayush Tewari¹ Weipeng Xu¹ Justus Thies³
Matthias Nießner³ Patrick Pérez² Christian Richardt⁴ Michael Zollhöfer⁵ Christian Theobalt¹

¹ *MPI Informatics*



² *Technicolor*



³ *Technical University of Munich*



⁴ *University of Bath*



⁵ *Stanford University*



Vision task I – Inverse graphics network

- Live Speech Portraits (SIGGRAPH/ASA 2021)



Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation

Yuanxun Lu¹ Jinxiang Chai² Xun Cao¹

¹Nanjing University

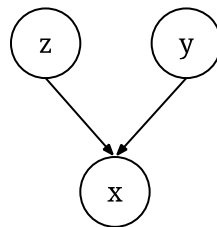
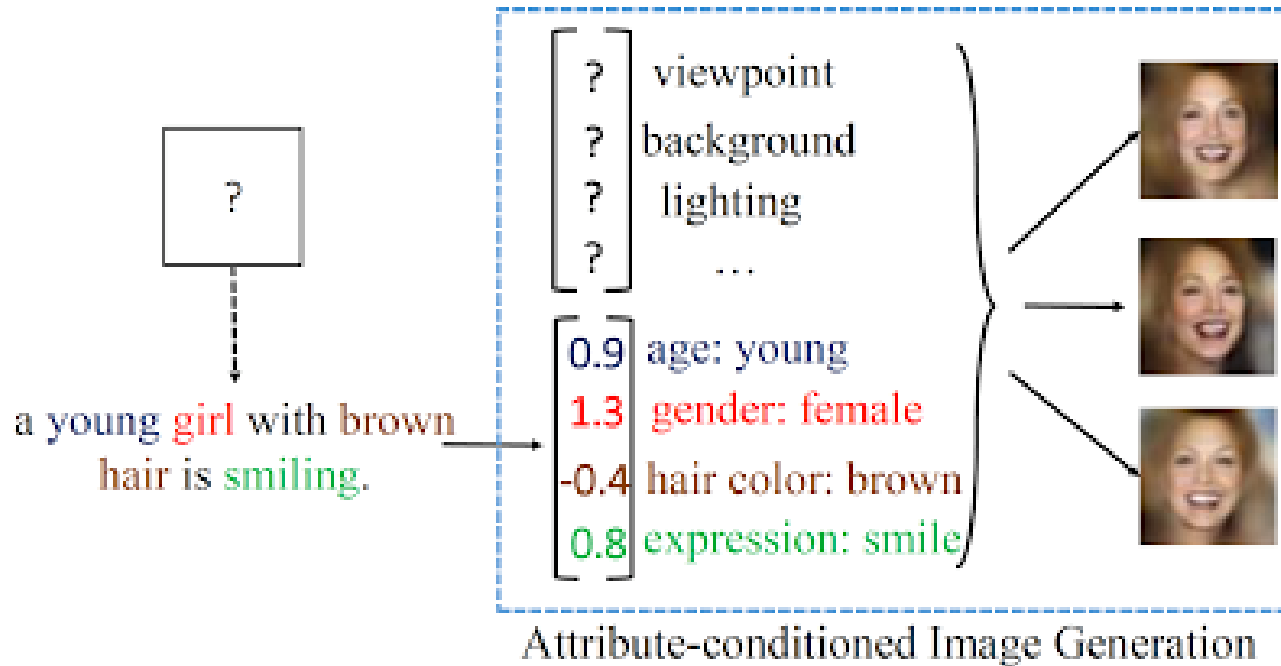
²Xmov



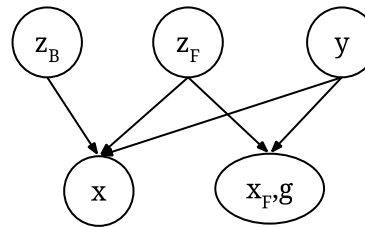
(with Audio)

Vision task II – Attribute2Image

■ Main ideas



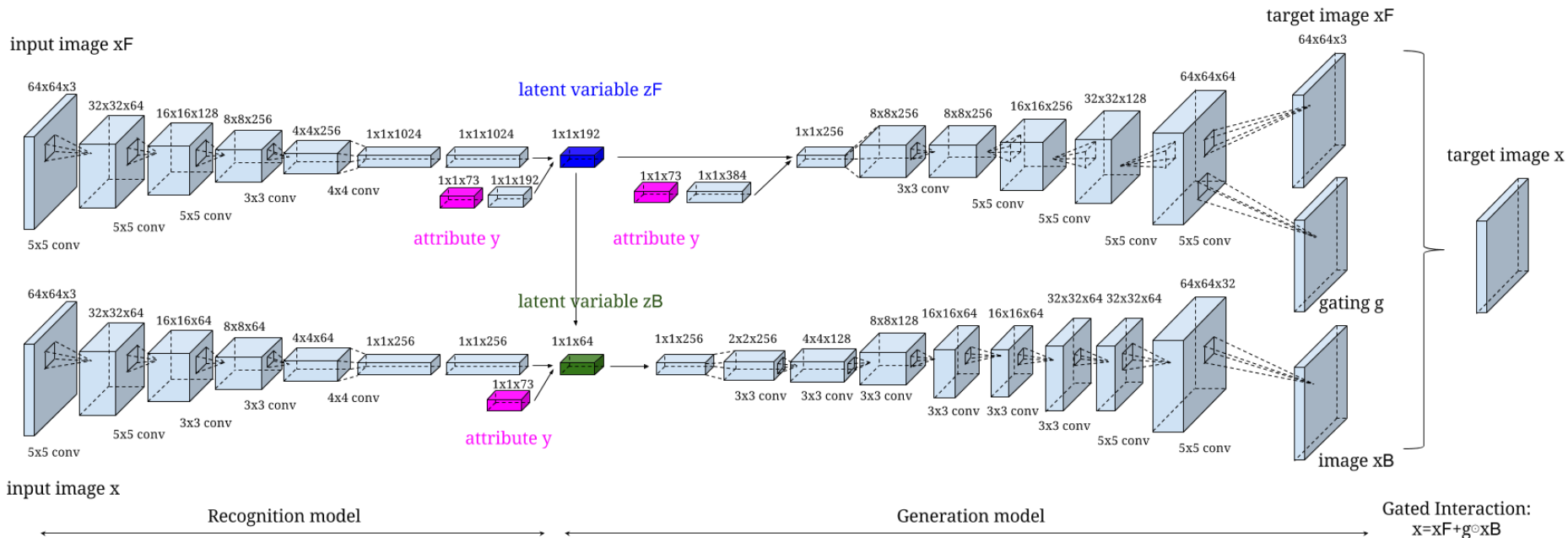
(a) CVAE: $p_{\theta}(x|y, z)$



(b) disCVAE: $p_{\theta}(x, x_F, g|y, z_F, z_B)$

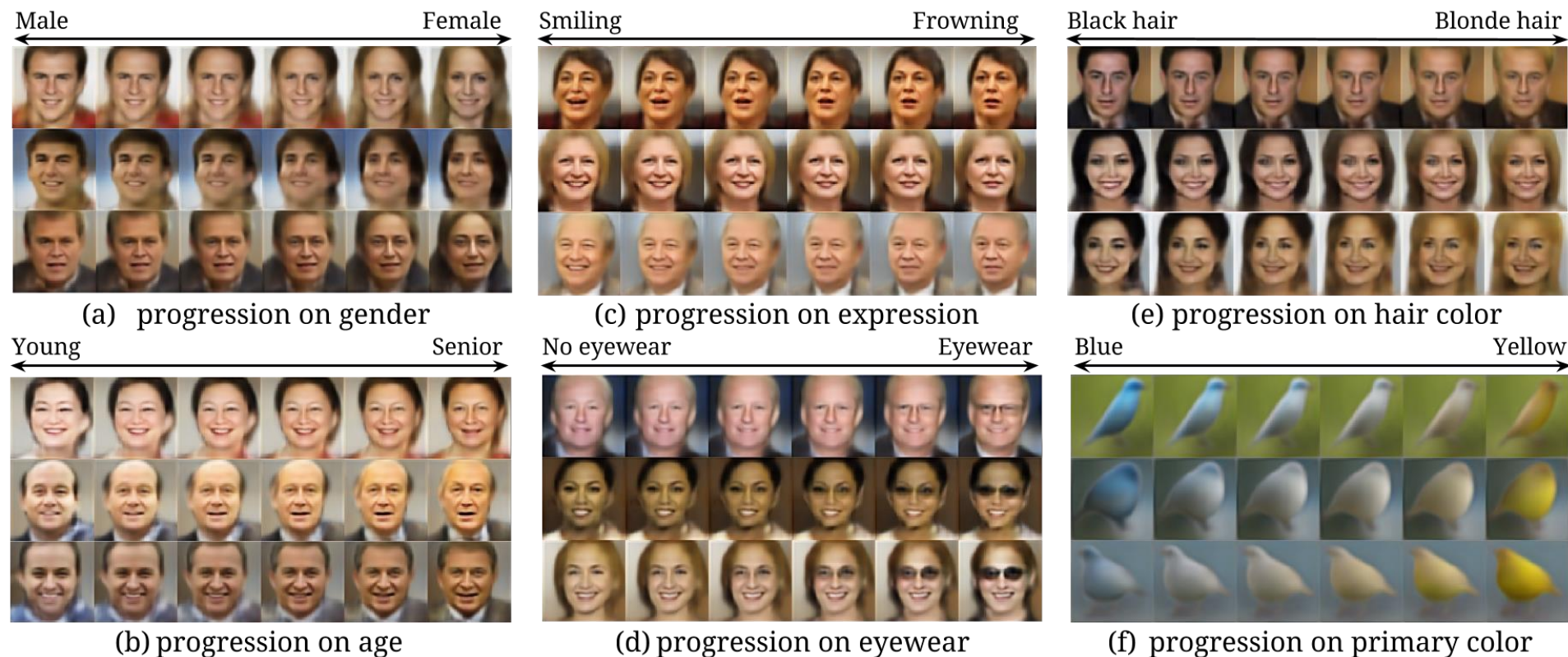
Vision task II – Attribute2Image

■ Network structure



Vision task II – Attribute2Image

■ Results



Problems of VAE

■ Model capacity

- Note that the VAE requires 2 tractable distributions to be used:
 - The prior distribution $p(z)$ must be easy to sample from
 - The conditional likelihood $p(x|z, \theta)$ must be computable
- In practice this means that the 2 distributions of interest are often simple, for example uniform, Gaussian, or even isotropic Gaussian

Problems of VAE

■ Blurry images



<https://blog.openai.com/generative-models/>

- The samples from the VAE look blurry
- Three plausible explanations for this
 - Maximizing the likelihood
 - Restrictions on the family of distributions
 - The lower bound approximation

Problems of VAE

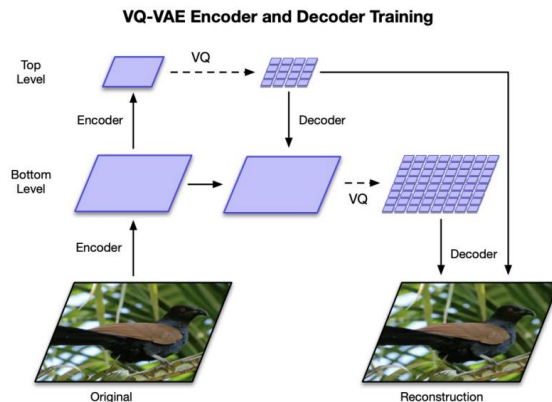
■ Blurry images

- Recent investigations suggest that both the simple probability distributions and the variational approximation lead to blurry images
- [Kingma & colleagues: Improving Variational Inference with Inverse Autoregressive Flow](#)
- [Zhao & colleagues: Towards a Deeper Understanding of Variational Autoencoding Models](#)
- [Nowozin & colleagues: f-gan: Training generative neural samplers using variational divergence minimization](#)

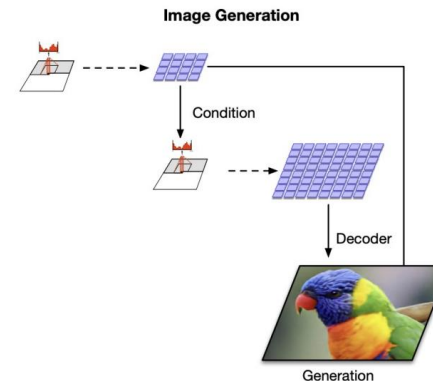
VQ-VAE

- Vector-Quantized Variational Autoencoder
- Combining VAE + Autoregressive: VQ-VAE1,2

Train a VAE-like model to generate multiscale grids of latent codes



Use a multiscale PixelCNN to sample in latent code space



Aaron van den Oord et al, "Neural Discrete Representation Learning ", NeurIPS 2017
Razavi et al, "Generating Diverse High-Fidelity Images with VQ-VAE-2", NeurIPS 2019

VQ-VAE

- 1024 x 1024 generated faces, trained on FFHQ

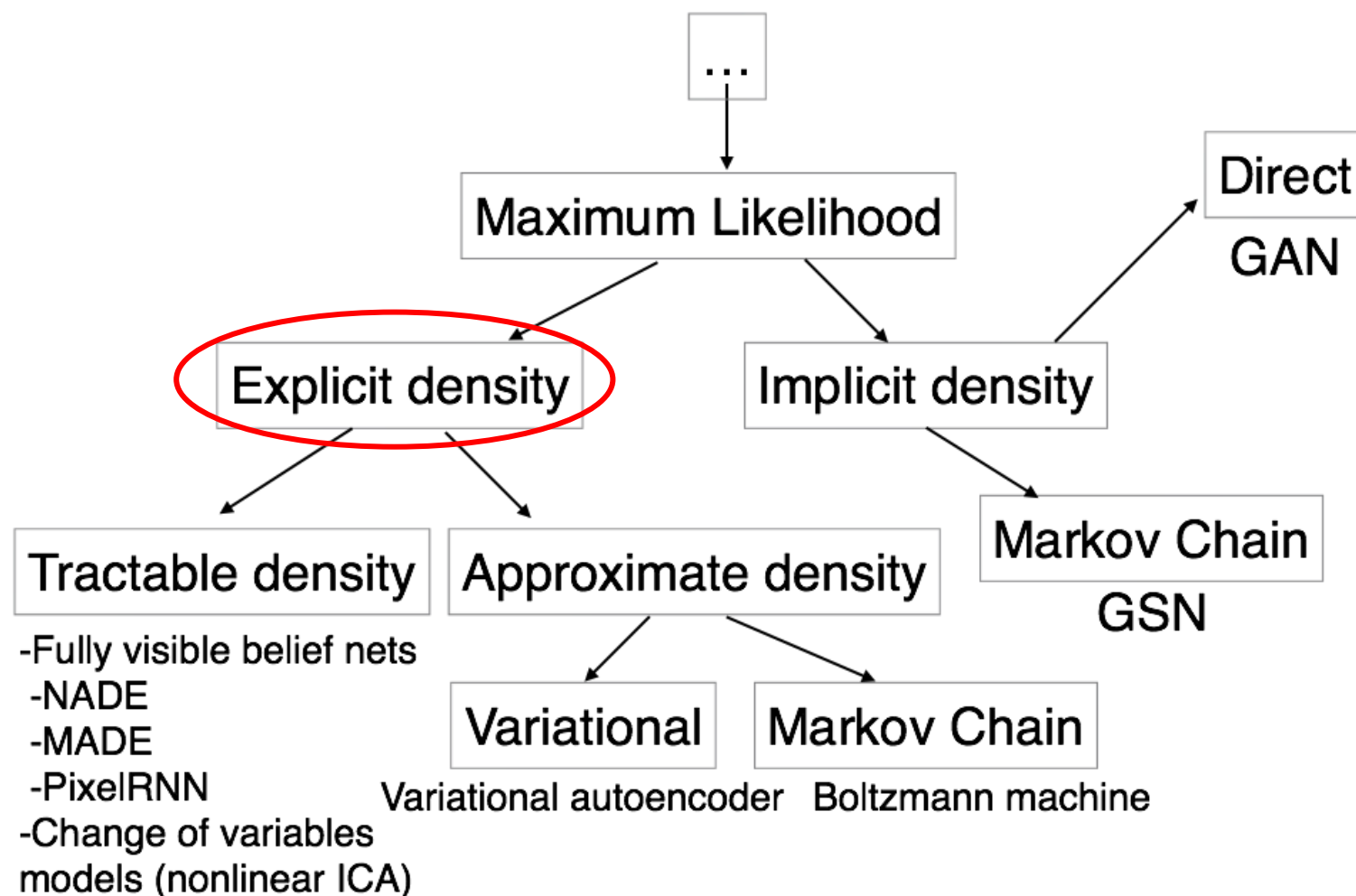


Outline

- Vision applications of VAEs
 - Inverse graphics network
 - Attribute2Image
- Generative Adversarial Networks
 - Implicit generative models
 - Adversarial learning

Acknowledgement: Feifei Li et al's cs231n notes

Taxonomy of Generative Models

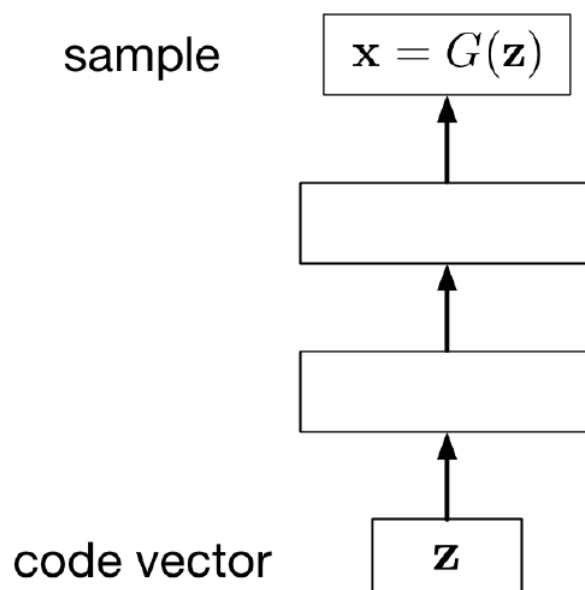


Implicit Generative Models

- Working with explicit model $p(x)$ could be expensive
 - Variational Autoencoder (variational inference)
 - Boltzmann Machines (MCMC)
- Representation learning may not require $p(x)$
 - Sometimes we are more interested in taking samples from $p(x)$ instead of p itself

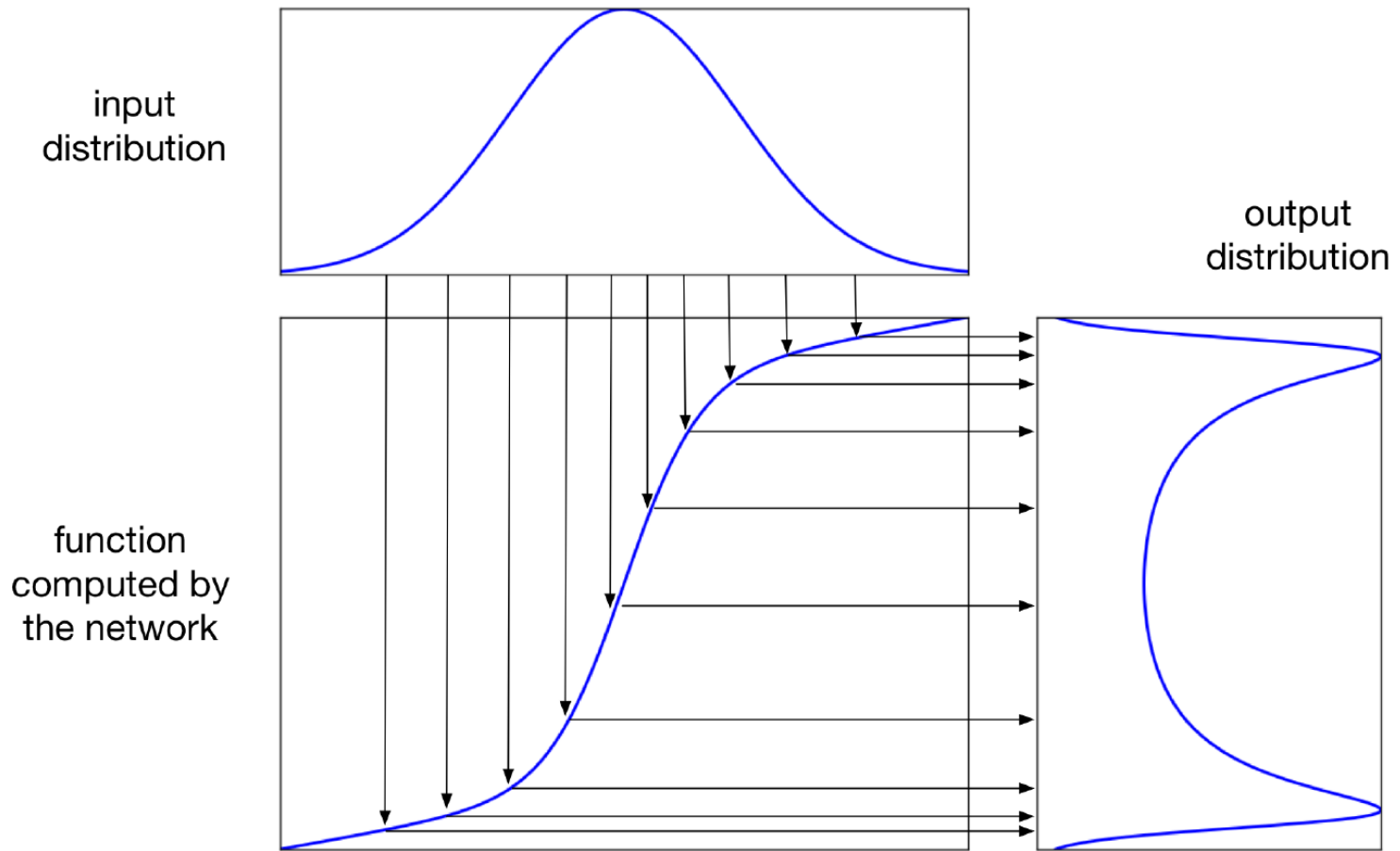
Implicit Generative Models

- Implicitly define a probability distribution
- Start by sampling the code vector z from a fixed, simple distribution
- A generator network computes a differentiable function G mapping z to an x in data space



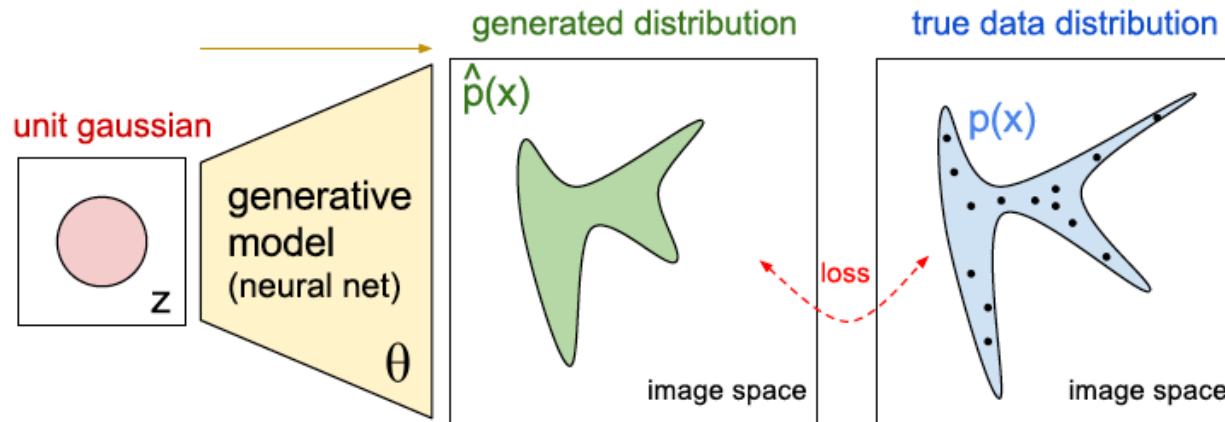
Implicit Generative Models

■ Intuition: 1D example



Implicit Generative Models

■ Intuition

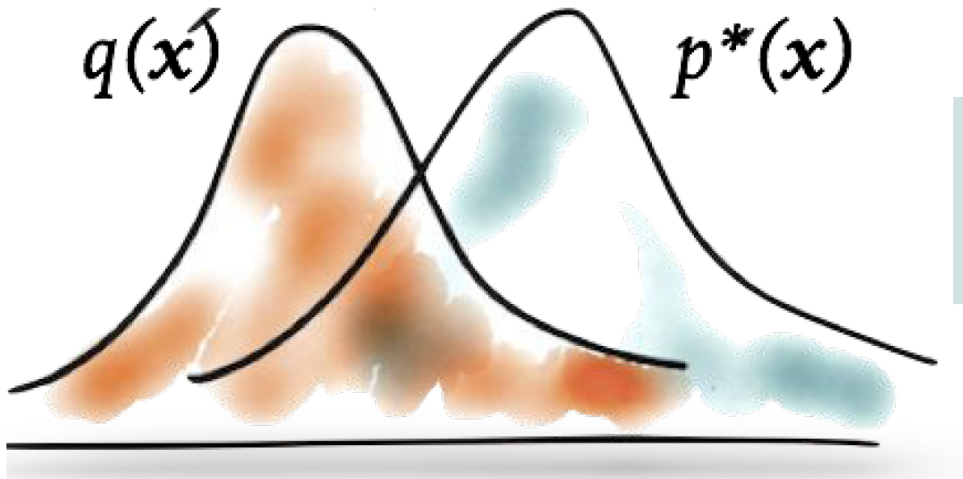
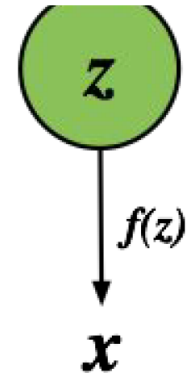


advocate/penalize samples within the blue/white region.

Learning by comparison

■ Basic idea

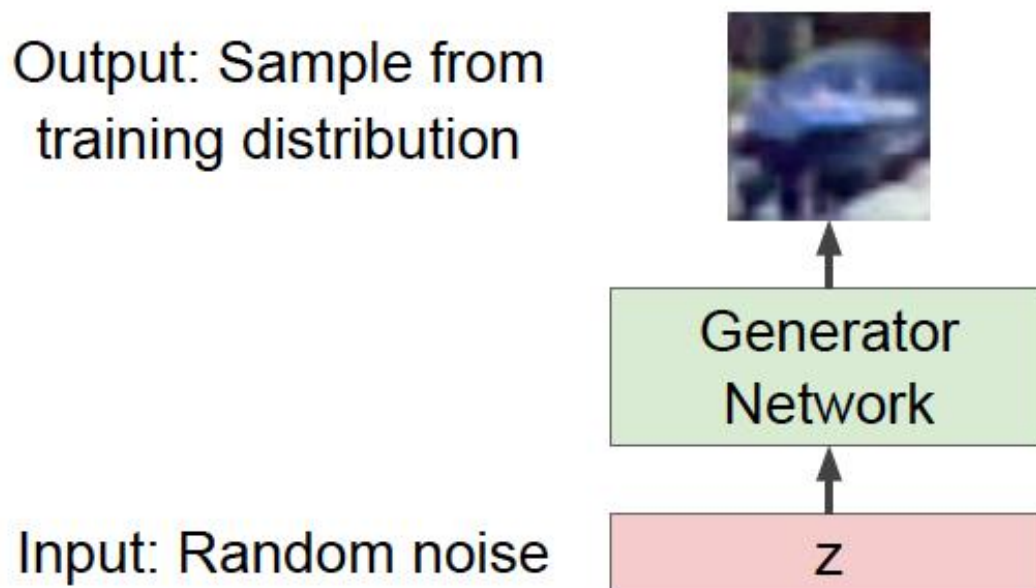
For some models, we only have access to an unnormalised probability, partial knowledge of the distribution, or a simulator of data.



We compare the estimated distribution $q(x)$ to the true distribution $p^*(x)$ using samples.

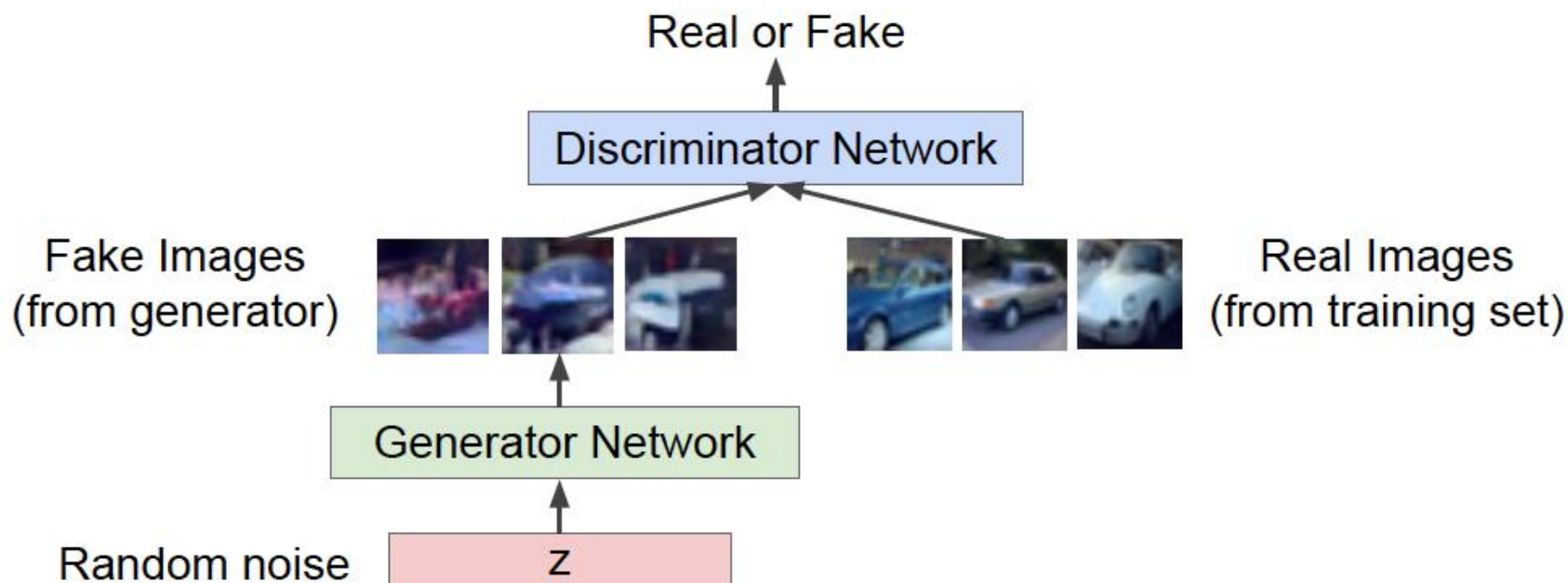
Generative Adversarial Networks

- Using a neural network to generate data



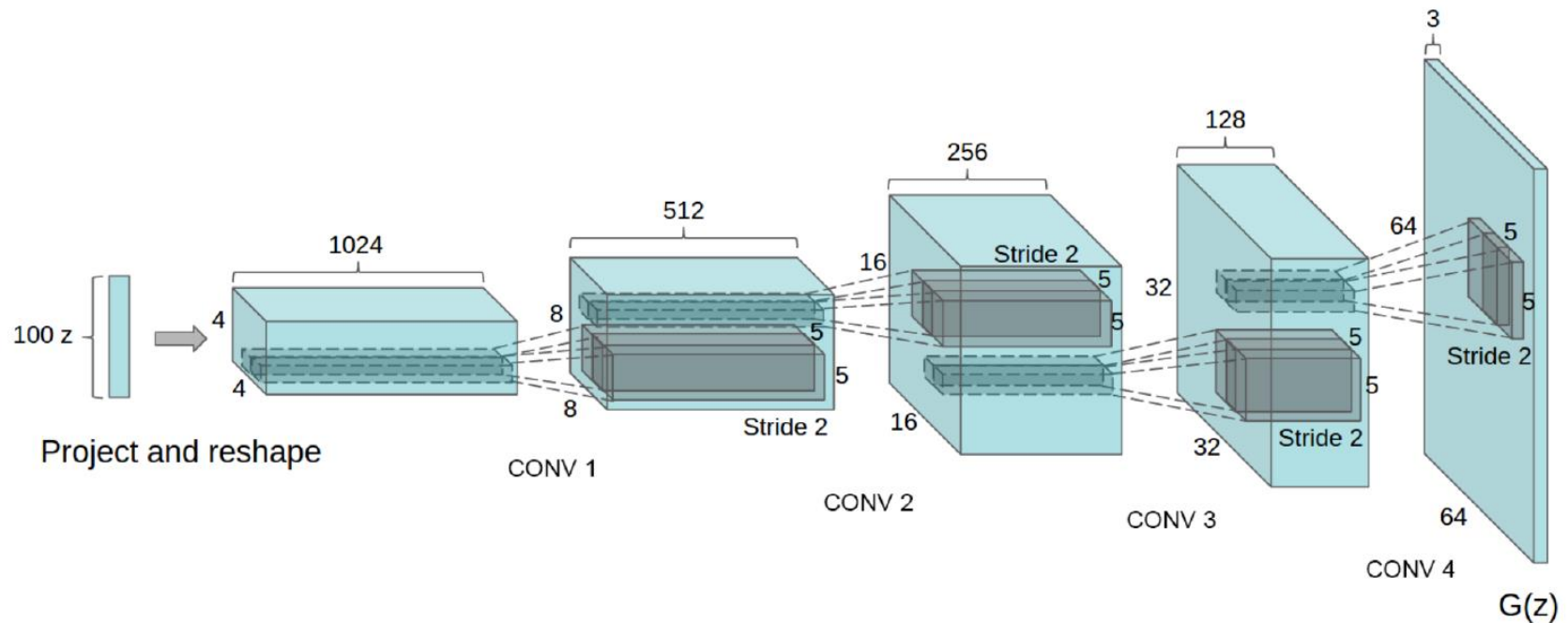
Generative Adversarial Networks

- Using another neural network to determine if the data is real or not



Typical generator architecture

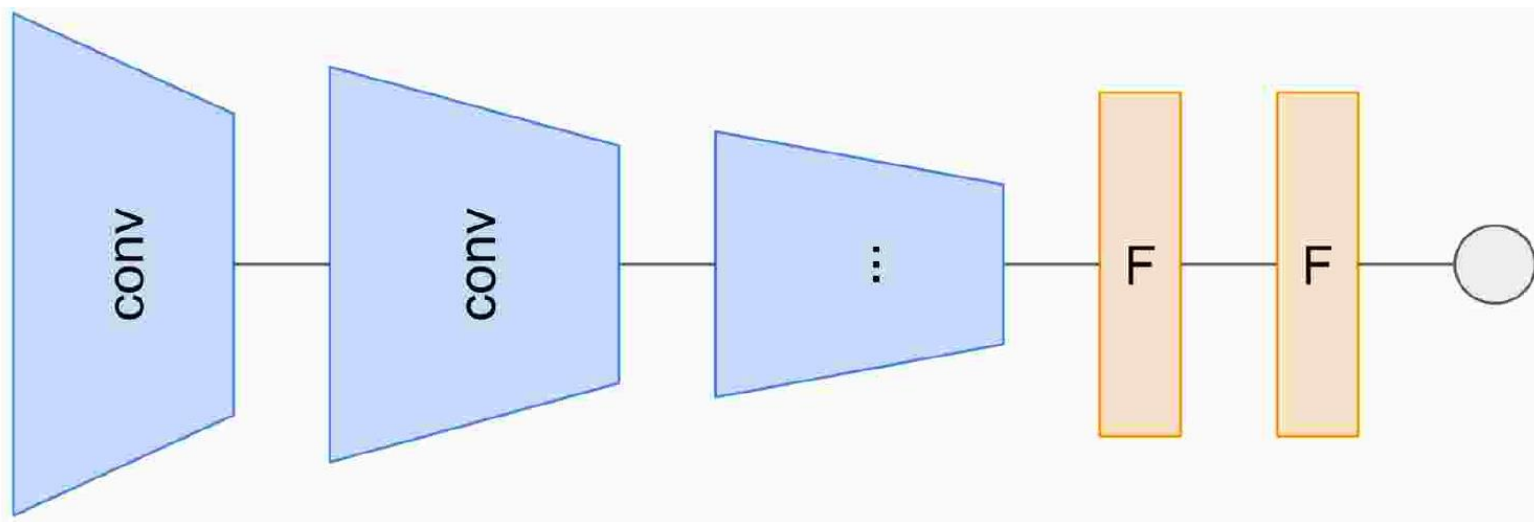
- For images



- ▶ Unit Gaussian distribution on z , typically 10-100 dim.
- ▶ Up-convolutional deep network (reverse recognition CNN)

Typical discriminator architecture

- For images



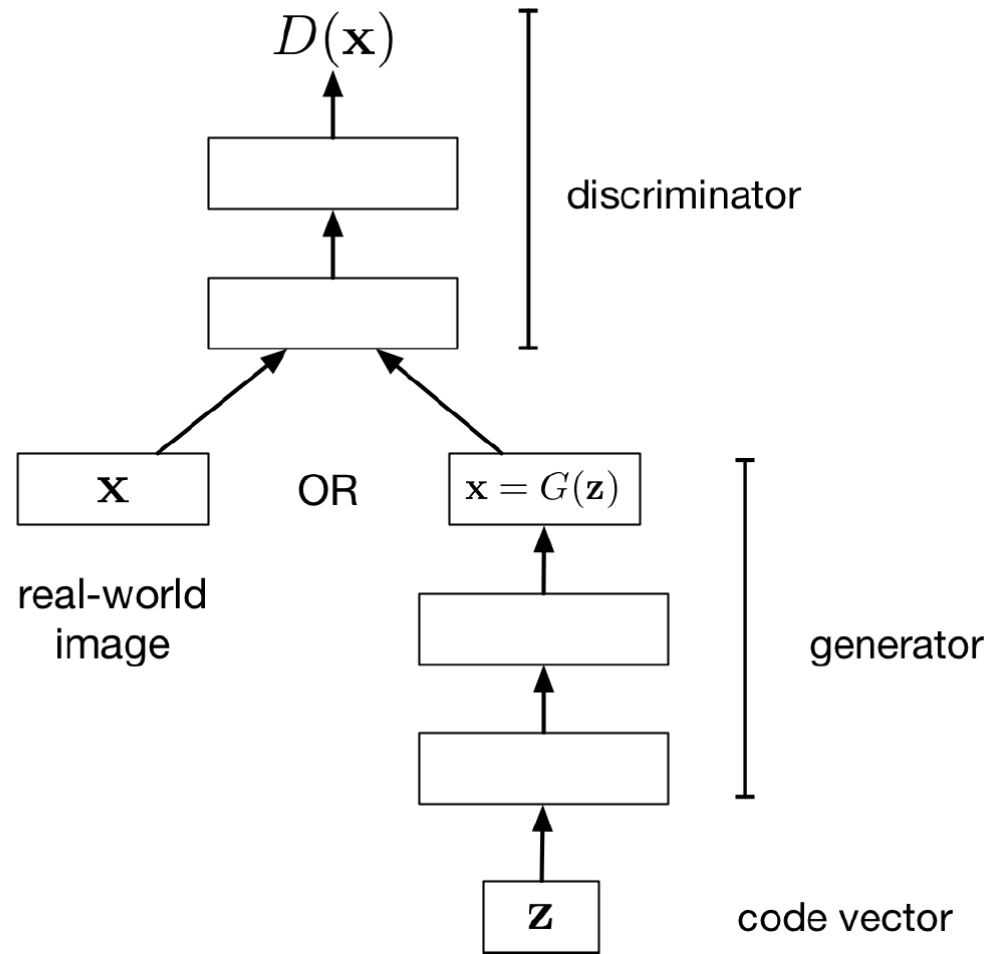
- ▶ Recognition CNN model
- ▶ Binary classification output: real / synthetic

Adversarial learning

- GAN objective for the generator is some complicated objective function defined by a neural network.
 - This means a new way of thinking about "distance".
 - We are training networks to minimize the "distance" or "divergence" between generated images and real images.
 - Instead of some hand-crafted distance metric like L1 or L2, we can make something completely new.
 - A neural network, with the right architecture, is arguably the definition of perceptual similarity (assuming our visual system is some sort of neural network).

Adversarial Learning

- Adversarial loss



Adversarial Learning

- Let D denote the discriminator's predicted probability of being real data
- Discriminator's cost function: cross-entropy loss for task of classifying real vs. fake images

$$\mathcal{J}_D = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[-\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[-\log(1 - D(G(\mathbf{z})))]$$

- One possible cost function for the generator: the opposite of the discriminator's

$$\begin{aligned}\mathcal{J}_G &= -\mathcal{J}_D \\ &= \text{const} + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))]\end{aligned}$$

Two-player game

■ Minimax formulation

- The generator and discriminator are playing a zero-sum game against each other

$$\min_G \max_D \mathcal{J}_D$$

- Using parametric models

Discriminator outputs likelihood in (0,1) of real image

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log \underbrace{D_{\theta_d}(x)}_{\substack{\text{Discriminator output} \\ \text{for real data } x}} + \mathbb{E}_{z \sim p(z)} \log(1 - \underbrace{D_{\theta_d}(G_{\theta_g}(z))}_{\substack{\text{Discriminator output for} \\ \text{generated fake data } G(z)}}) \right]$$

Summary

- Variational Autoencoders (VAEs)
 - VAE objective
 - VAE: Vision applications
- Generative Adversarial Networks (GAN)
 - Adversarial learning
- Next time:
 - GAN
- Quiz:
 - If online, send it to zhaochf2022@shanghaitech.edu.cn
- Reading material
 - <https://arxiv.org/pdf/1503.03167.pdf> (Deep Convolutional Inverse Graphics Network)
 - <https://arxiv.org/pdf/2011.10063.pdf> (Dual Contradistinctive Generative Autoencoder)
 - <https://taesung.me/SwappingAutoencoder/> (Swapping Autoencoder for Deep Image Manipulation)