Generative Modeling

Merging VAE and GANs

Denis Derkach, Artem Ryzhikov, Maxim Artemev

Laboratory for methods of big data analysis

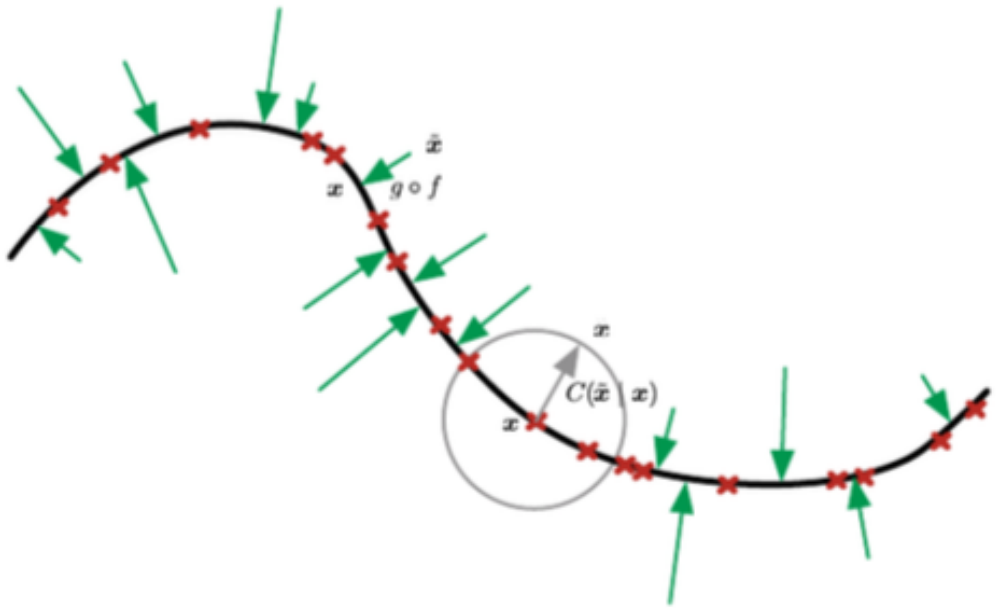
LAMBDA • HSE

Spring 2021

In this Lecture

- Special Discriminator Structures
  - Energy-based Generative Adversarial Network
  - Boundary Equilibrium Generative Adversarial Networks
  - Discriminator Rejection Sampling
- Merging VAE and GAN:
  - VEEGAN, AAE, VGH etc.
- Outlook

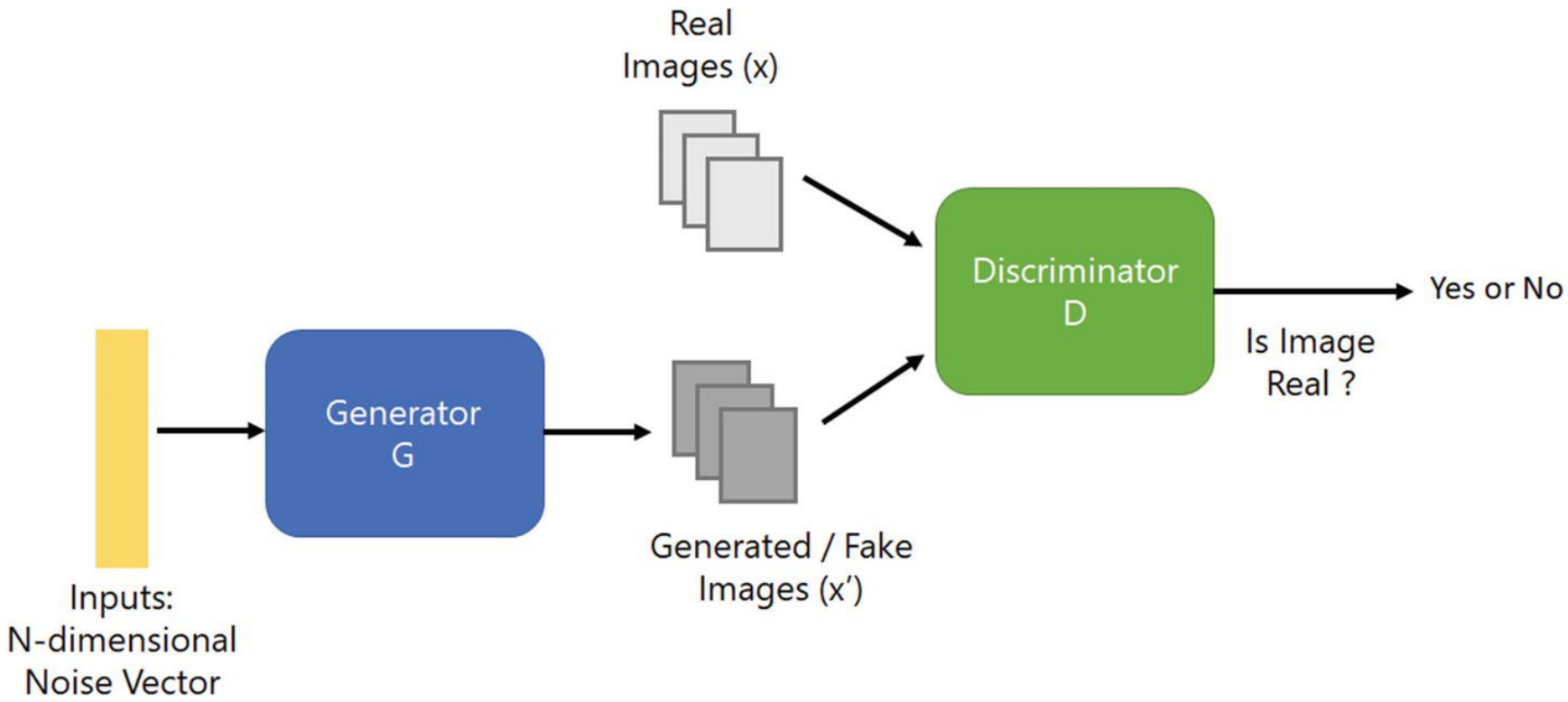
Reminder Contractive Denoising Autoencoders



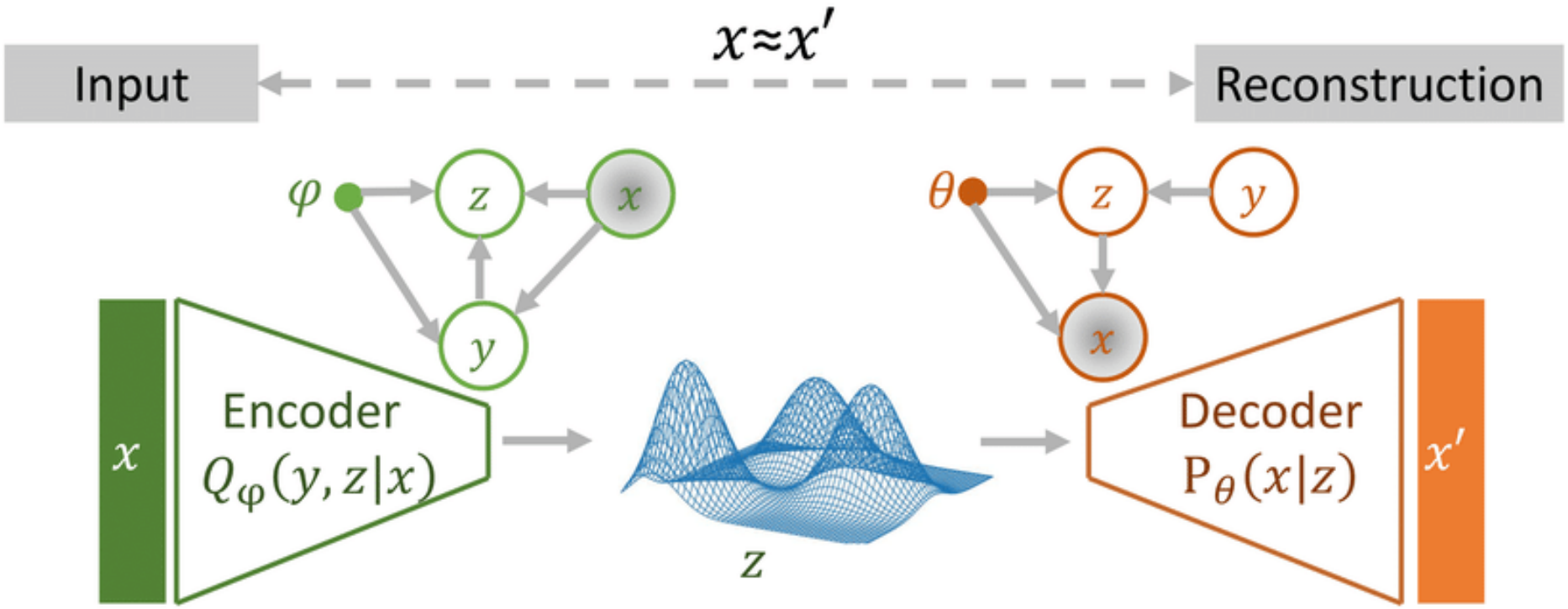
- The true signal is always situated on a manifold inside the R^D space.
- Denoising autoencoder is trained to map a corrupted data point $\tilde{x}$ back to the original data point x.

https://arxiv.org/abs/1305.6663

Vanilla GAN


Real
Images (x)
Discriminator
D
Yes or No
Is Image
Real ?
Generator
G
Generated / Fake
Images (x')
Inputs:
N-dimensional
Noise Vector

VAE



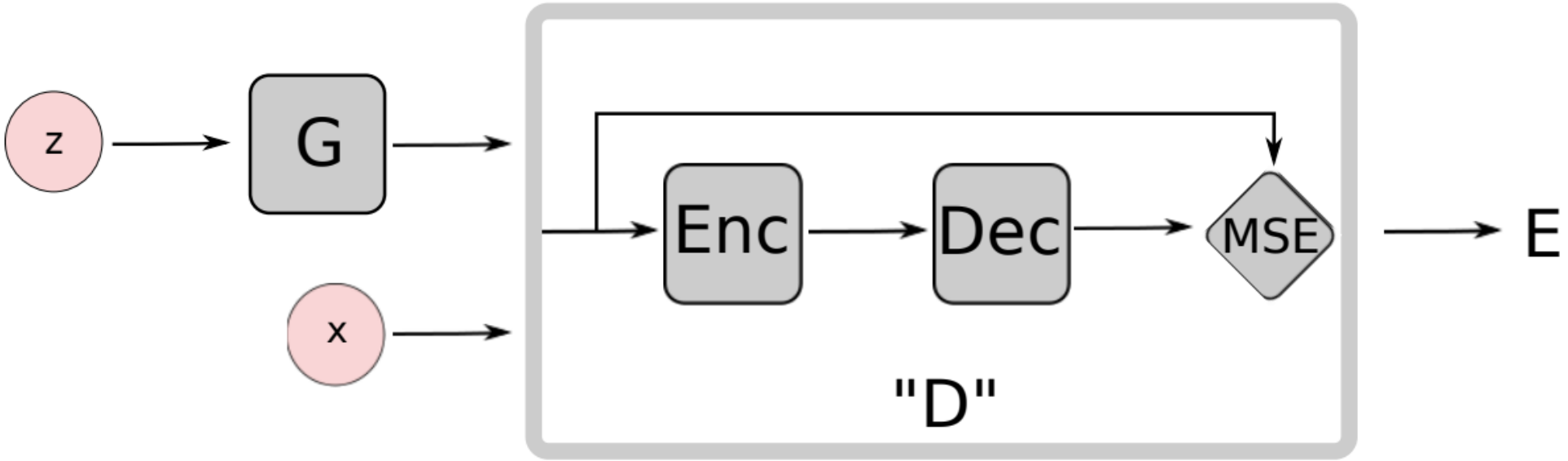
$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log P_\theta(x, y, z) - \log Q_\phi(y, z|x)\right]$$

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{Q_\phi(y,z|x)}\left[\log \frac{P(y)}{Q_\phi(y|x)} + \log \frac{P_\theta(z|y)}{Q_\phi(z|x,y)} + \log P_\theta(x|z)\right]$$

Entropy | Regularization | Reconstruction

Energy-based GAN

Autoencoder Discriminator



- Use AE to extract latent features of the input image by an encoder and reconstruct it again with the decoder with MSE loss:

$$D(x) = ||Dec\big(Enc(x)\big) - x||$$

https://arxiv.org/abs/1609.03126

EB-GAN training

For $[.]^+ = \max(0, .)$:

$$\mathcal{L}_D(x, z) = D(x) + [m - D(G(z))]^+;$$
$$\mathcal{L}_G(x, z) = D(G(z)),$$

- **autoencoder**: reconstruction cost $D(x)$ for real images is low;
- $D(x)$ is trained first several rounds;
- once $G(z)$ generates sufficiently good images $D(x)$ training resumes;
- repelling loss to address AE collapse problem:

$$f_{PT} = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \frac{S_i^T S_j}{||S_i|| ||S_j||}.$$

where $S \in \mathbb{R}^{s \times N}$ a batch of sample of size N representations taken from the encoder output layer.

GAN energy interpretation

For $[.]^+ = \max(0, .)$:

$$\mathcal{L}_D(x, z) = D(x) + [m - D(G(z))]^+;$$
$$\mathcal{L}_G(x, z) = D(G(z)),$$

- **autoencoder**: reconstruction cost $D(x)$ for real images is low;
- $D(x)$ does not have probability interpretation;
- one can use **energy interpretation** instead.

EB-GAN results

Figure 4: Generation from the grid search on MNIST. Left(a): Best GAN model; Middle(b): Best EBGAN model. Right(c): Best EBGAN-PT model.

Boundary equilibrium GAN

Wasserstein Distance lower bound

- **Wasserstein distance:**

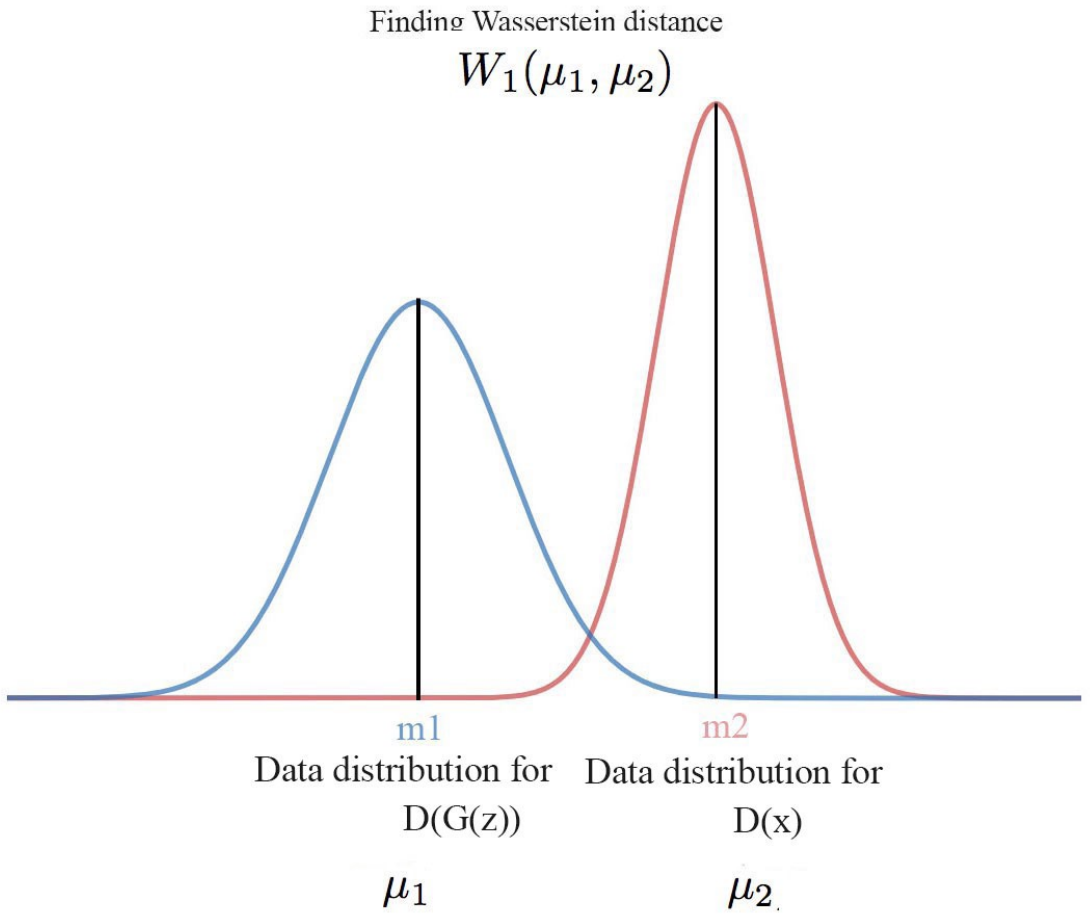
$$W(\mu 1, \mu 2) = \inf_{\gamma \in \Pi} \mathbb{E}_{(x,y)\sim\gamma} ||x - y||$$

- **Jensen's inequality**:

$$W(\mu 1, \mu 2) \geq \inf_{\gamma \in \Pi} |\mathbb{E}_{(x,y)\sim\gamma} |x - y|| =$$

$$= |m_1 - m_2|,$$

where m_i are the mean on μ_i.



https://arxiv.org/abs/1703.10717

Wasserstein Discriminator

- We have $D(x)$ as AE:

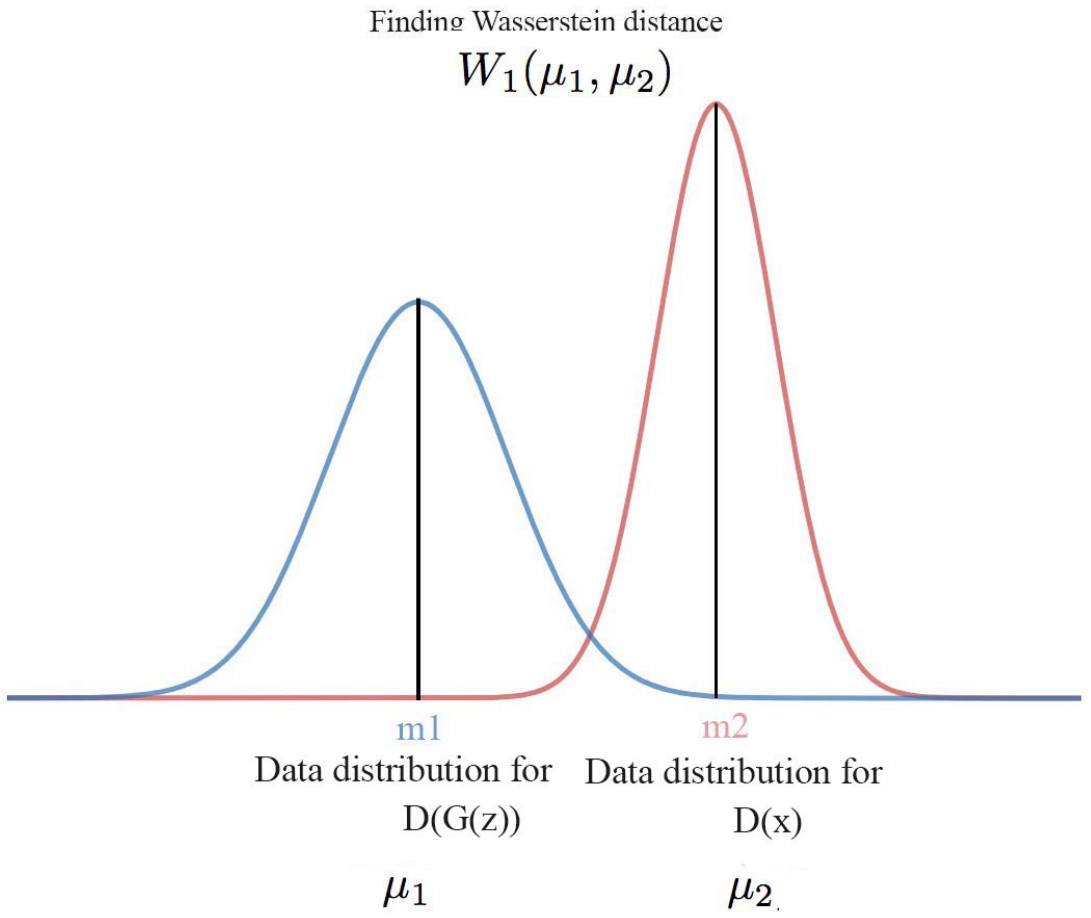
$$D(x) = ||Dec\big(Enc(x)\big) - x||$$

$$\mathcal{L}_D = W(\mu_1\mu_2) \geq |m_1 - m_2|.$$

- We can use $D(x)$ in minibatch instead of mean:

$$\mathcal{L}_D = D(x) - D\big(G(z)\big).$$

- We thus optimize W between losses.
- No need for K-Lipshitz, since no Kantorovich-Rubinstein duality is used.



Equilibrium term

- we need to maintain balance between G and D:

$$\mathbb{E}(D(x)) = \mathbb{E}(D(G(z)))$$

- we thus can use a parameter to balance the impact:

$$\gamma = \frac{\mathbb{E}(D(x))}{\mathbb{E}(D(G(z))}.$$

- γ can be chosen to sharpen the image.

BEGAN formulation

- We thus can write full optimization for BEGAN

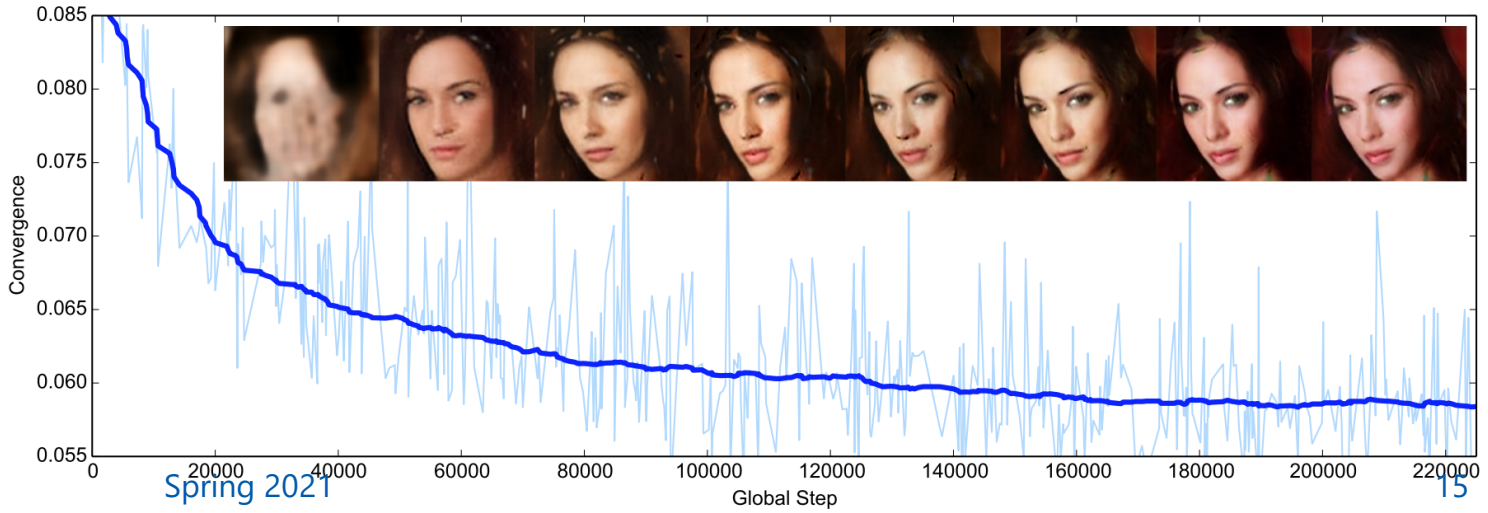
$$\mathcal{L}_D = D(x) - k_t D(G(z));$$

$$\mathcal{L}_G = D(G(z));$$

$$k_{t+1} = k_t + \lambda_k(\gamma D(x) - D(G(z))).$$

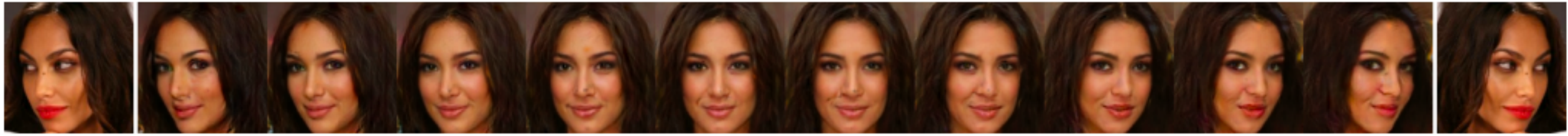
- Dropping γ leads to mode collapse.
- To monitor the convergence:

$$M_{global} = D(x) + (\gamma D(x) - D(G(z)))$$



BEGAN results

(c) Our results (128x128 with 128 filters)



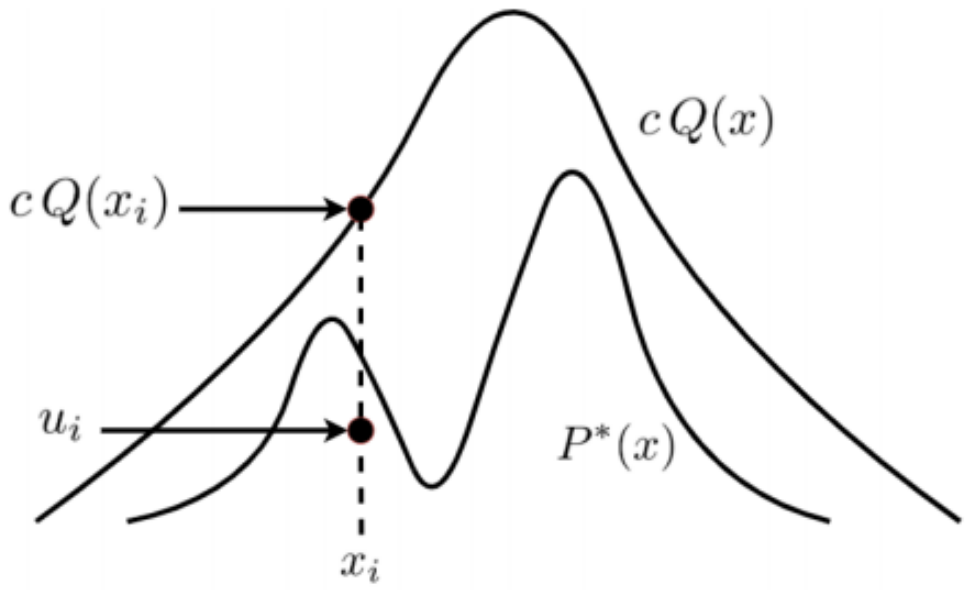
(d) Mirror interpolations (our results 128x128 with 128 filters)

Wrap-up

- We can change the architecture of discriminator.
- This might lead to change of the optimization idea.
- If we use autoencoder as discriminator we have access to the energy instead of probability.
- We can optimize Wasserstein distance not only for datasets, but also for results of function.

Rejection Sampling

Rejection Sampling



Input: $P^*(X), Q(X), c$
Output: $\mathcal{S} = \{x_i\}_{i=1}^n \sim P^*(X)$
$\mathcal{S} \leftarrow \varnothing$
for *sample index* i *from* 1 *to* n **do**
  $x_i \sim Q(X)$
  $u_i \sim U(0, cQ(x_i))$
  if $u_i < P^*(x_i)$ **then**
    Accept x_i: $\mathcal{S} \leftarrow \mathcal{S} \cup \{x_i\}$
  else
    Reject x_i: $i \leftarrow i - 1$

https://arxiv.org/pdf/2011.00901.pdf

Ideal Discriminator

- Ideal discriminator:

$$D^*(x) = \frac{p(x)}{p(x) + q_\theta(x)}.$$

- Remember f-GAN idea of last layer:

$$D^*(x) = \frac{1}{1 + e^{-\widetilde{D}^*(x)}} = \frac{p(x)}{p(x) + q_\theta(x)}.$$

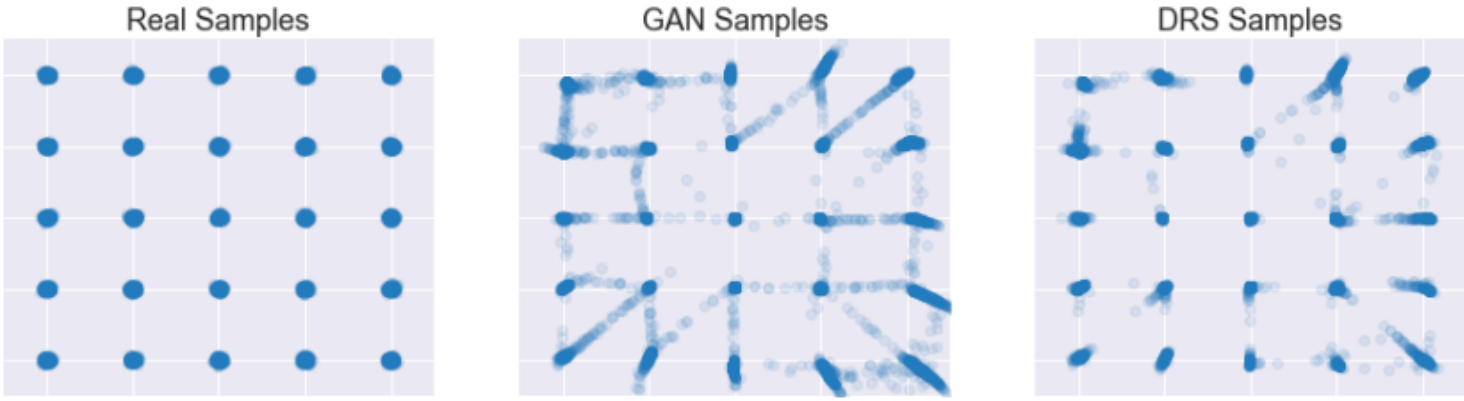
- Thus:

$$\frac{p(x)}{q_\theta(x)} = e^{\widetilde{D}^*(x)}.$$

- This defines constant for rejection sampling.

https://arxiv.org/pdf/1810.06758.pdf

Discriminator Rejection Sampling



ImageNet	IS	FID
Without DRS	52.34 ± 0.45	18.21 ± 0.14
With DRS	61.44 ± 0.09	17.14 ± 0.09

Results suggest that the quality of sampling is improved

https://arxiv.org/pdf/1810.06758.pdf

Your GAN is secretly an energy based model

- Previous results can be revisited:

$$\frac{p(x)}{q_\theta(x)} = e^{\widetilde{D}^*(x)}.$$

- And applied to the latent space. This will create a rule for new latent space distribution:

$$p_t(z) = p_0(z)r(z)/C.$$

- Which can be rewritten as:

$$p_t(z) = e^{-E(z)}/Z_0, \text{ with tractable } E(z):$$
$$E(z) = -\log p_0(z) - d\big(G(z)\big).$$

- This can be used to define MCMC in latent space and later obtain $x \sim G(z)$.

Energy-based sampling: results

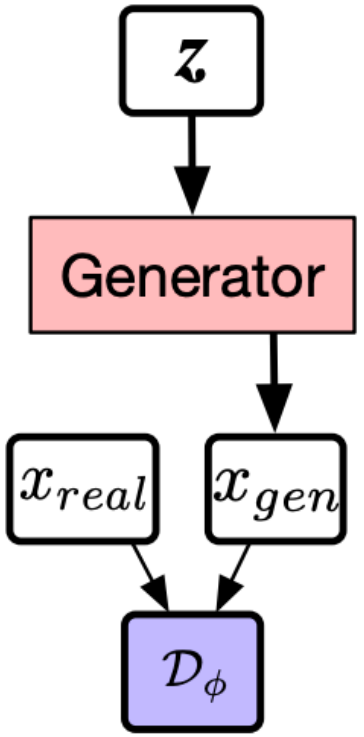
Figure 4. Top-5 nearest neighbor images (right columns) of generated samples (left column).

Discussion

- GAN's discriminator can enable better modeling of the data distribution with Discriminator Driven Latent Sampling.
- The major advantage of DDLS is that it allows MCMC sampling in the latent space.

VAE+GAN

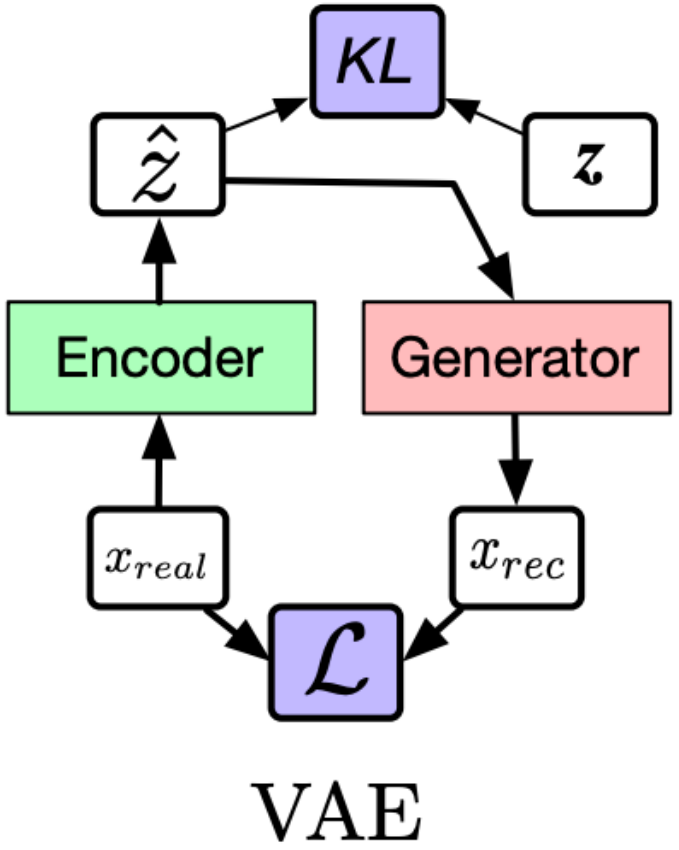
Improving GANs



DCGAN

- Main problems with GAN:
  - mode collapse;
  - intractable likelihood.
- Idea:
  - use likelihood-based mode;.
  - have easier inference;
  - diversify sampling.

Reminder: VAE structure



VAE

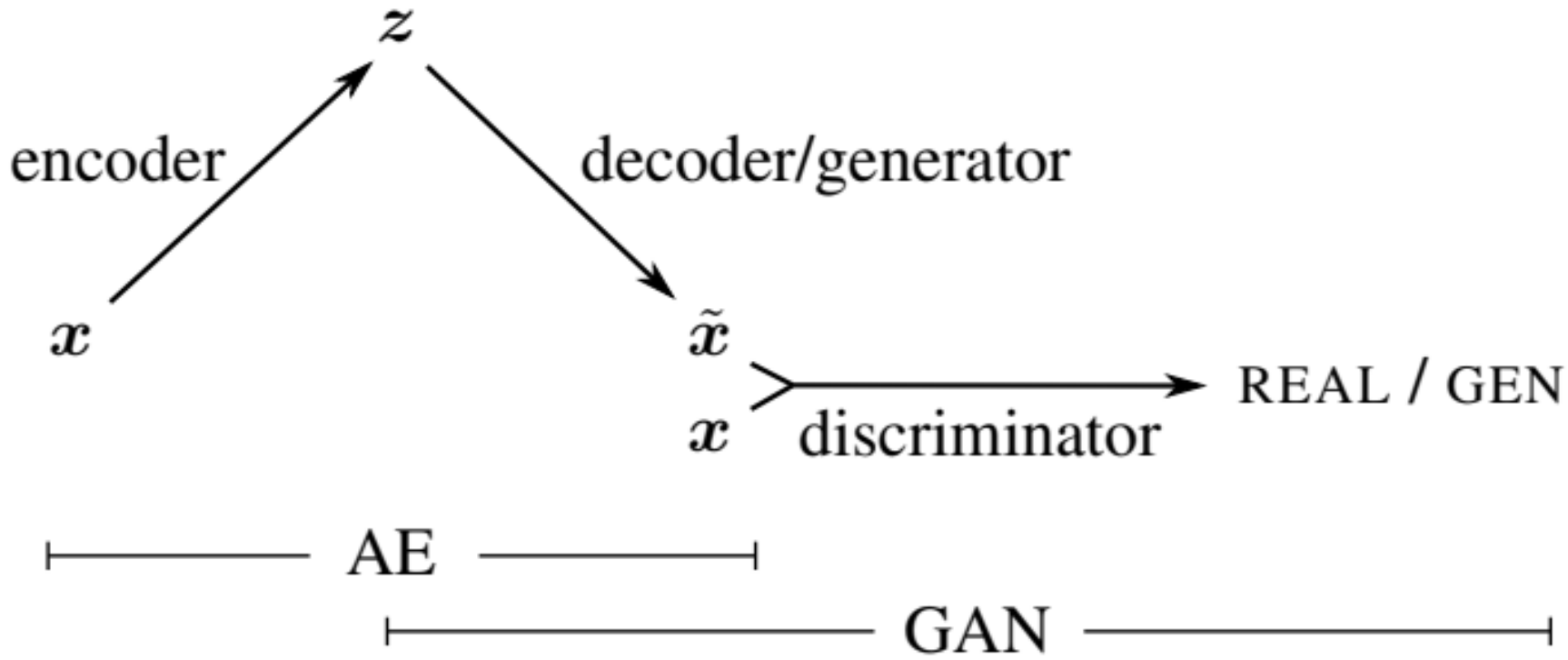
$$\mathcal{L}_{prior} = KL(q_n(z|x)||p(z))$$

$$+$$

$$-\mathcal{L}_{pixel} = \mathbb{E}_{q_n(z|x)}(\log p_\theta(x|z)$$

$$= \mathcal{L}_{VAE}$$

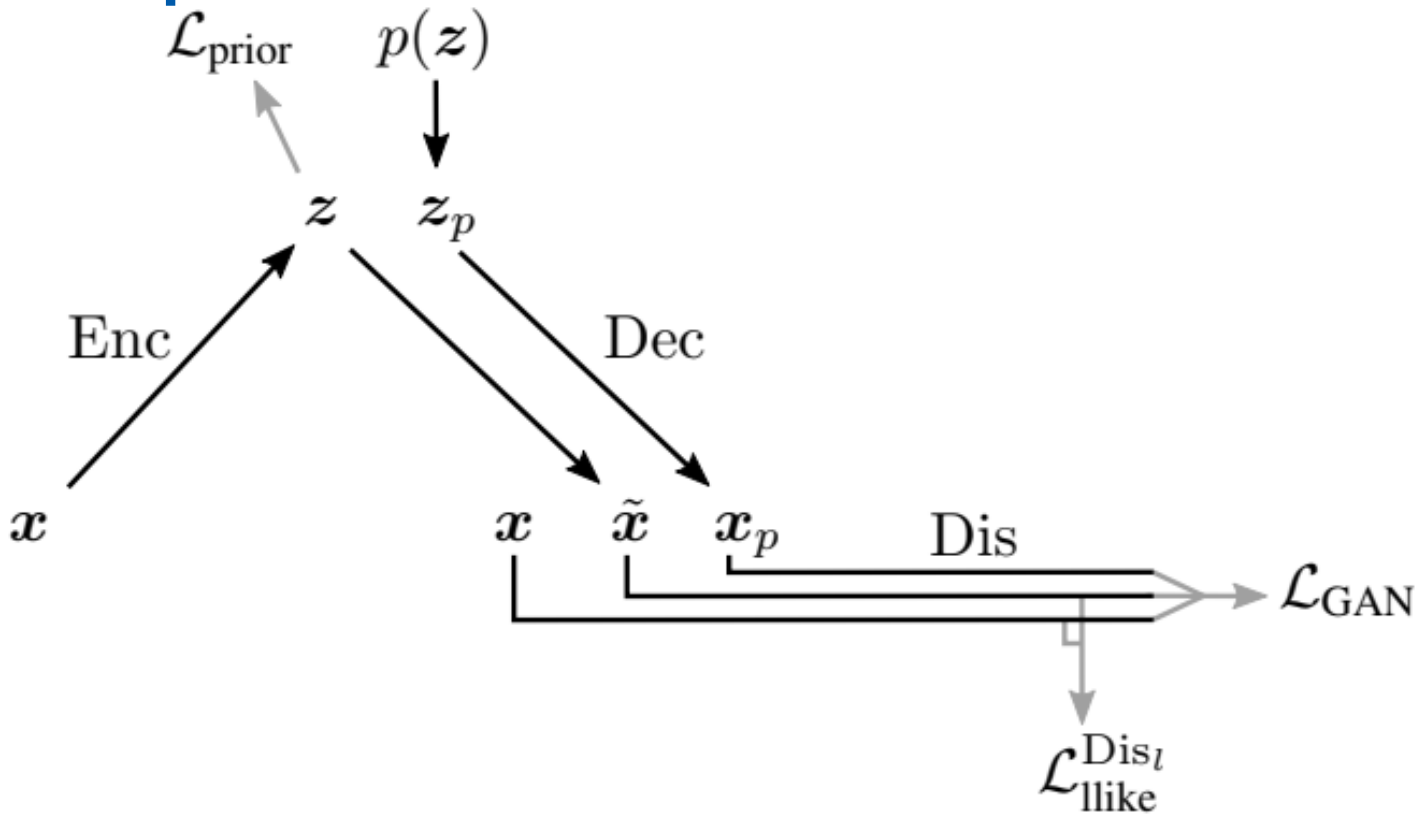
VAE+GAN



- problem of VAE is blurry output
- GAN overcomes due to the use of discriminator
- add GAN loss to VAE

https://arxiv.org/pdf/1512.09300.pdf

VAE+GAN expanded

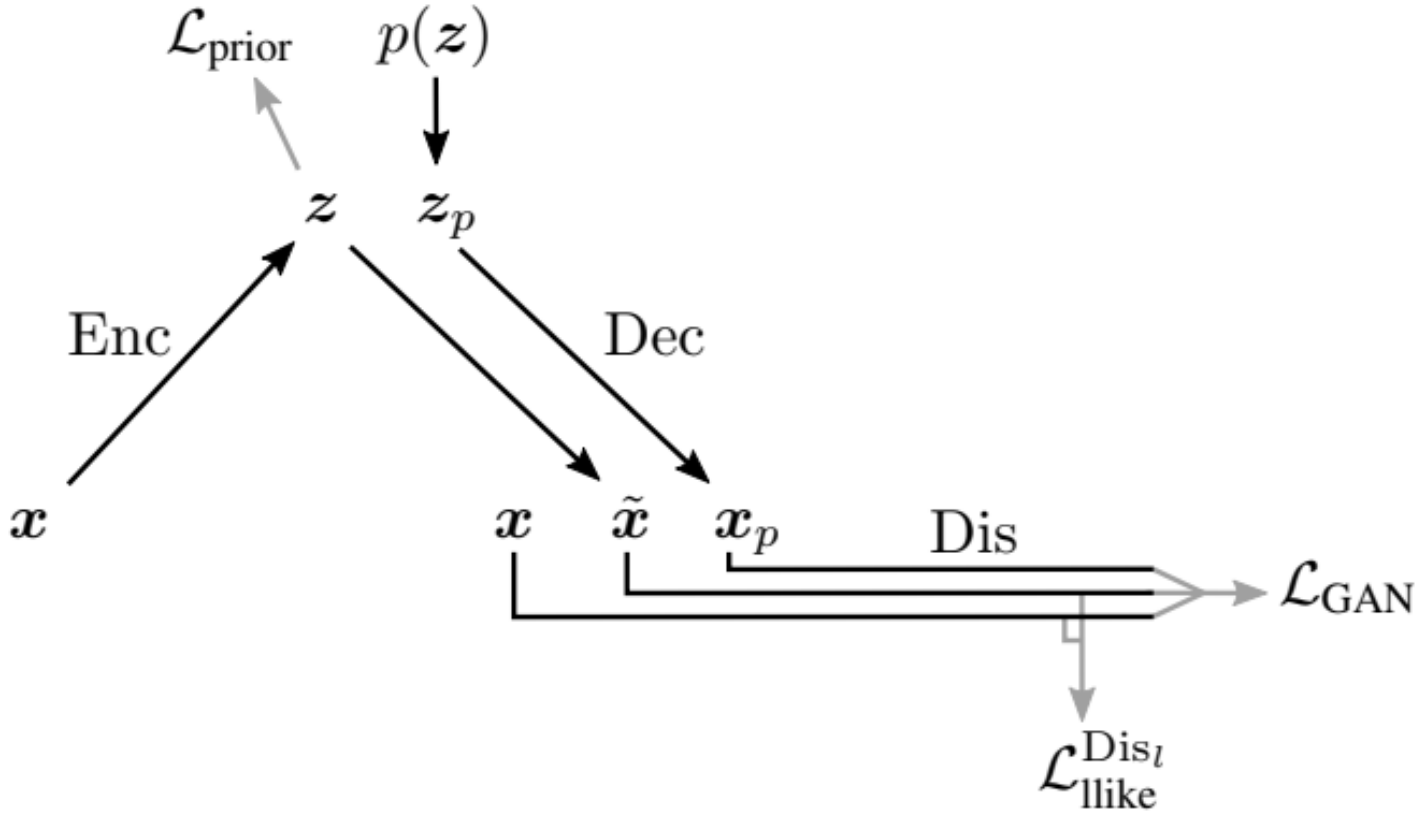


- Consider image as a whole:

$$-\mathcal{L}_{DIS} = \mathbb{E}_{q_n(z|x)} p(Dis_l(x)|z) = \mathbb{E}_{q_n(z|x)} N\left(Dis_l(x)|\, Disl(\tilde{x}), I\right), \qquad \tilde{x} \sim Dec(z)$$

where $Dis_l(x)$ denote the hidden representation of the l-th layer of the discriminator.

VAE+GAN final loss



$$\mathcal{L}_{GAN+VAE} = \underline{\mathcal{L}_{prior} + \mathcal{L}_{DIS}} + \underline{\mathcal{L}_{GAN}}$$

Image **Style**

VAE+GAN results

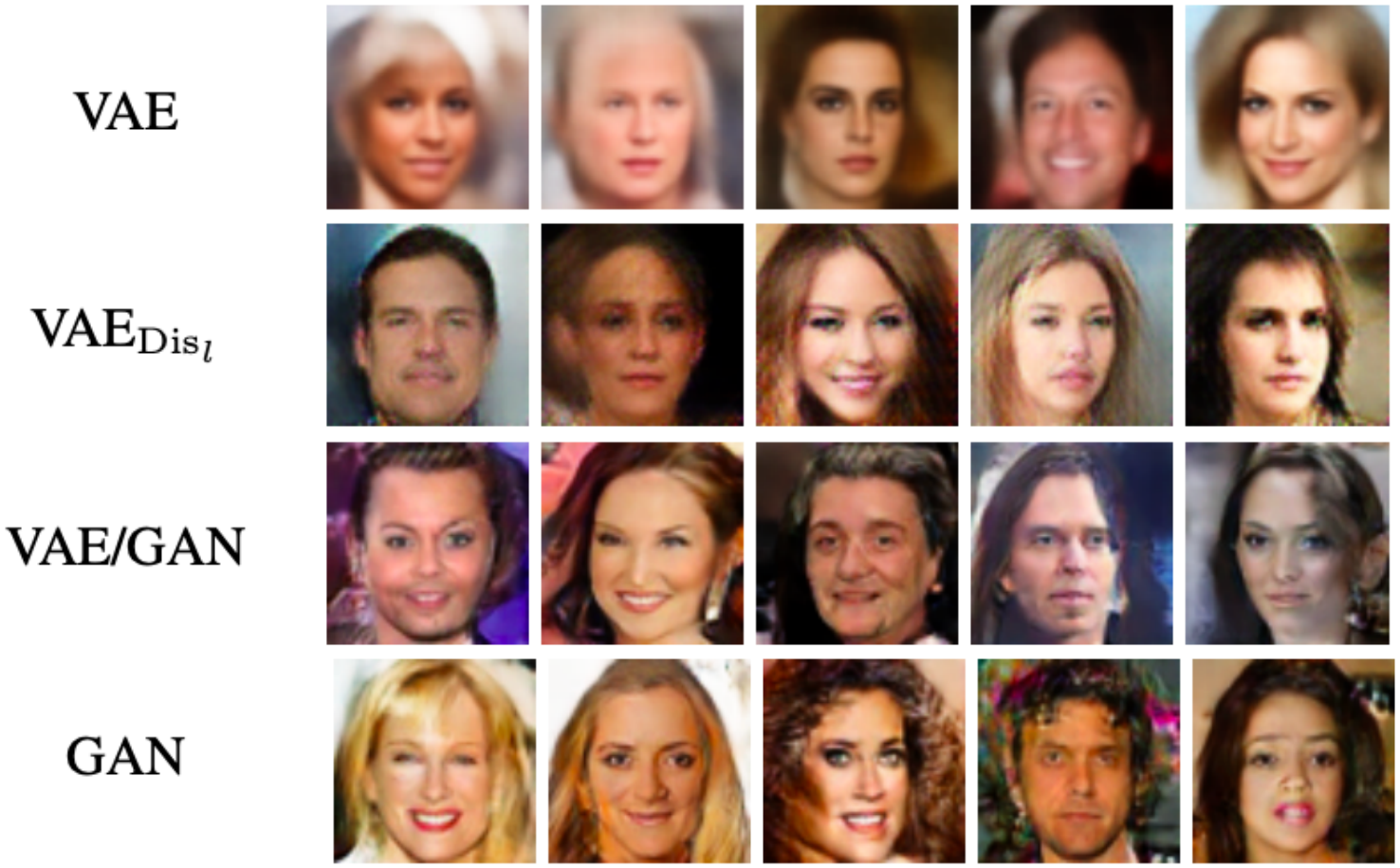


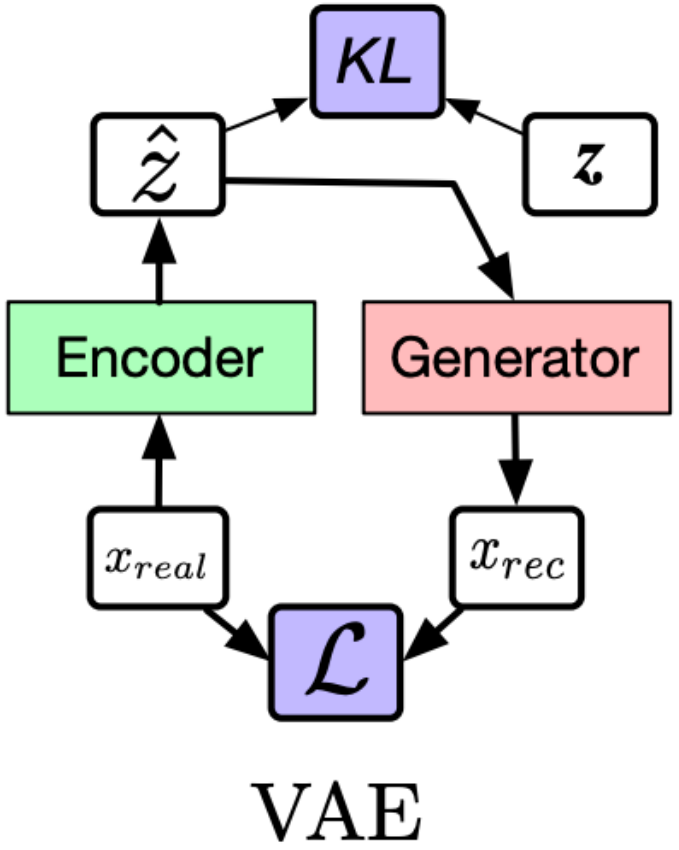
Figure 3. Samples from different generative models.

Wrap-up

- Benefits from combined loss and combined architecture.
- Better images than VAE or GAN standalone.
- Can be viewed as extension of VAE with adversarial selection.
- Can be viewed as extension of GAN with better latent space construction.

More VAEGANs

VAE structure

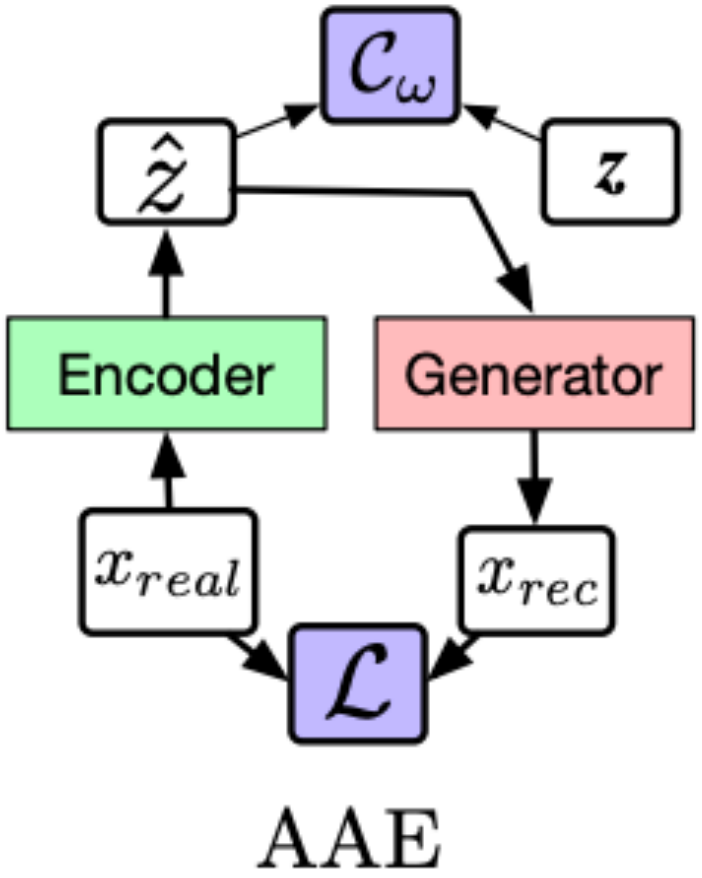


VAE

- KL divergence is considered as one of blurriness causes.
- We can switch this objective to GAN-like.

https://arxiv.org/pdf/1511.05644.pdf

AAE structure



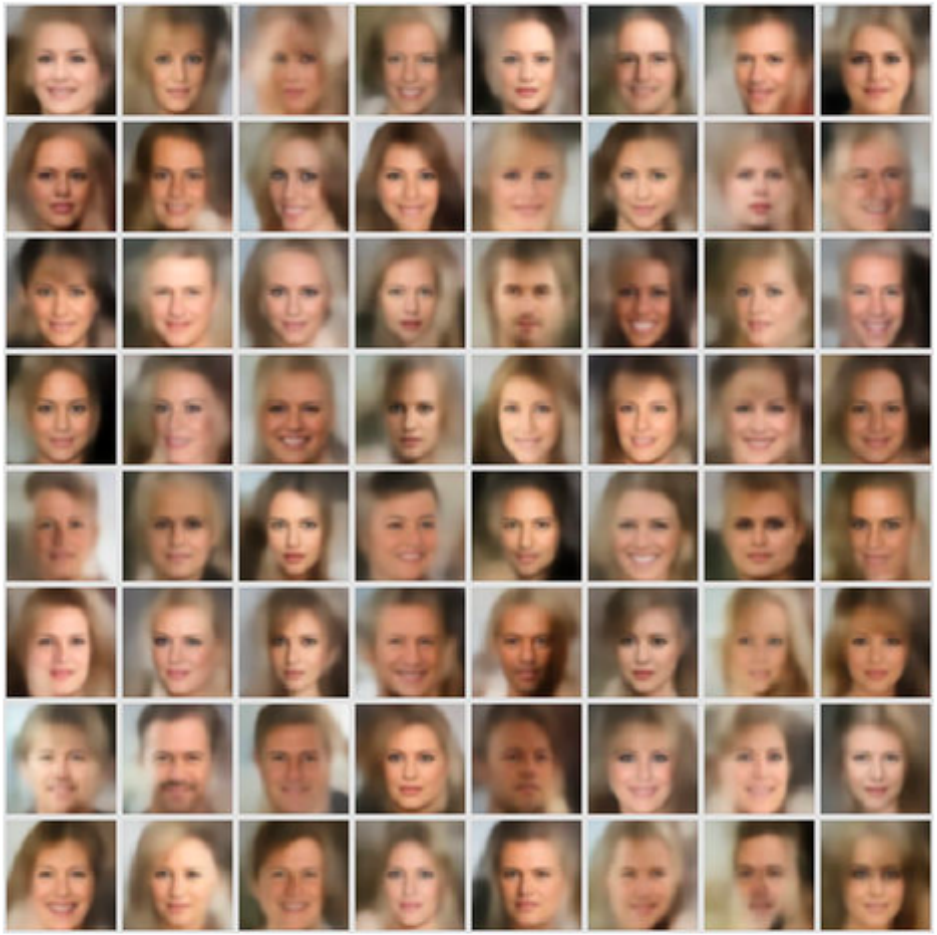
AAE

- KL divergence is considered as one of blurriness causes.
- We can switch this objective to GAN-like.
- Price to pay: ELBO becomes less constrained.

https://arxiv.org/pdf/1511.05644.pdf

AAE results

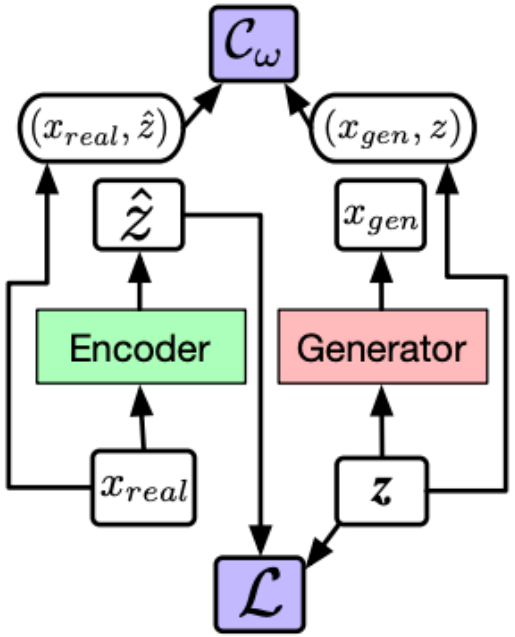
(b) VAE samples



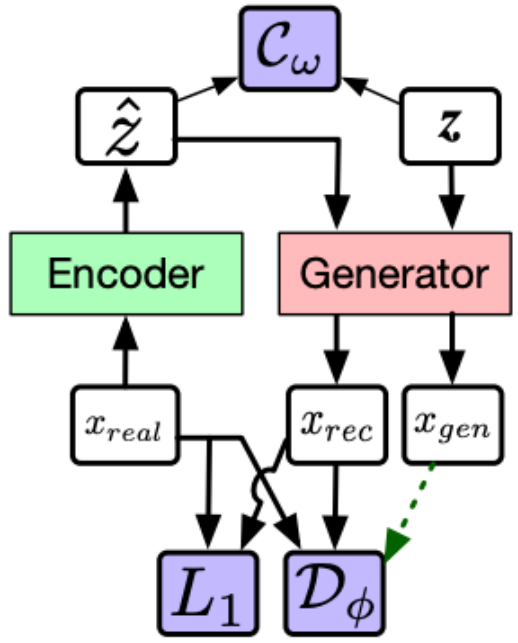
(c) AAE samples

https://arxiv.org/pdf/1802.06847.pdf

VEEGAN and VGH



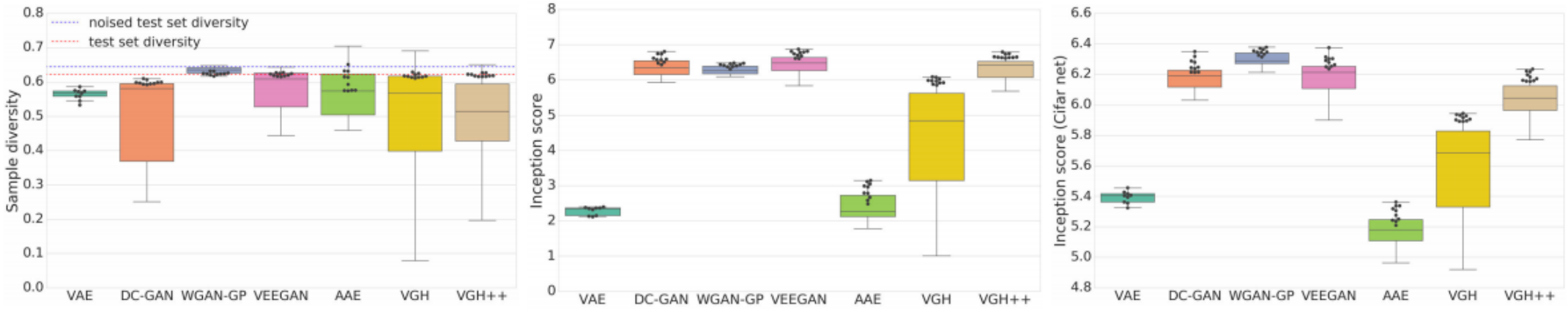
VEEGAN



VGH/VGH++

- VGH: Marginal matching and implicit distributions using GANs both in latent and visible space;
- VEEGAN: Directly match in joint space.

VAE+GAN merged



(a) Diversity score (CelebA) (b) Inception score (ImageNet) (c) Inception score (CIFAR)

Figure 8: (Left) Sample diversity on CelebA, and is viewed relative to test set: too much diversity shows failure to capture the data distribution, too little is indicative of mode collapse. We also report the diversity obtained on a noised-version of the test set, which has a higher diversity than the test set. (Middle) Inception scores on CIFAR-10. (Right) Inception scores computed using a VGG-style network on CIFAR-10. For inception scores, higher values are better. For test data, diversity score: 0.621, inception score: 11.25, inception score (using CIFAR-10 trained net): 9.18. Best results are shown with black dots, and box plots show the hyperparamter sensitivity.

- Our aim was to improve GAN mode collapse and VAE blurriness.
- The results are yet to be improved.