Martin Mundt, Dr. Iuliia Pliushch, Prof. Dr. Visvanathan Ramesh

Pattern Analysis & Machine Intelligence
Praktikum: MLPR-SS21

Week 09: Adversarial Training

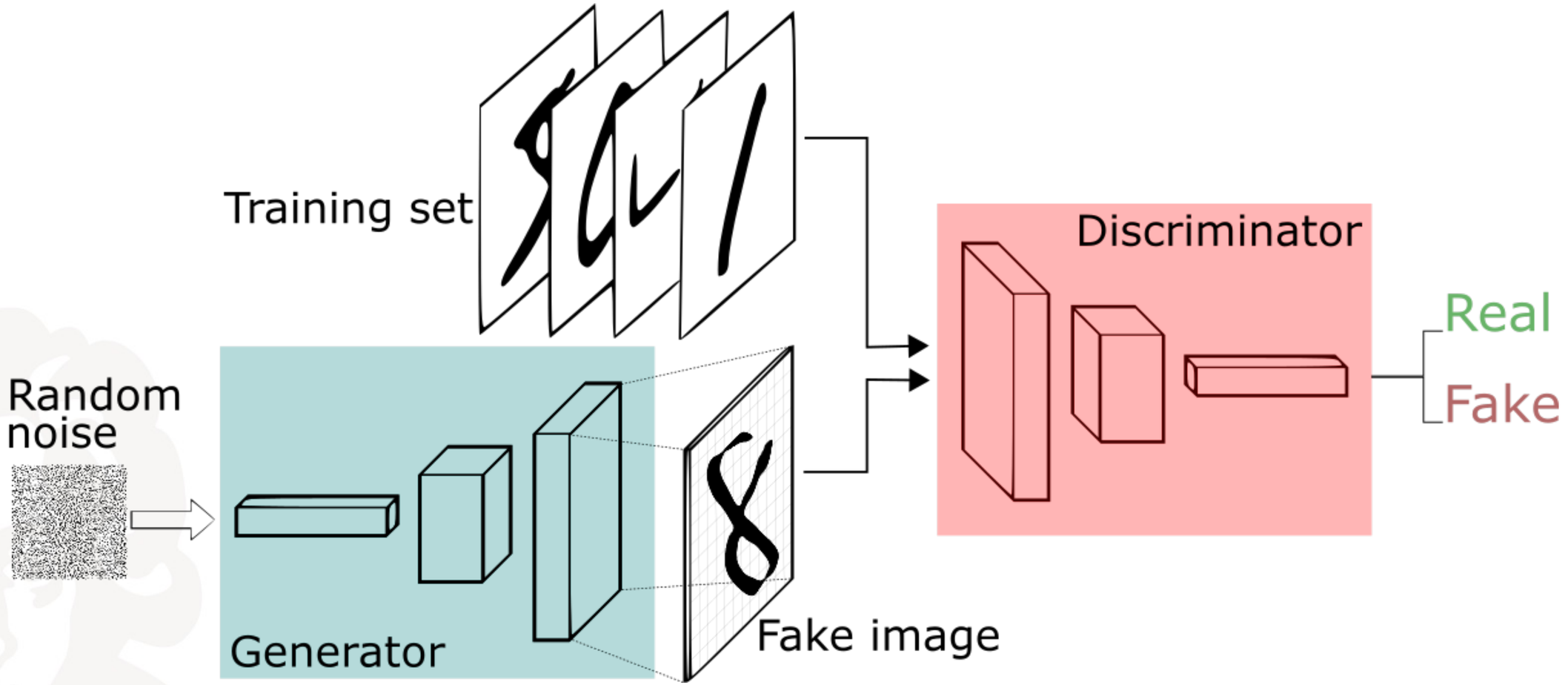


Image taken from: https://skymind.ai/images/wiki/GANs.png

We have already seen VAEs, why should we look at GANs?

We have already seen VAEs, why should we look at GANs?

All images created by BigGAN

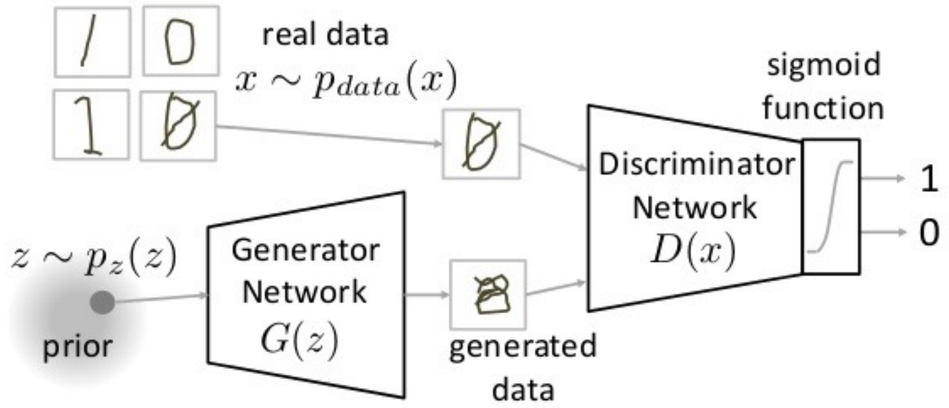
Image taken from: https://cdn-images-1.medium.com/max/2600/1*Yw2KxjmIkj8yqS-ykLCQCQ.png

How do GANs work? Alternating training

Generative Adversarial Networks

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

$$V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z))]$$



Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

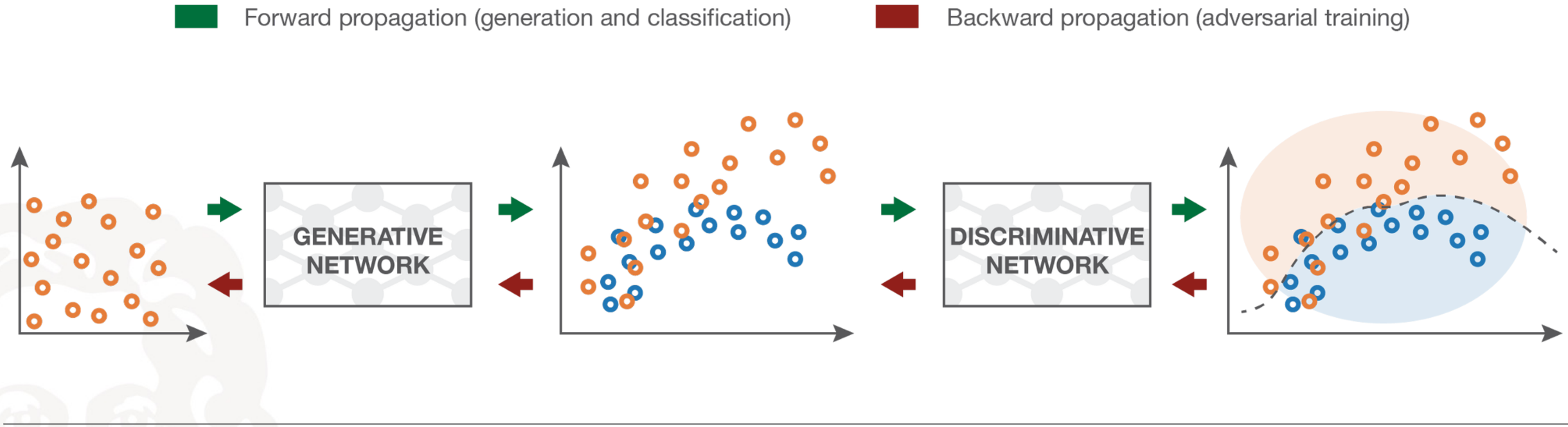
end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

Taken from: https://image.slidesharecdn.com/generativeadversarialnetworks-161121164827/95/generative-adversarial-networks-11-638.jpg?cb=1480242452

From original 2014 NeurIPS paper by Goodfellow et. Al.
https://arxiv.org/pdf/1406.2661.pdf

How do GANs work? Alternating training



Taken from: https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

Variants of GANs

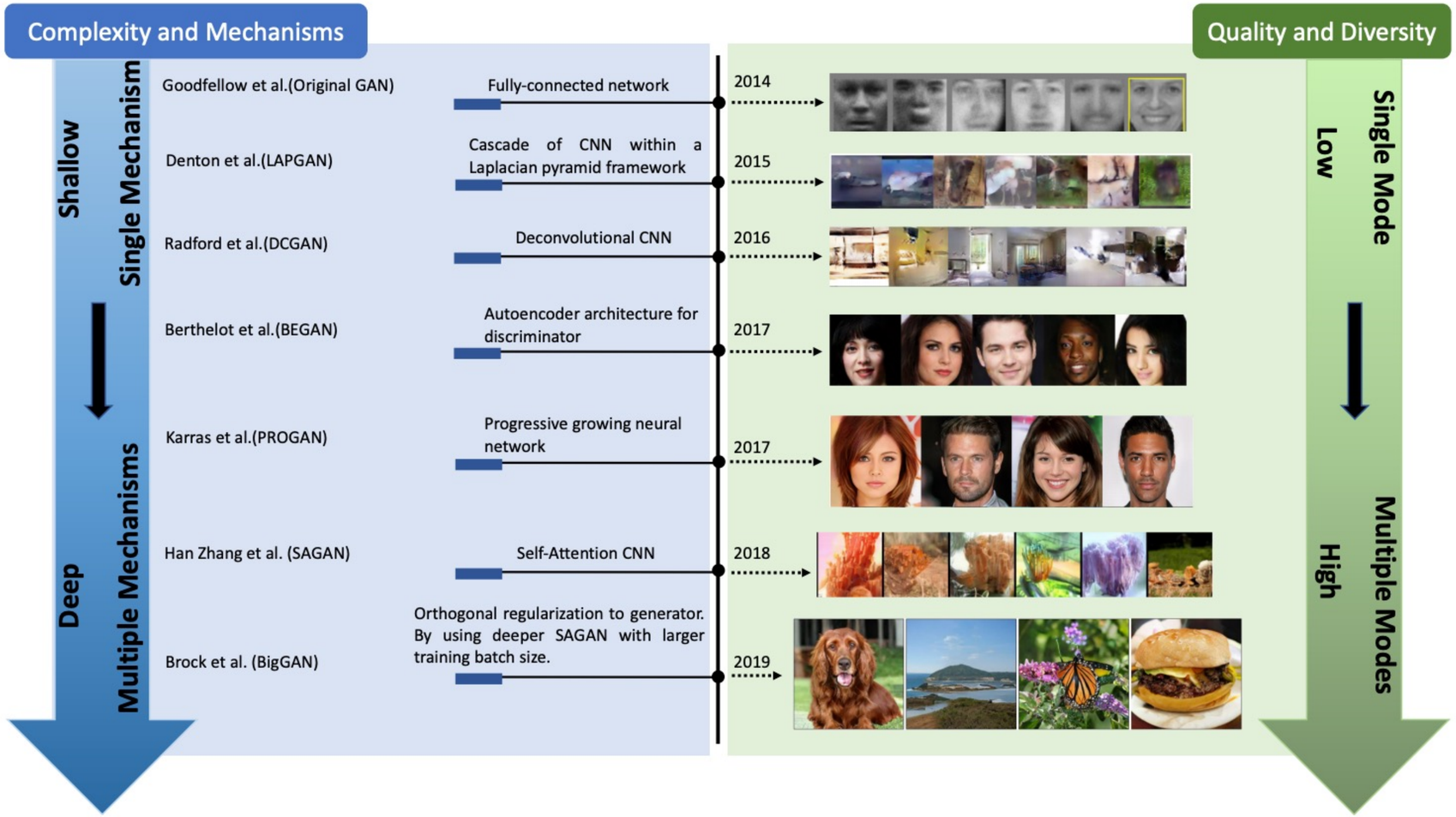


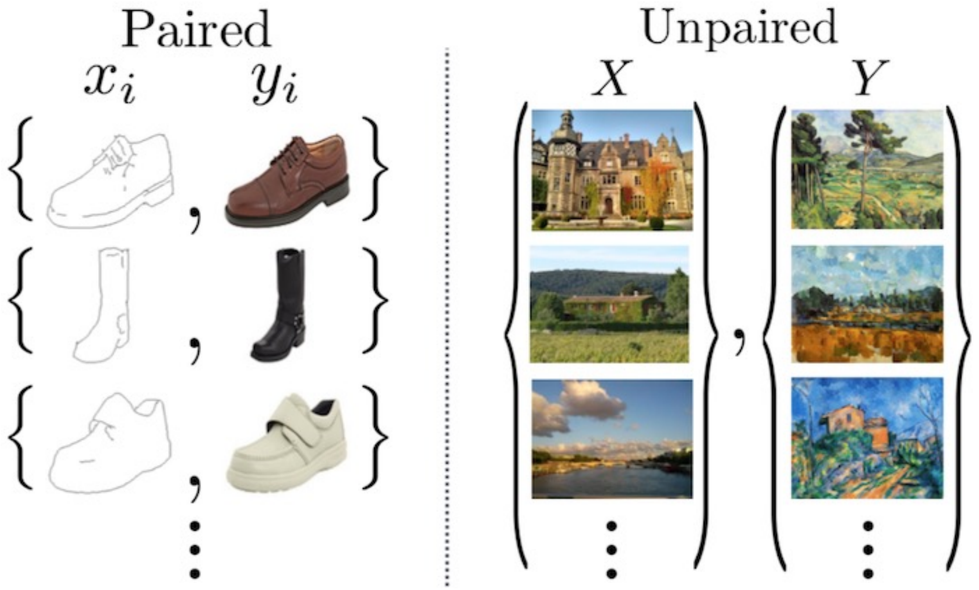
Fig. 2. Timeline of architecture-variant GANs. Complexity in blue stream refers to size of the architecture and computational cost such as batch size. Mechanisms refer to the number of types of models used in the architecture (e.g., BEGAN uses an autoencoder architecture for its discriminator while a deconvolutional neural network is used for the generator. In this case, two mechanisms are used).

Taken from Wang et. al 2019 “Generative Adversarial Networks: A Survey and Taxonomy“: https://arxiv.org/pdf/1906.01529.pdf

GAN examples

Taken from: https://medium.com/datadriveninvestor/432-000-painting-by-ai-sold-at-christies-my-thoughts-4a33cd94f782

Taken from: https://cdn-images-1.medium.com/max/1600/1*5DG4hHjxAyWTfV1J3mRH_A.png



Taken from: https://cdn-images-1.medium.com/max/1200/1*oZsw1JaGkKPxWKKvVUWlyg.png

Taken from: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Mode Collapse in GANs

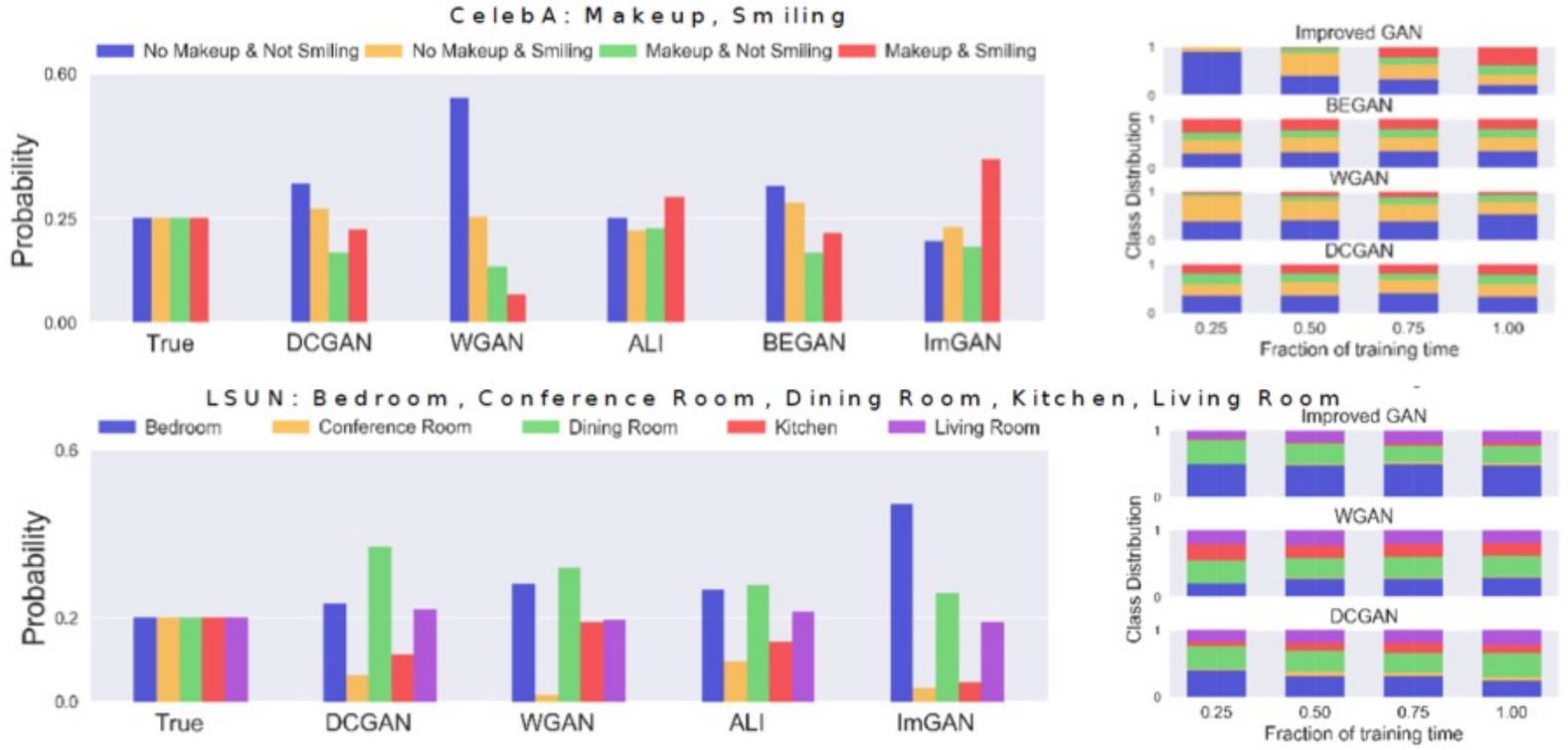


Figure 18: Illustration of mode collapse in GANs trained on select subsets of CelebA and LSUN datasets using the technique in [42]. Left panel shows the relative distribution of modes in samples drawn from the GANs, and compares is to the true data distribution (leftmost

Figure from “Pros and Cons of GAN Evaluation Measures”, Ali Borji, Computer Vision and Image Understanding, Volume 179, p. 51-64, 2019

VAE GAN hybrids: beyond pixel similarities

- If GANs seem to generate prettier examples, does that mean they are practically superior to VAEs? No → mode collapse

- If mode collapse is not a severe issue in VAEs, does that mean they are practically superior to GANs? No → human perceived generated quality

- The key really lies in the metric which is used to measure our unsupervised goal. Recall that in VAEs we typically use a reconstruction loss, whereas the loss in GANs could be seen as more perceptual (only judging fake vs. real independent of precise pixel values)

→ There is several works advancing VAEs and GANs through combination of adversarial training with variational inference!

VAE-GAN

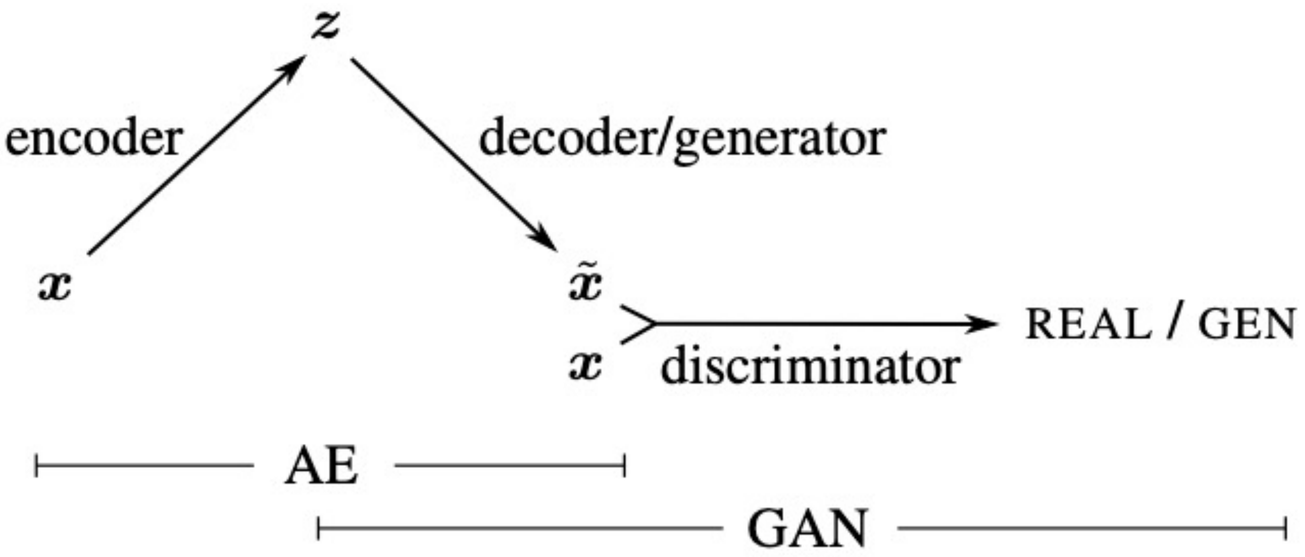


Figure 1. Overview of our network. We combine a VAE with a GAN by collapsing the decoder and the generator into one.

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})}\right] = \mathcal{L}_{\text{llike}}^{\text{pixel}} + \mathcal{L}_{\text{prior}}$$

$$\mathcal{L}_{\text{GAN}} = \log(\text{Dis}(\boldsymbol{x})) + \log(1 - \text{Dis}(\text{Gen}(\boldsymbol{z}))) ,$$

$$\mathcal{L}_{\text{llike}}^{\text{Dis}_l} = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\text{Dis}_l(\boldsymbol{x})|\boldsymbol{z})\right]$$

$$\mathcal{L} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{llike}}^{\text{Dis}_l} + \mathcal{L}_{\text{GAN}}$$

Figure & equations from “Autoencoding beyond pixels using a learned similarity metric”, Larsen et al, ICML 2016

IntroVAE: introspection for adversarial training

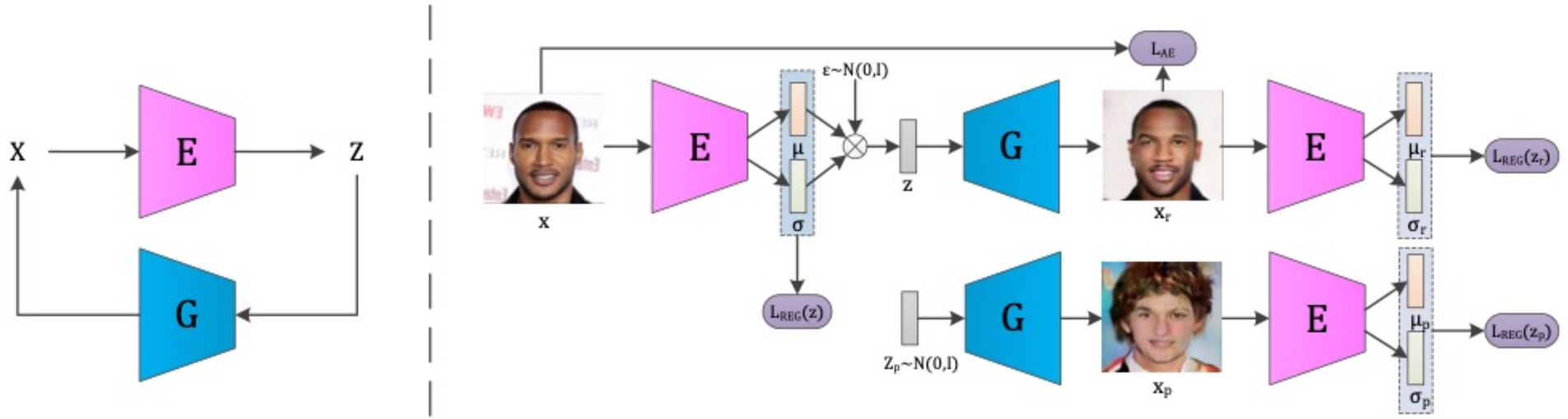


Figure 2: The architecture and training flow of IntroVAE. The left part shows that the model consists of two components, the inference model E and the generator G, in a circulation loop. The right part is the unrolled training flow of the proposed method.

$$L_E = L_{REG}(z) + \alpha \sum_{s=r,p} [m - L_{REG}(z_s)]^+ + \beta L_{AE}(x, x_r)$$

$$L_G = \alpha \sum_{s=r,p} L_{REG}(Enc(x_s)) + \beta L_{AE}(x, x_r)$$

Figure & equations from “IntroVAE: Introspective Variational Auto-Encoders for Photographic Image Synthesis”, Huang et al, NeurIPS 2018

How do we evaluate generative models?

- There unfortunately is no present answer to this question and it depends heavily on the application context

- In a recent paper, 27 metrics have been summarized and commented:
  “Pros and Cons of GAN Evaluation Measures”, Ali Borji, Computer Vision and Image Understanding, Volume 179, Pages 51-64, 2019

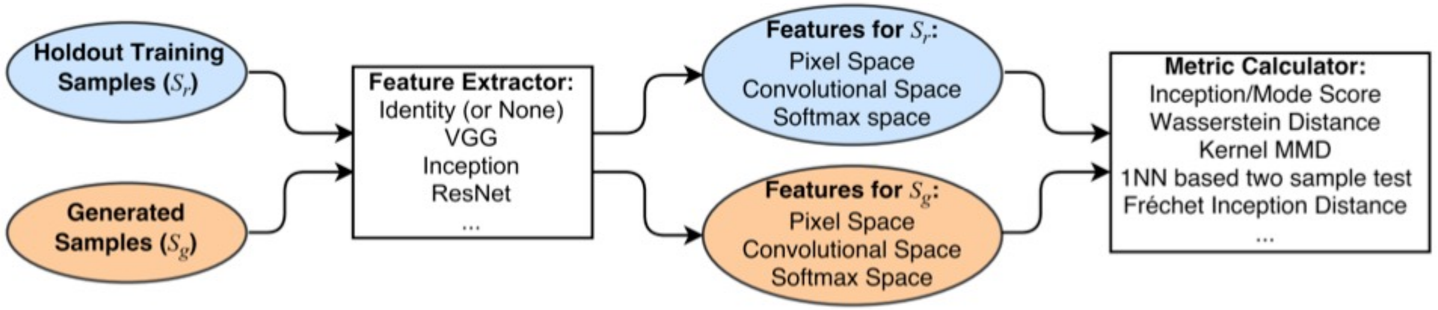


Figure 1: A schematic layout of the typical approach for sample based GAN evaluation. S_r and S_g represent real and generated samples, respectively. Figure from [26].

Figure from “Pros and Cons of GAN Evaluation Measures”, Ali Borji, Computer Vision and Image Understanding, Volume 179, p. 51-64, 2019

How do we evaluate generative models?

- Let's take a look at the formulated desiderata, as formulated by Borji:

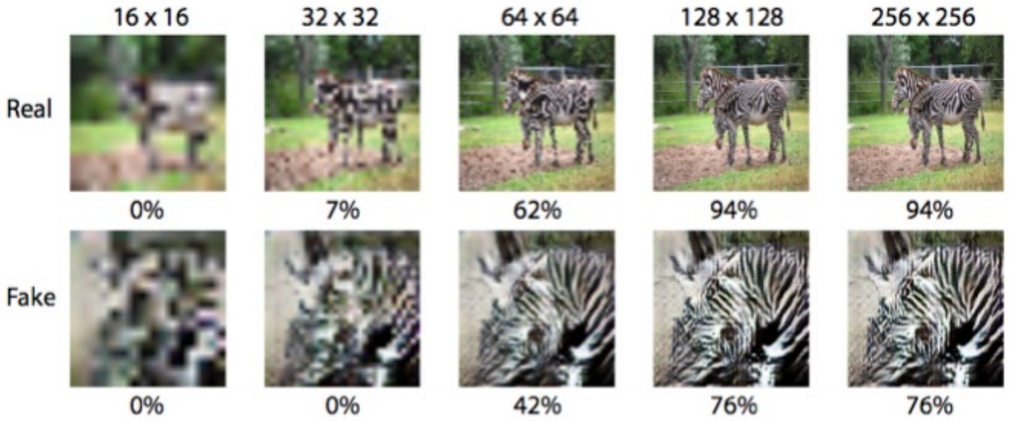
1. favor models that generate high fidelity samples (*i.e.* ability to distinguish generated samples from real ones; discriminability),
2. favor models that generate diverse samples (and thus is sensitive to over-fitting, mode collapse and mode drop, and can undermine trivial models such as the memory GAN),
3. favor models with disentangled latent spaces as well as space continuity (*a.k.a* controllable sampling),
4. have well-defined bounds (lower, upper, and chance),
5. be sensitive to image distortions and transformations. GANs are often applied to image datasets where certain transformations to the input do not change semantic meanings. Thus, an ideal measure should be invariant to such transformations. For instance, score of a generator trained on CelebA face dataset should not change much if its generated faces are shifted by a few pixels or rotated by a small angle.
6. agree with human perceptual judgments and human rankings of models, and
7. have low sample and computational complexity.

Taken from "Pros and Cons of GAN Evaluation Measures", Ali Borji, Computer Vision and Image Understanding, Volume 179, p. 51-64, 2019

How do we evaluate generative models?

- Let's take a look at the 2 most popular metrics, as summarized by Borji:

Inception Score (IS). Proposed by Salimans *et al.* [3], it is perhaps the most widely adopted score for GAN evaluation (*e.g.* in [67]). It uses a pre-trained neural network (the Inception Net [68] trained on the ImageNet [69]) to capture the desirable properties of generated samples: *highly classifiable* and *diverse* with respect to class labels. It measures the average KL divergence between the conditional label distribution $p(y|\mathbf{x})$ of samples (expected to have low entropy for easily classifiable samples; better sample quality) and the marginal distribution $p(y)$ obtained from all the samples (expected to have high entropy if all classes are equally represented in the set of samples; high diversity). It favors low entropy of $p(y|\mathbf{x})$ but a large entropy of $p(y)$.

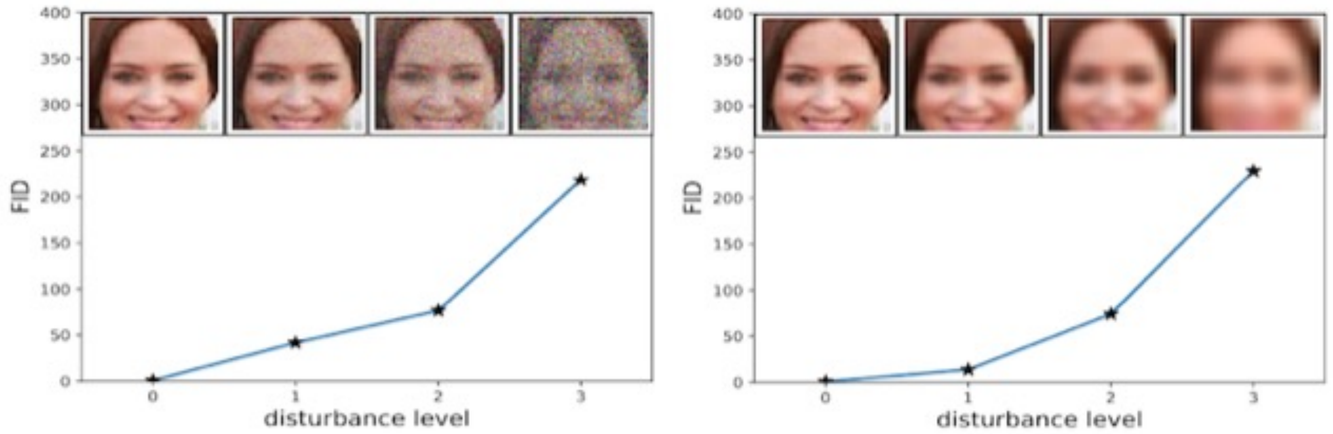


Taken from "Pros and Cons of GAN Evaluation Measures", Ali Borji, Computer Vision and Image Understanding, Volume 179, p. 51-64, 2019

How do we evaluate generative models?

- Let's take a look at the 2 most popular metrics, as summarized by Borji:

Fréchet Inception Distance (FID). Introduced by Heusel *et al.* [37], FID embeds a set of generated samples into a feature space given by a specific layer of Inception Net (or any CNN). Viewing the embedding layer as a continuous multivariate Gaussian, the mean and covariance are estimated for both the generated data and the real data. The Fréchet distance between these two Gaussians (*a.k.a* Wasserstein-2 distance) is then used to quantify the quality of generated samples, *i.e.* ,



Taken from "Pros and Cons of GAN Evaluation Measures", Ali Borji, Computer Vision and Image Understanding, Volume 179, p. 51-64, 2019

How do we evaluate generative models?

- What is the best measure, according to Borji?

(a) only two measures are designed to explicitly address overfitting,
(b) the majority of the measures do not consider disentangled representations,
(c) few measures have both lower and upper bounds,
(d) the agreement between the measures and human perceptual judgments is less clear,
(e) several highly regarded measures have high sample and computational efficiencies, and
(f) the sensitivity of measures to image distortions is less explored.

→ “Seeking appropriate measures for evaluation continues to be an important open problem”

Taken from “Pros and Cons of GAN Evaluation Measures”, Ali Borji, Computer Vision and Image Understanding, Volume 179, p. 51-64, 2019