Variational Autoencoders (VAEs)

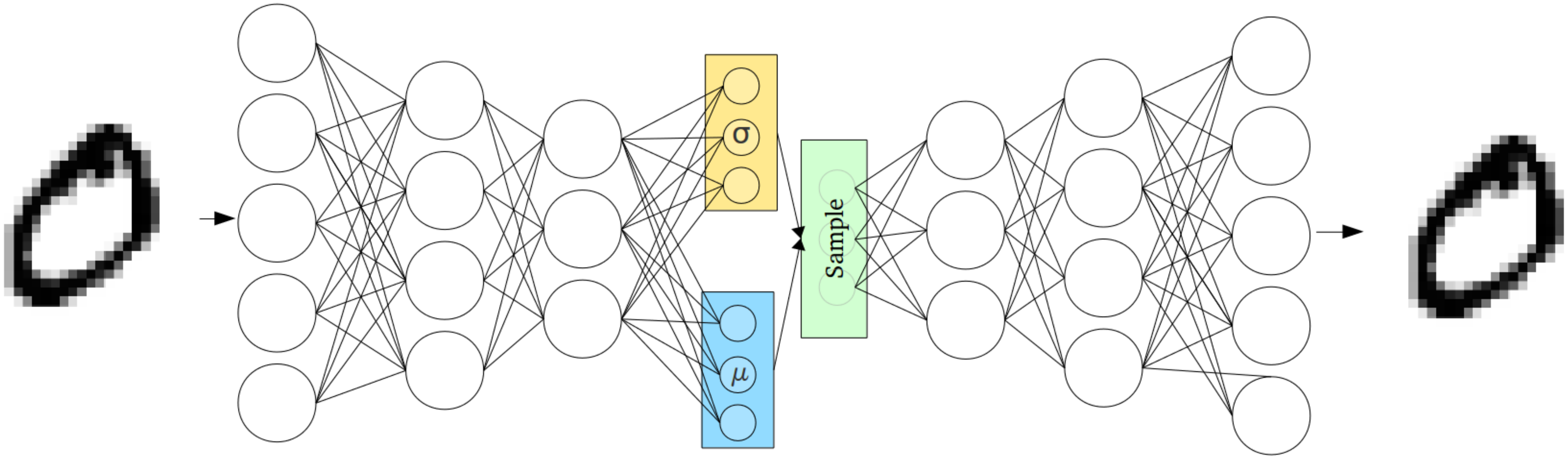


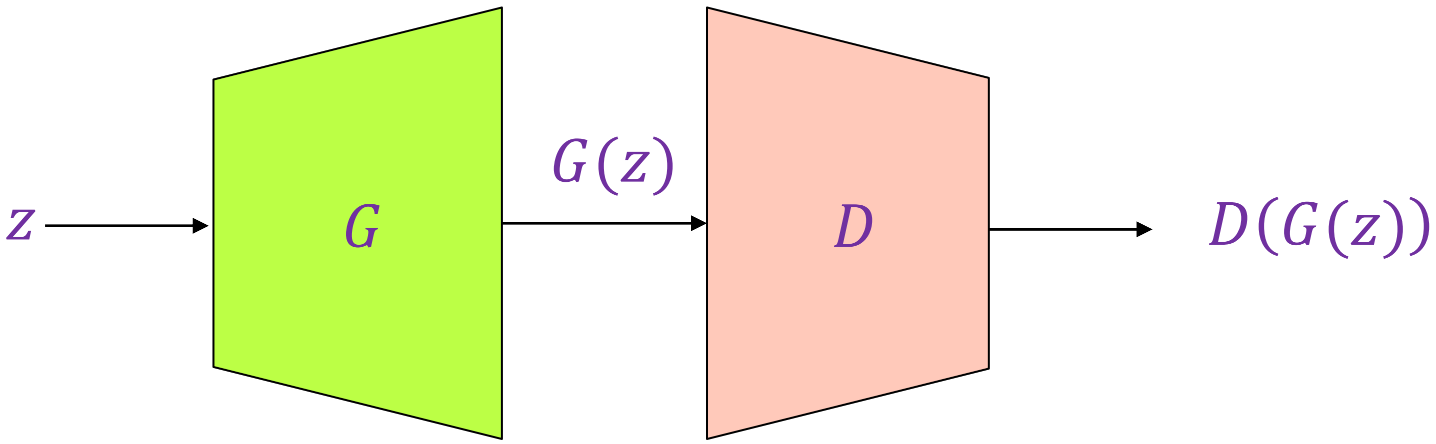
Image Source

Outline

- Basic VAE formulation
- Highlights of recent work

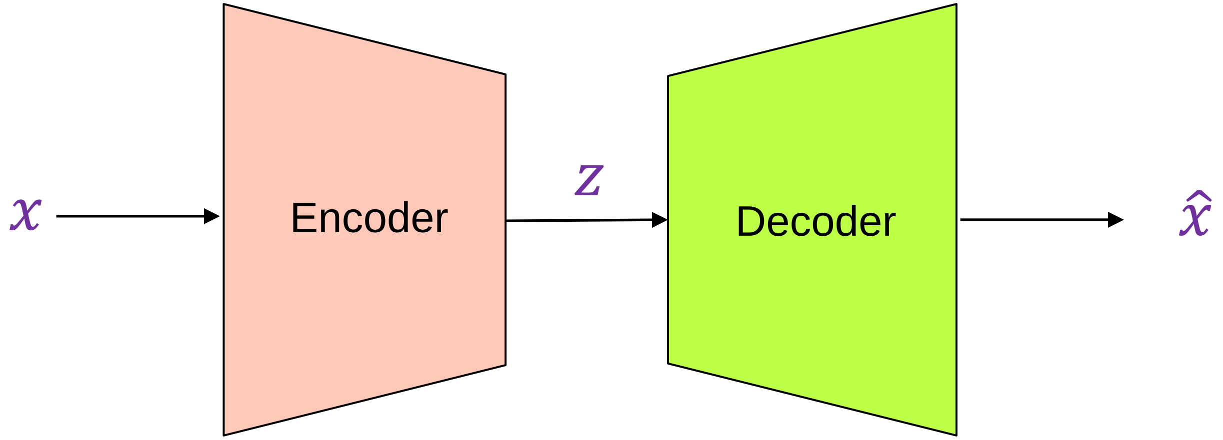
Recall: GANs

- Training:
  - Discriminator: low scores for fake data, high scores for real data
  - Generator: increase discriminator score on fake data
- Test time: discard discriminator and use generator to sample from learned distribution



Variational autoencoders: Overview

- Probabilistic formulation based on *variational Bayes* framework
- At training time, jointly learn *encoder* and *decoder* by maximizing (a bound on) the data likelihood
- At test time, discard encoder and use decoder to sample from the learned distribution



D. Kingma and M. Welling, Auto-Encoding Variational Bayes, ICLR 2014

Variational autoencoders: Overview

- Probabilistic generative model of the data distribution:

Latent code Data point

$$z \longrightarrow x$$

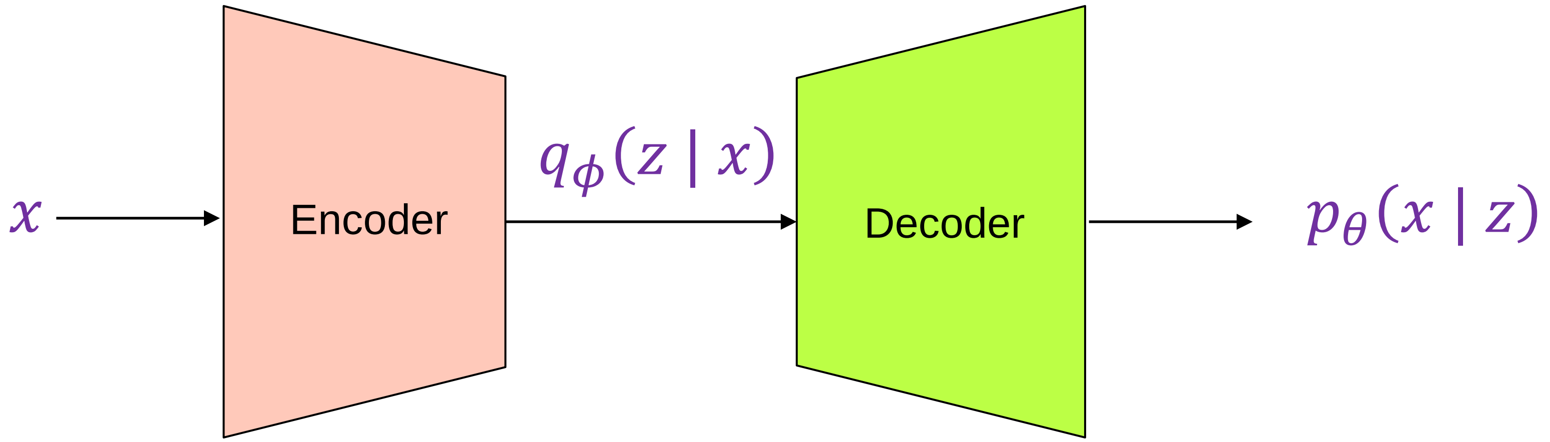
Sample z from *prior* $p(z)$
(usually Gaussian)

Sample x from
conditional $p(x|z)$

Try to approximate the
conditional with neural
network

Variational autoencoders: Overview

- At training time, jointly learn *encoder* and *decoder*
- **Encoder:** given inputs x, output $q_\phi(z \mid x)$
  - Specifically, output mean and (diagonal) covariance, or $\mu_{z|x}$ and $\Sigma_{z|x}$, so that $q_\phi(z \mid x) = N(\mu_{z|x}, \Sigma_{z|x})$
- **Decoder:** given z, output $p_\theta(x \mid z)$
  - Specifically, output $\mu_{x|z}$ and $\Sigma_{x|z}$ so that $p_\theta(x \mid z) = N(\mu_{x|z}, \Sigma_{x|z})$
- **Training objective:** (a bound on) data likelihood under the model



Variational autoencoders: Overview

- VAE makes its encoder not output an encoding vector of size n, rather, outputting two vectors of size n: a vector of means, μ, and another vector of standard deviations, σ.

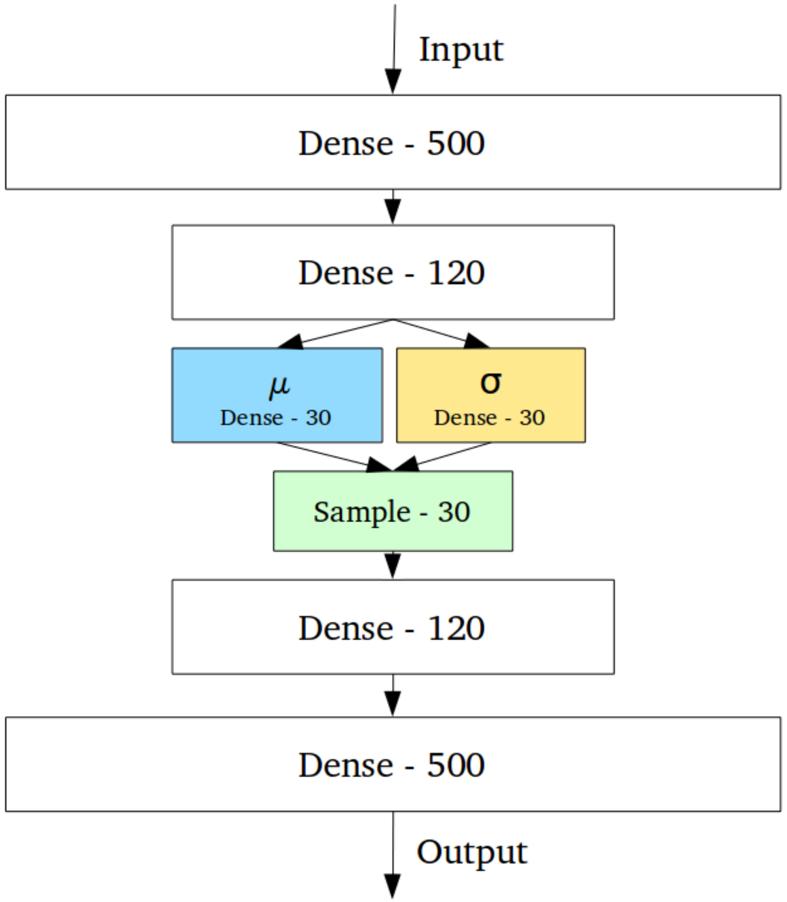


Image Source

Variational autoencoders: Overview

- VAE makes its encoder not output an encoding vector of size n, rather, outputting two vectors of size n: a vector of means, μ, and another vector of standard deviations, σ.

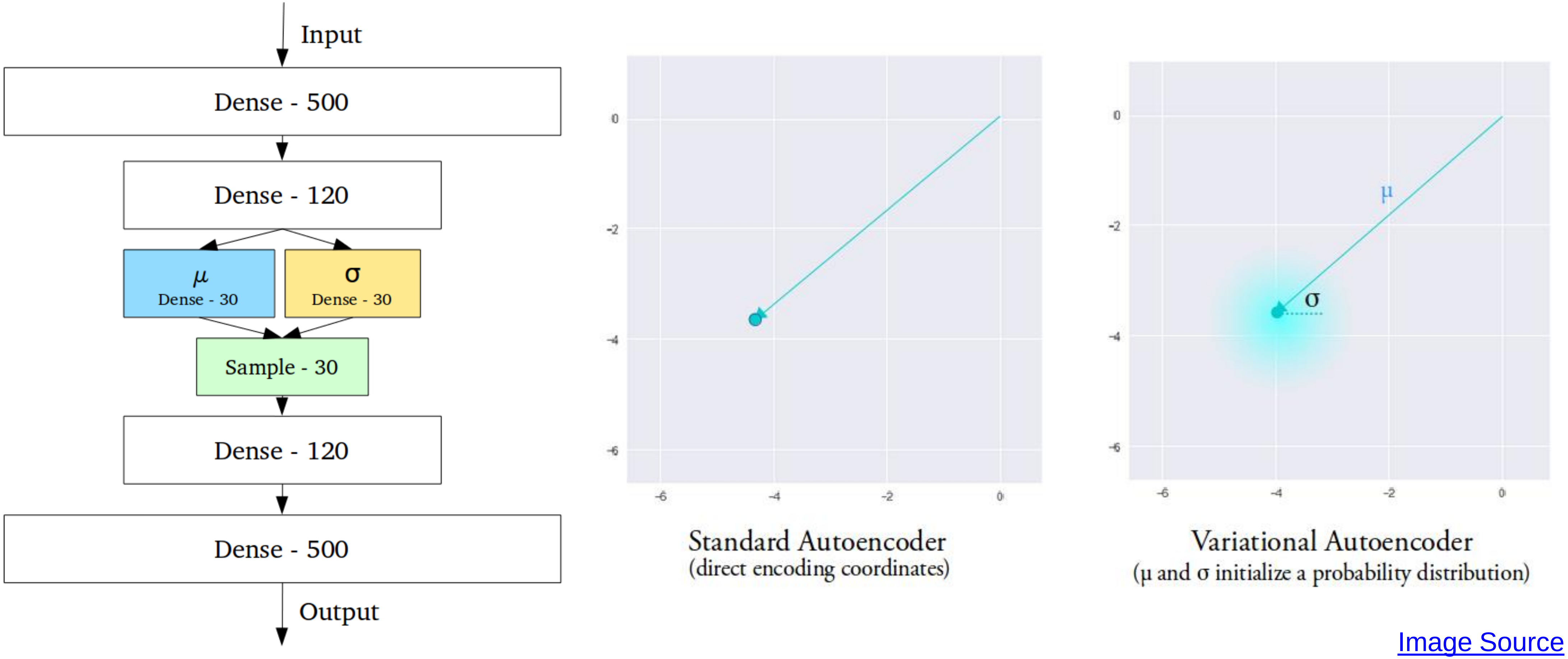


Image Source

Variational autoencoders: Overview

- VAE makes its encoder not output an encoding vector of size n, rather, outputting two vectors of size n: a vector of means, μ, and another vector of standard deviations, σ.

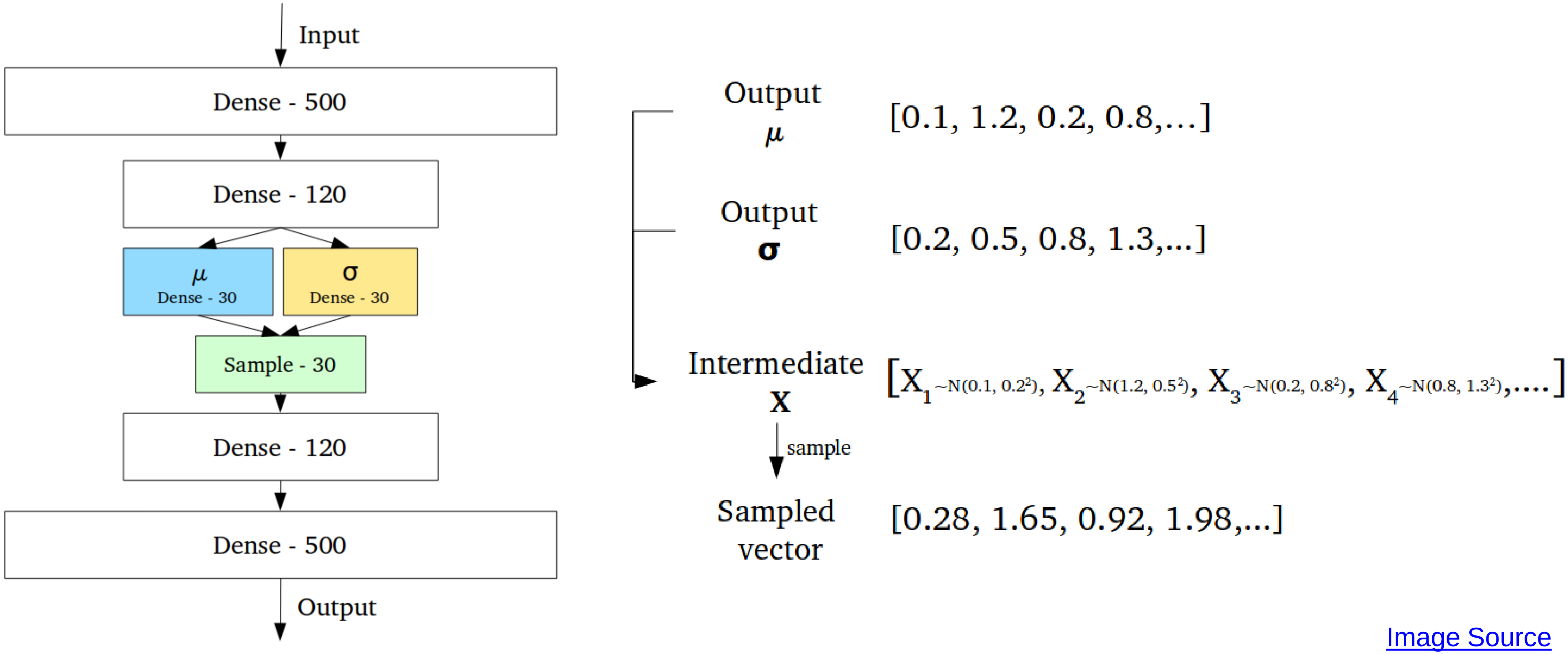


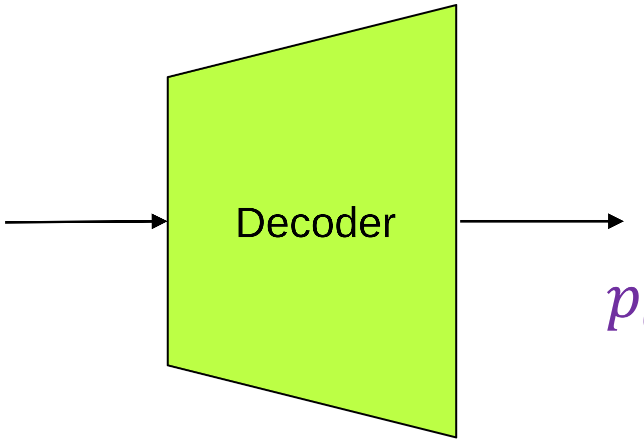
Image Source

Variational autoencoders: Overview

- At test time, discard encoder and use decoder to sample from $p_\theta(x \mid z) = N(\mu_{x|z}, \Sigma_{x|z})$

Sample z from prior (Gaussian)

z



Sample x from $p_\theta(x \mid z) = N(\mu_{x|z}, \Sigma_{x|z})$

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \;\geq\; \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] \;-\; D_{KL}\big(q_\phi(z|x), p(z)\big)$$

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \;\geq\; \mathbb{E}_{z \sim \boxed{q_\phi(z|x)}}[\log p_\theta(x|z)] \;-\; D_{KL}\left(\boxed{q_\phi(z|x)}, p(z)\right)$$

1. Run training point x through encoder to get a distribution over latent codes z

Adapted from J. Johnson

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] \boxed{- D_{KL}(q_\phi(z|x), p(z))}$$

1. Run training point x through encoder to get a distribution over latent codes z
2. Encoder output should match the prior $p(z)$

Adapted from J. Johnson

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \geq \mathbb{E}_{\boxed{z\sim q_\phi(z|x)}}[\log p_\theta(x|z)] - D_{KL}\left(q_\phi(z|x), p(z)\right)$$

1. Run training point x through encoder to get a distribution over latent codes z
2. Encoder output should match the prior $p(z)$
3. Sample code z from encoder output

Adapted from J. Johnson

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q_\phi(z|x)}[\log \boxed{p_\theta(x|z)}] - D_{KL}\big(q_\phi(z|x), p(z)\big)$$

1. Run training point x through encoder to get a distribution over latent codes z
2. Encoder output should match the prior $p(z)$
3. Sample code z from encoder output
4. Run sampled z through decoder to get a distribution over data samples

Adapted from J. Johnson

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q_\phi(z|x)}[\boxed{\log p_\theta(x|z)}] - D_{KL}\big(q_\phi(z|x), p(z)\big)$$

1. Run training point x through encoder to get a distribution over latent codes z
2. Encoder output should match the prior $p(z)$
3. Sample code z from encoder output
4. Run sampled z through decoder to get a distribution over data samples
5. Original input should be likely under the distribution output in (4)

Adapted from J. Johnson

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{Data likelihood}} - \underbrace{D_{KL}\big(q_\phi(z|x), p(z)\big)}_{\text{Regularization}}$$

Variational autoencoders: Training

- Objective: maximize the *variational lower bound* on the data likelihood:

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{Data likelihood}} - \underbrace{D_{KL}\big(q_\phi(z|x), p(z)\big)}_{\text{Regularization}}$$

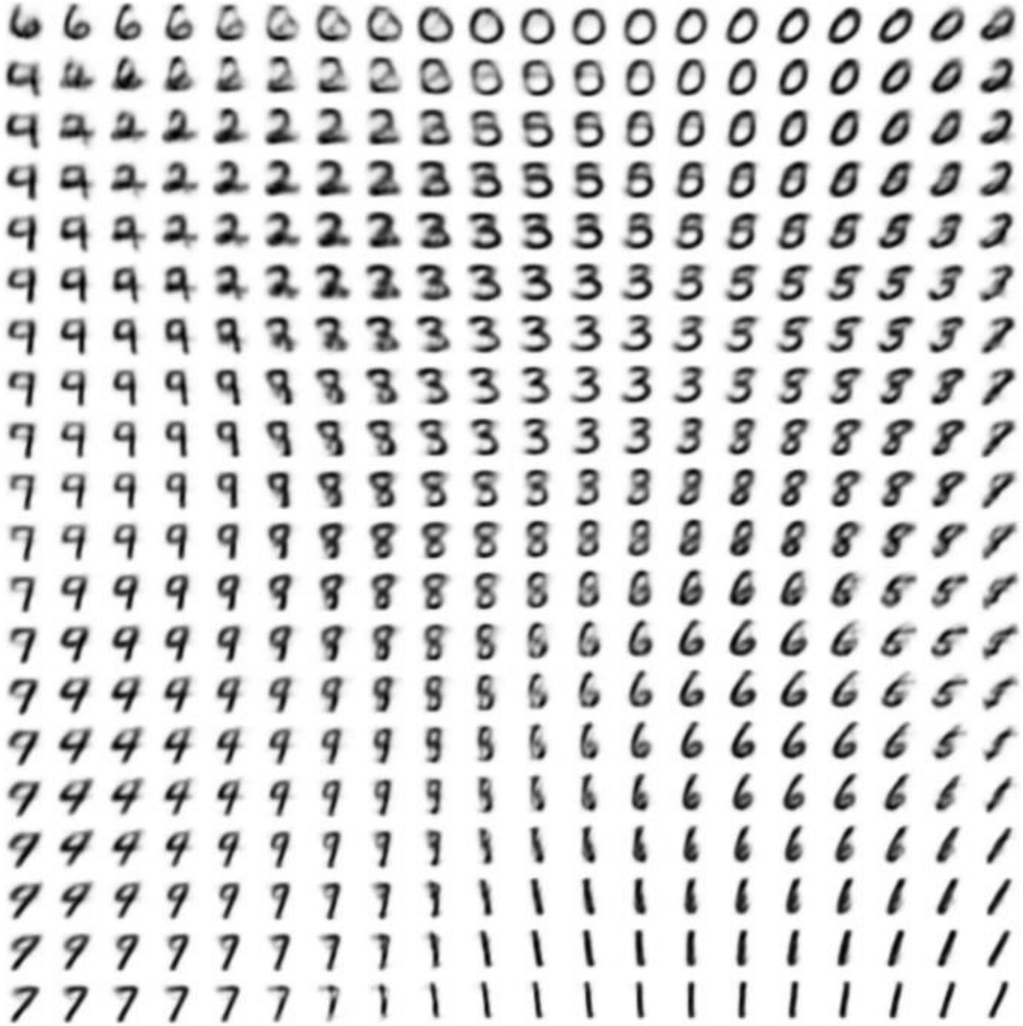
- Objective for the entire dataset:

$$\mathbb{E}_{x\sim D}\left[\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}\big(q_\phi(z|x), p(z)\big)\right]$$

For further details, see: C. Doersch, [Tutorial on Variational Autoencoders](), 2016

Original results

- Learned 2D manifolds:



D. Kingma and M. Welling, [Auto-Encoding Variational Bayes](), ICLR 2014

Variational autoencoders: Generating data

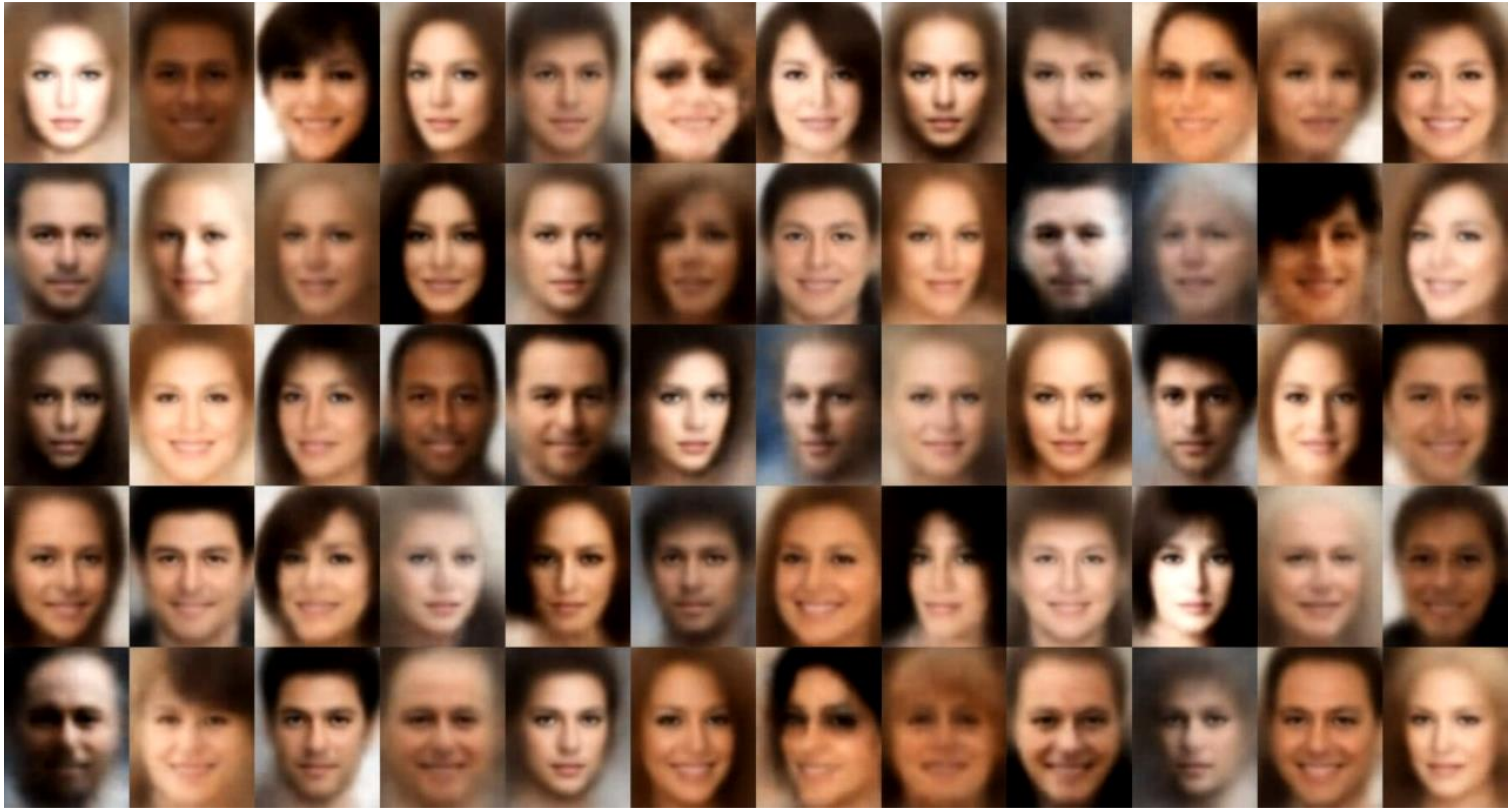


Image source

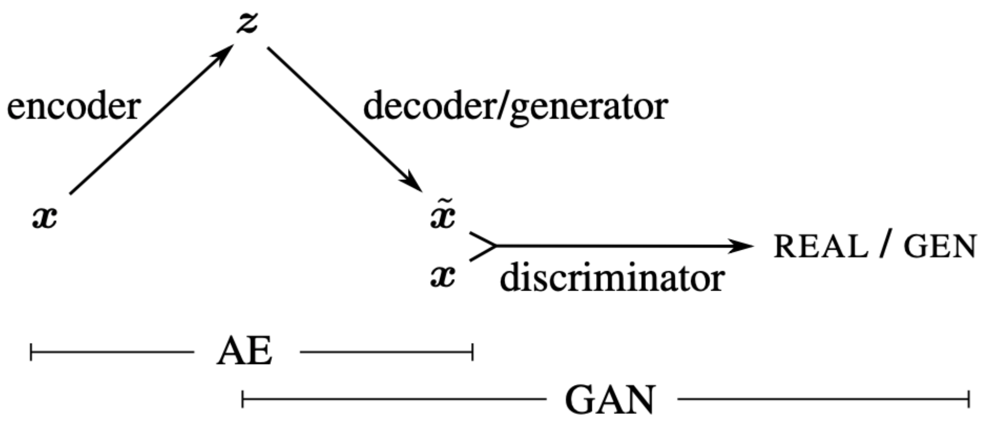
Basic VAE framework: Summary

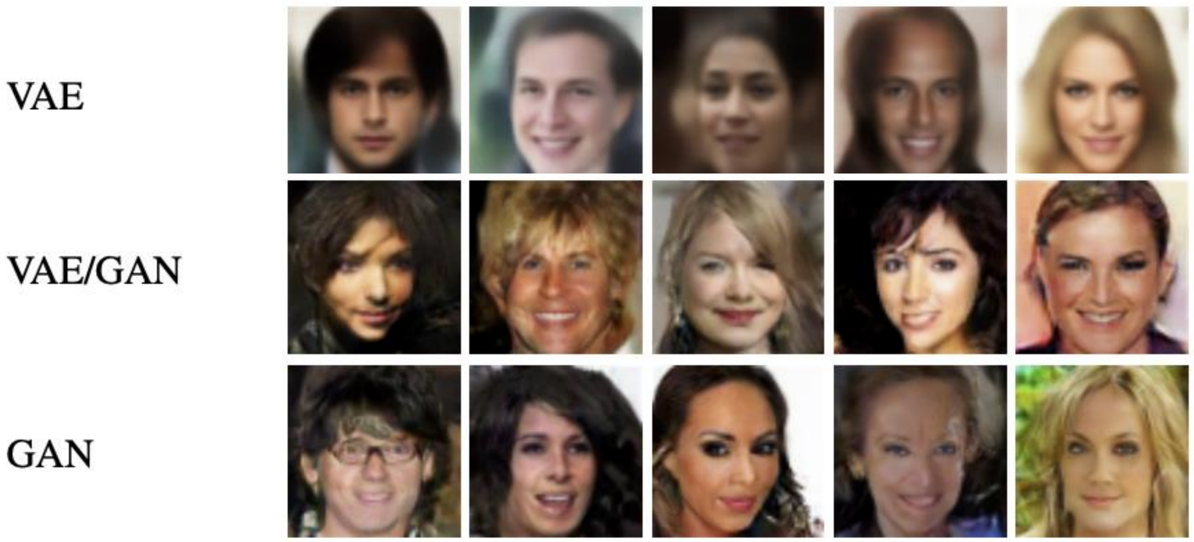
- Pros:
  - Principled mathematical formalism for generative models
  - Allows inference of code given image, can be useful for controlling the latent space
- Cons:
  - Samples blurrier and lower quality compared to GANs
- Active areas of research:
  - More powerful and flexible approximations for relevant probability distributions
  - Combining VAEs and GANs
  - Incorporating structure in latent variables, e.g., hierarchical or categorical distributions

Adapted from [J. Johnson]

Combining VAEs and GANs

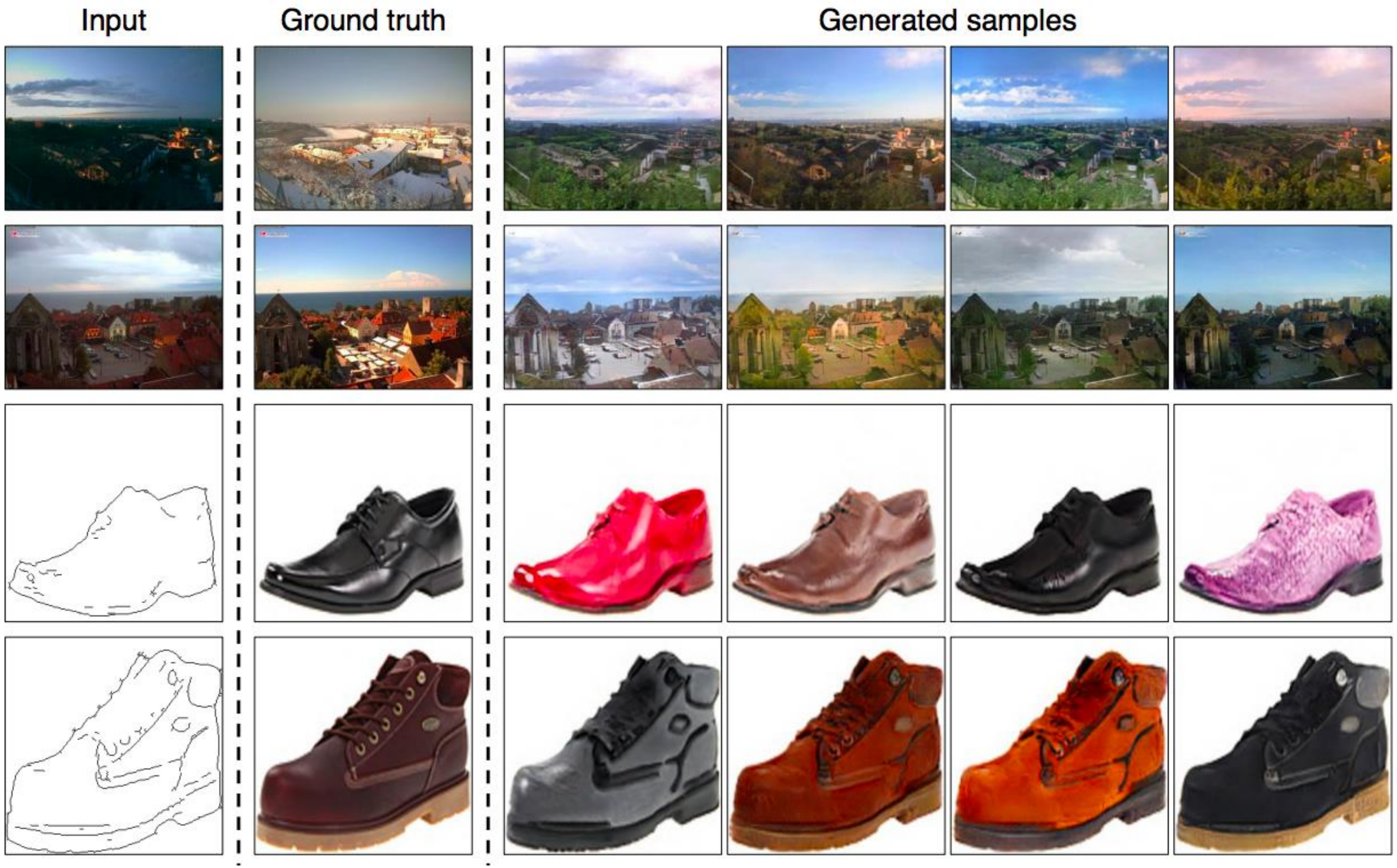
- Define decoder probability model $p_\theta(x|z)$ not in terms of reconstruction errors in pixel space, but in terms of errors in discriminator feature space





A. Larsen, S. Sonderby, H. Larochelle, O. Winther, [Autoencoding beyond pixels using a learned similarity metric](), ICML 2016

Combining VAEs and GANs: BicycleGANs



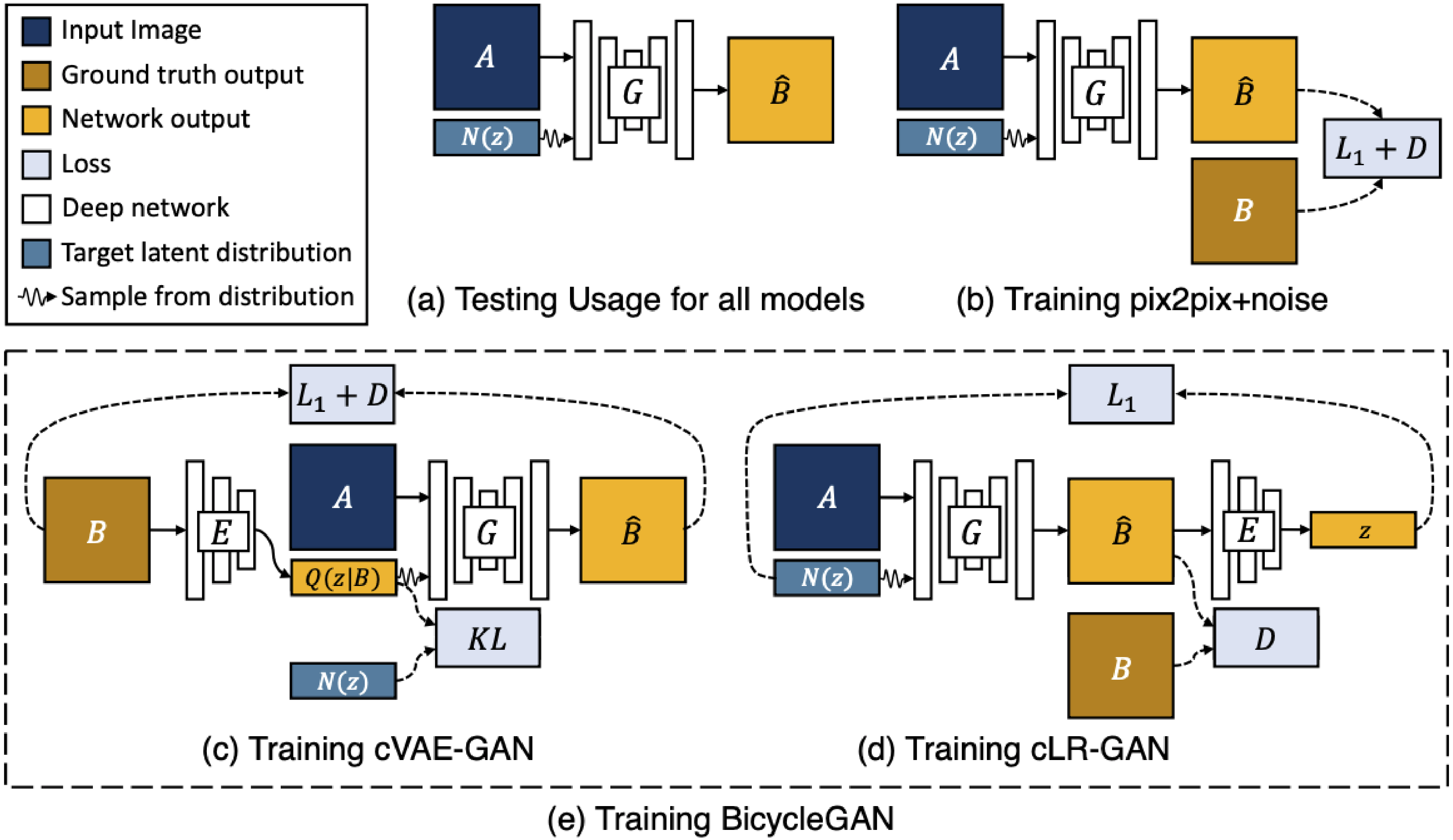
J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, Toward Multimodal Image-to-Image Translation, NIPS 2017

Combining VAEs and GANs: BicycleGANs

- Key ideas:
  - Image-to-image translation is a *one-to-many* problem. Need to model conditional distribution of output given input parametrized by z
  - To prevent mode collapse (or many-to-one mapping from z to output), need to encourage the mapping between output and latent code to be invertible
  - Propose BicycleGAN framework to simultaneously learn mappings in both directions

J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman,
Toward Multimodal Image-to-Image Translation, NIPS 2017

Combining VAEs and GANs: BicycleGANs



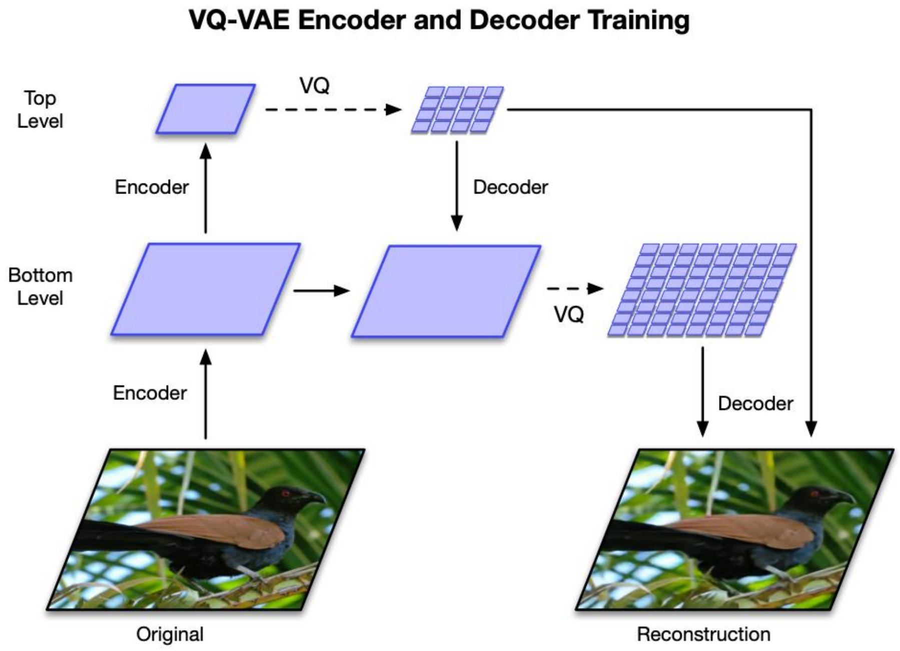
J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, [Toward Multimodal Image-to-Image Translation](), NIPS 2017

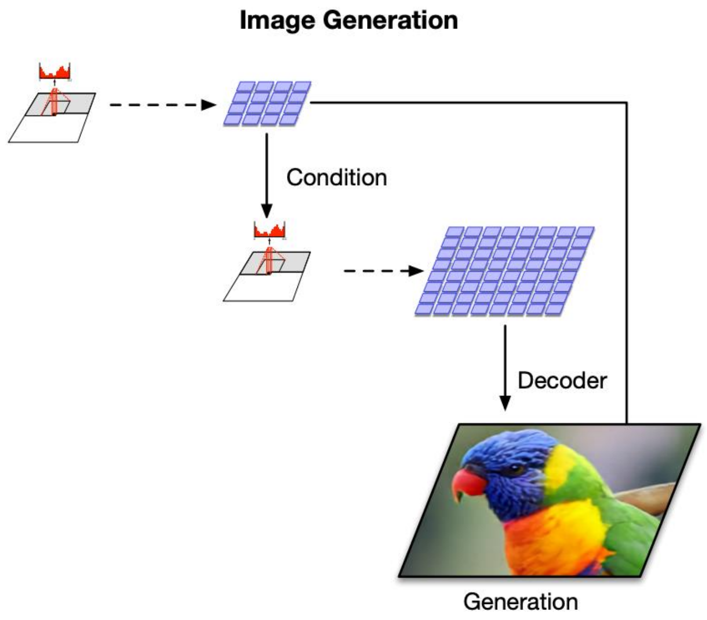
Generating better samples: VQ-VAE-2

- Combining VAE and autoregressive models:

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

Use a multiscale autoregressive model (PixelCNN) to sample in latent code space

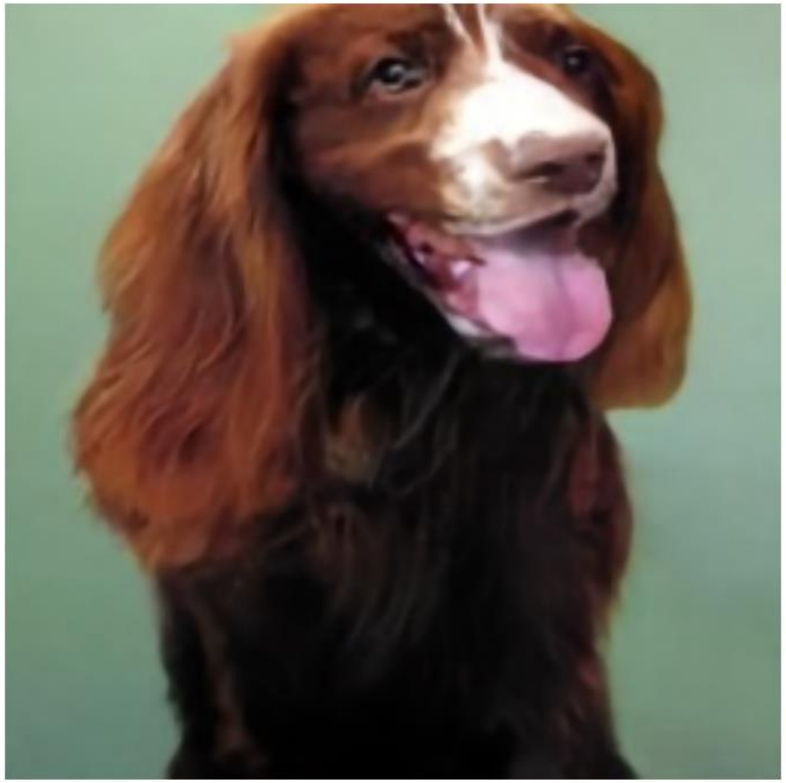




A. Razavi, A. van den Oord, O. Vinyals, [Generating Diverse High-Fidelity Images with VQ-VAE-2](), NeurIPS 2019

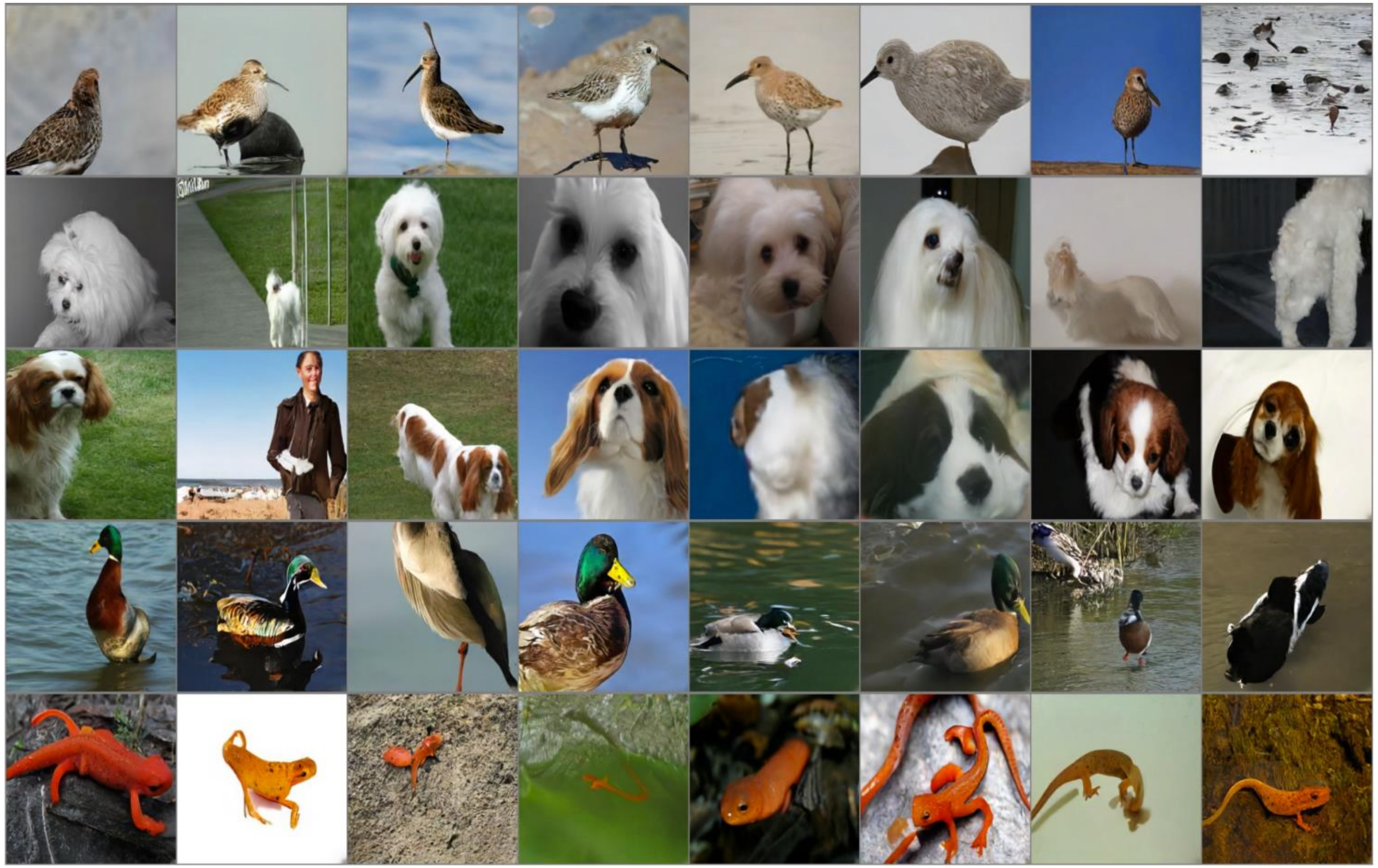
Generating better samples: VQ-VAE-2

- 256 x 256 class-conditional samples, trained on ImageNet:



Generating better samples: VQ-VAE-2

- 256 x 256 class-conditional samples, trained on ImageNet:



Redshank

Pekinese

Papillon

Drake

Spotted Salamander

Generating better samples: VQ-VAE-2

- 1024 x 1024 generated faces, trained on FFHQ:

Generating better samples: Hierarchical VAE

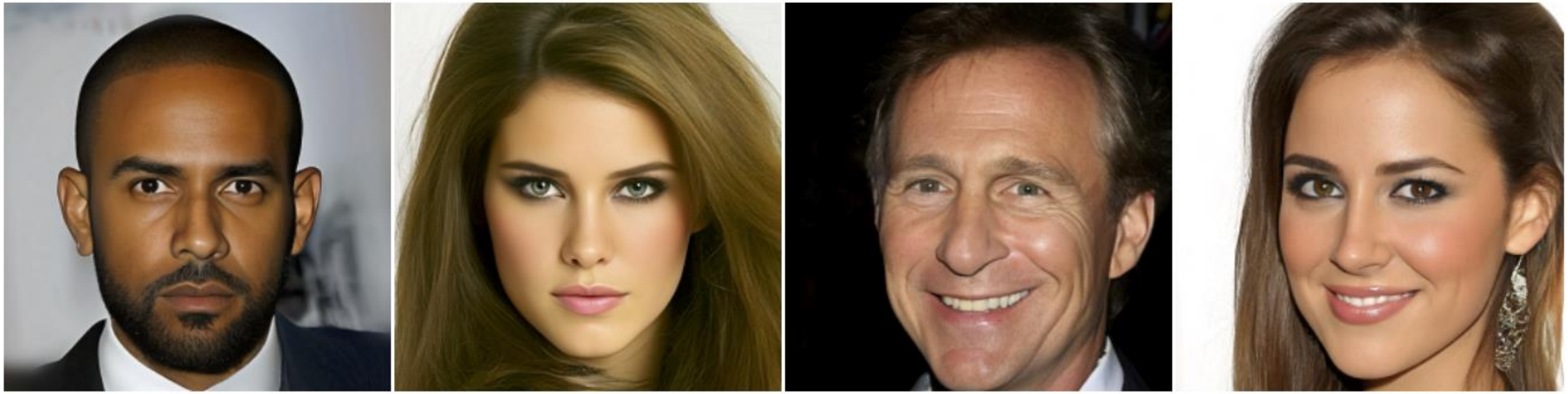


Figure 1: 256×256-pixel samples generated by NVAE, trained on CelebA HQ [28].

A. Vahdat, J. Kautz, [NVAE: A Deep Hierarchical Variational Autoencoder](), arXiv 2020

Acknowledgement

Thanks to the following courses and corresponding researchers for making their teaching/research material online

- Deep Learning, Stanford University
- Introduction to Deep Learning, University of Illinois at Urbana-Champaign
- Introduction to Deep Learning, Carnegie Mellon University
- Convolutional Neural Networks for Visual Recognition, Stanford University
- Natural Language Processing with Deep Learning, Stanford University
- And Many More ......