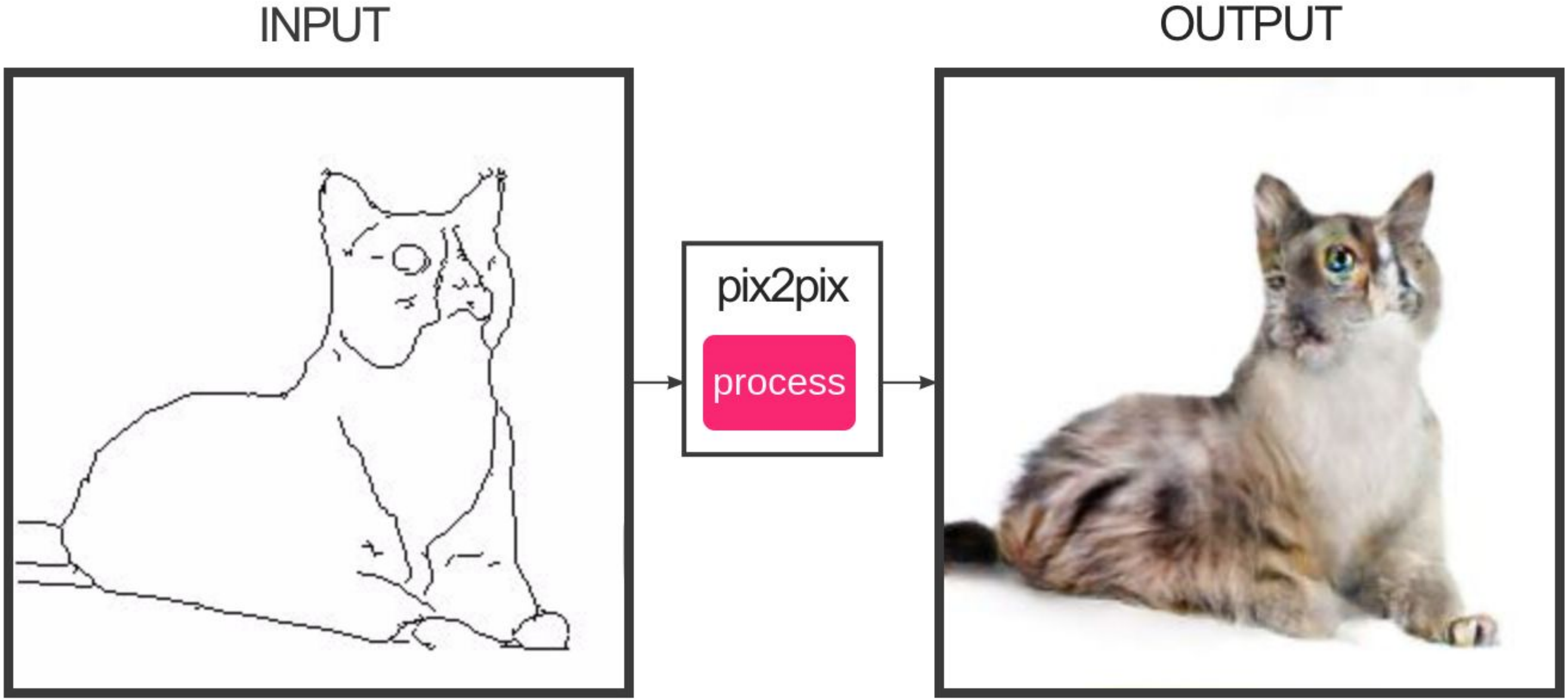
Generative Adversarial Networks

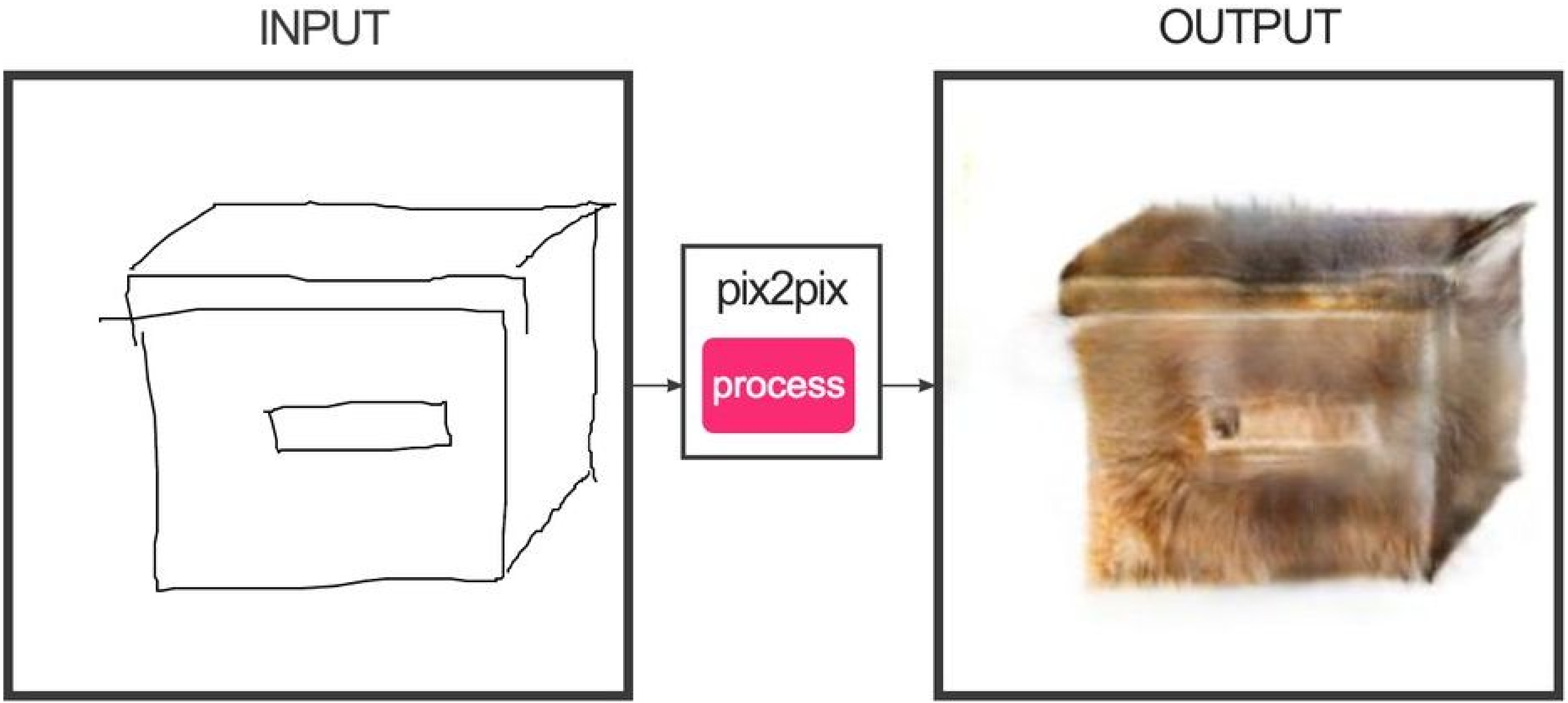
Michal Šustr

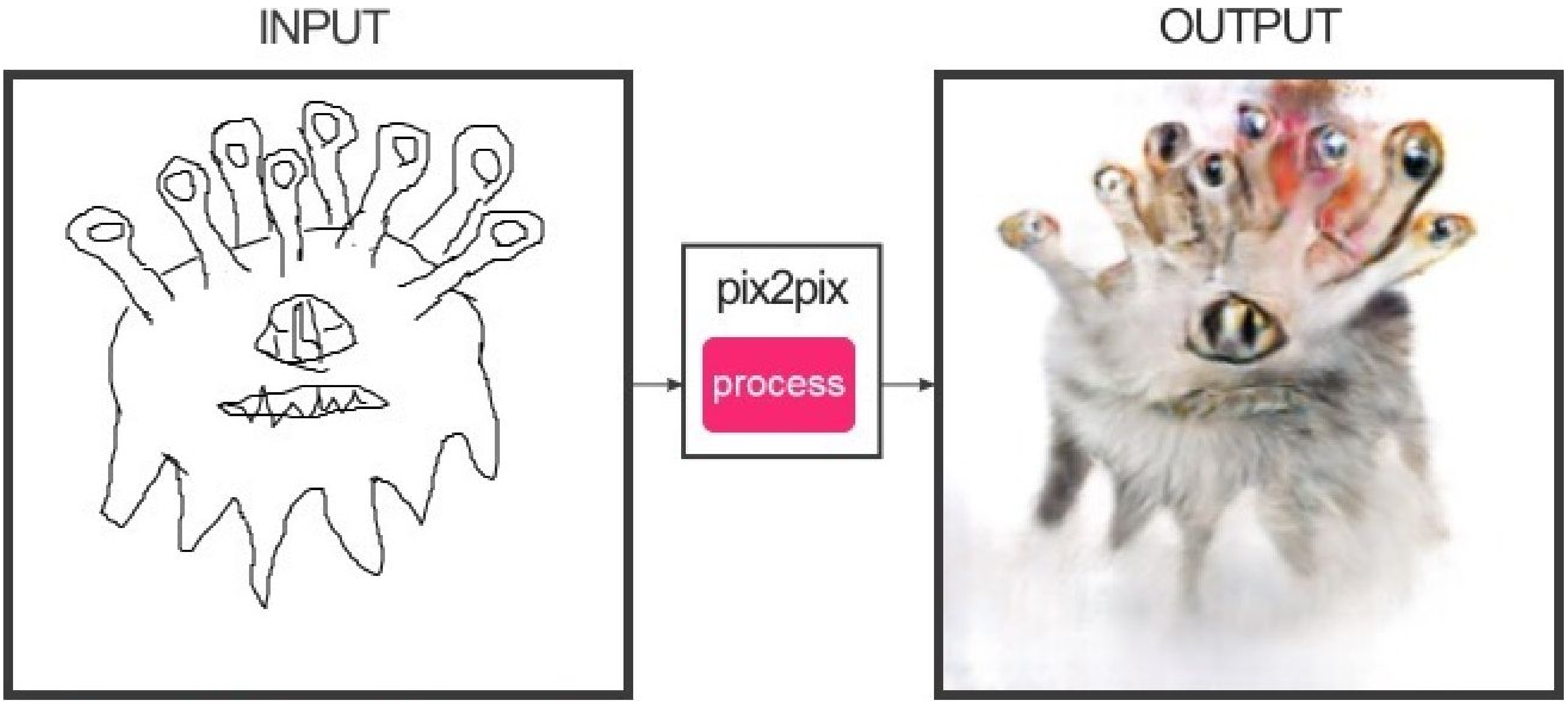
michal.sustr@gmail.com
http://michal.sustr.sk
http://lectures.ai/gan_mlmu
http://twitter.com/michal_sustr

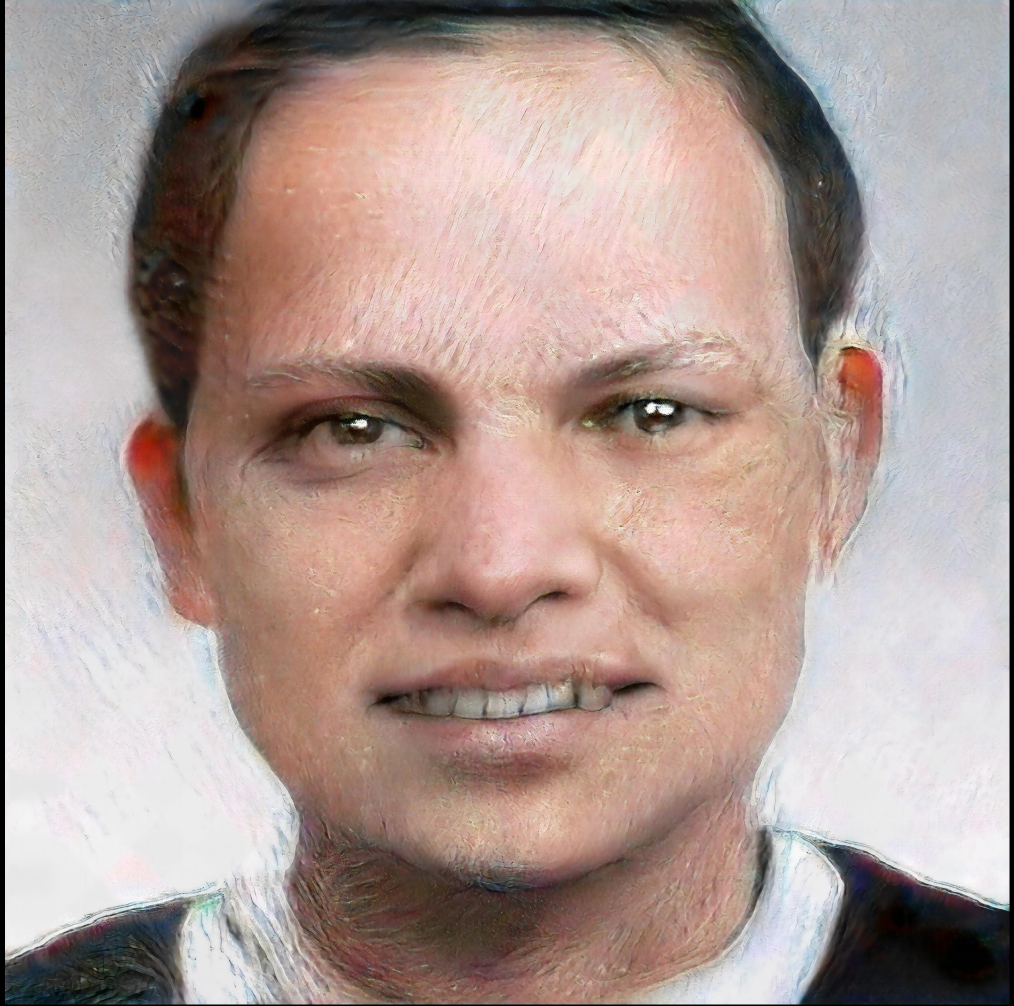
MLMU, Prague, March 22nd 2017

Contents


INPUT
pix2pix
process
OUTPUT


INPUT
pix2pix
process
OUTPUT


INPUT
pix2pix
process
OUTPUT



http://mtyka.github.io/machine/learning/2017/06/06/highres-gan-faces.html

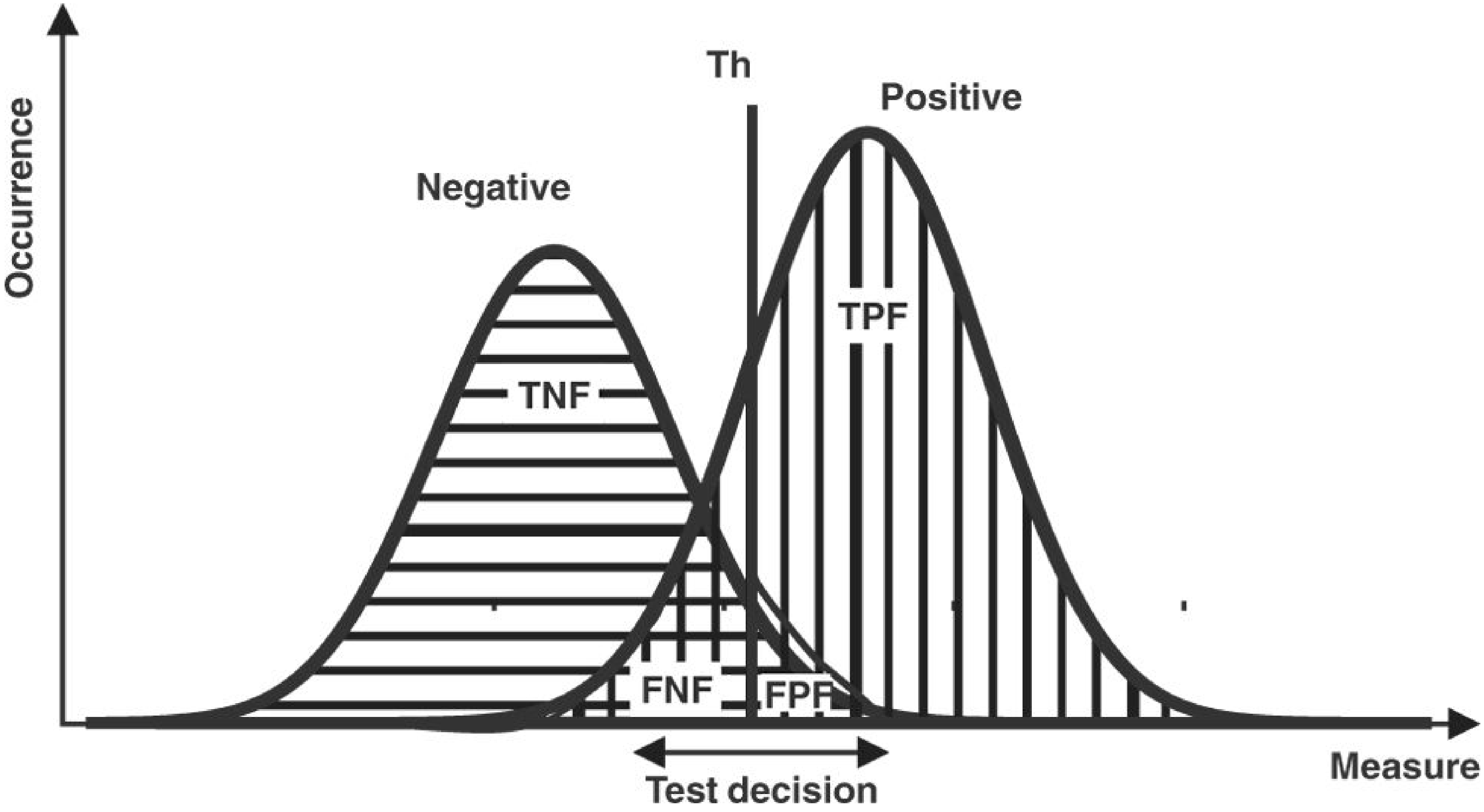
Unsupervised learning

The problem this paper is concerned with is that of unsupervised learning. Mainly, what does it mean to learn a probability distribution? The classical answer to this is to learn a probability density. This is often done by defining a parametric family of densities $(P_\theta)_{\theta \in \mathbb{R}^d}$ and finding the one that maximized the likelihood on our data: if we have real data examples $\{x^{(i)}\}_{i=1}^m$, we would solve the problem

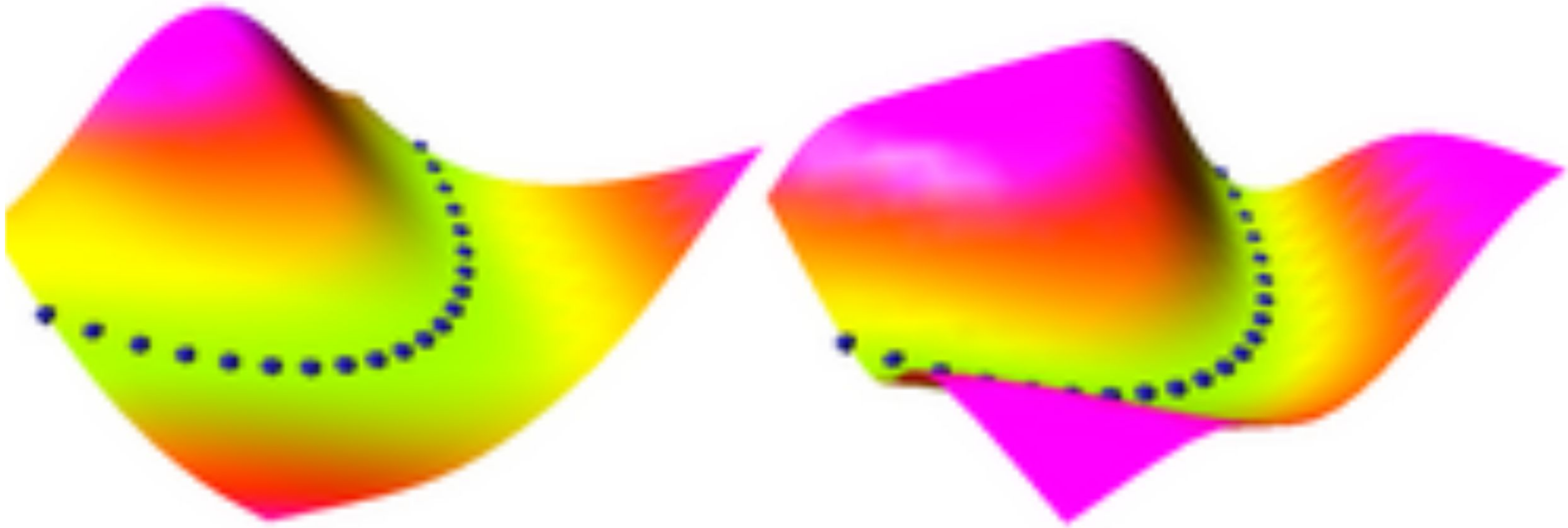
$$\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$

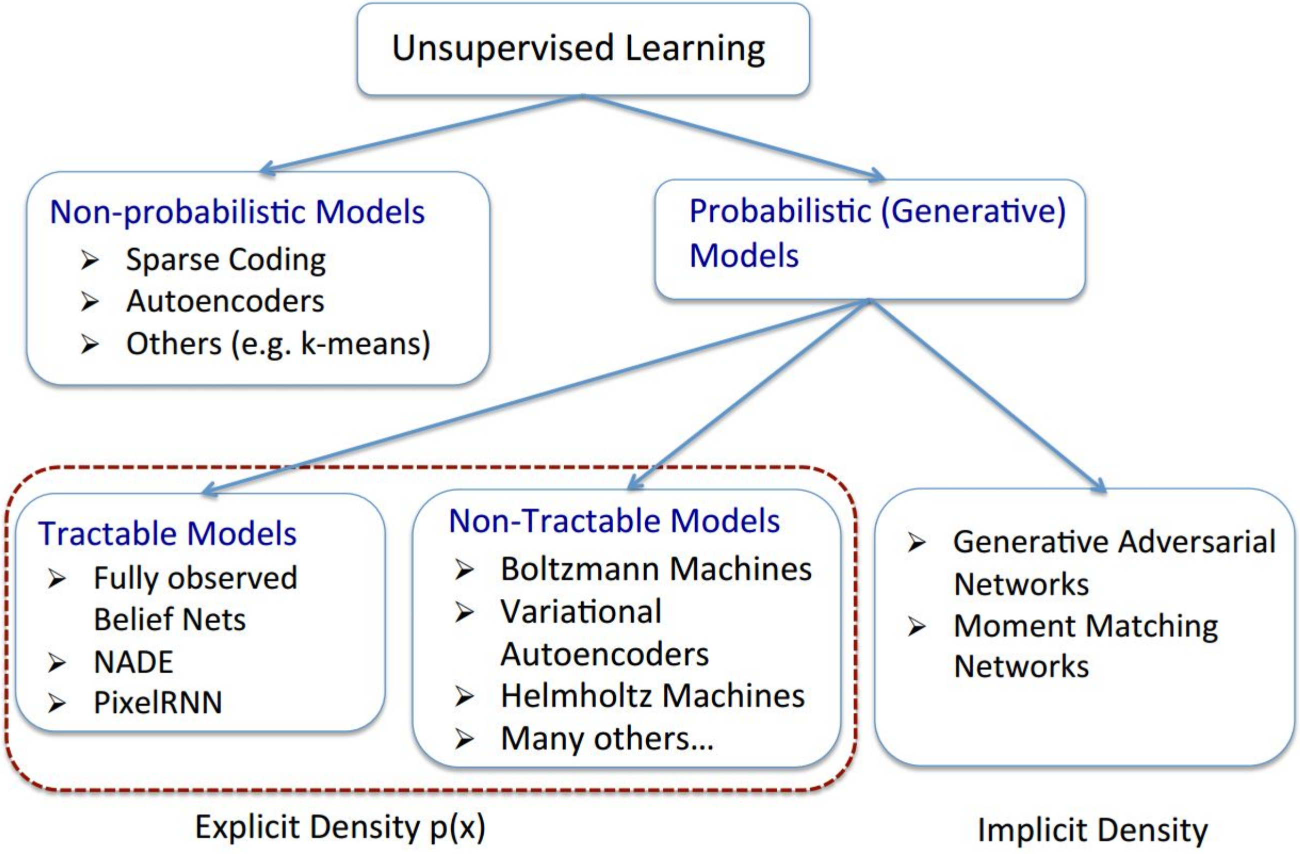
Wasserstein GAN, 2016

If we have PDFs:



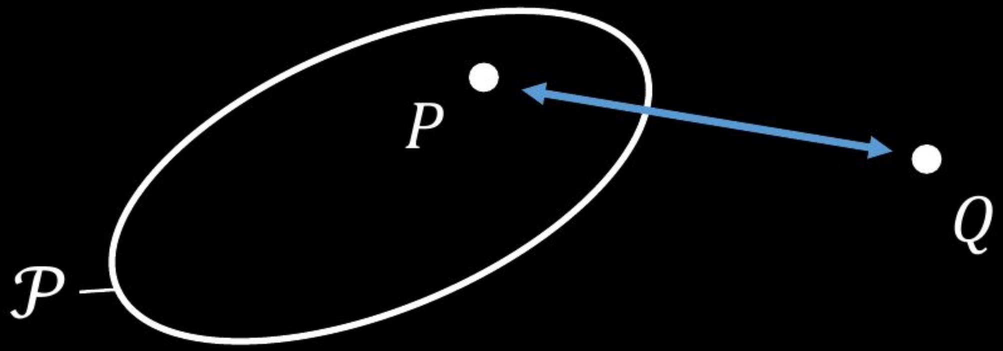
Manifold hypothesis





Salakhutdinov slides

Learning Probabilistic Models

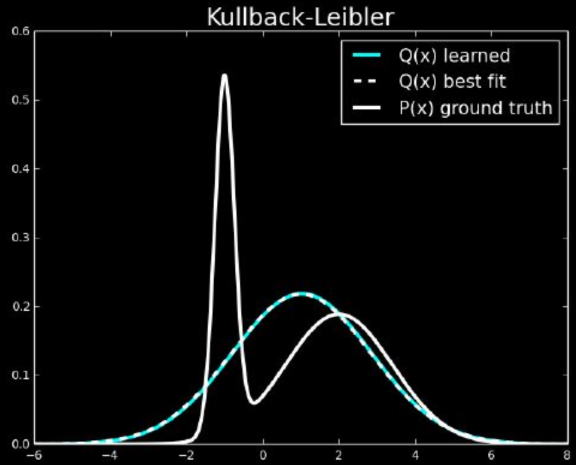


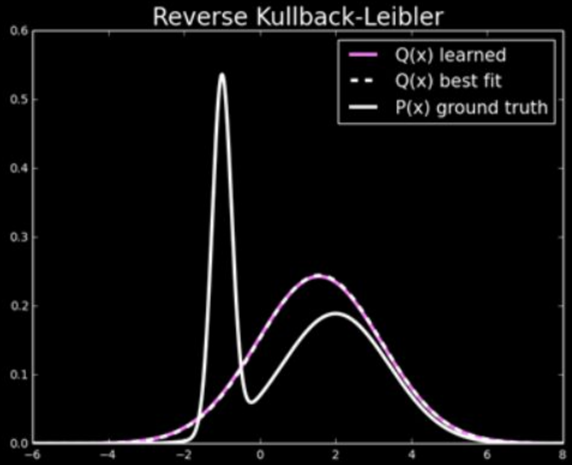
Kullback-Leibler divergence:

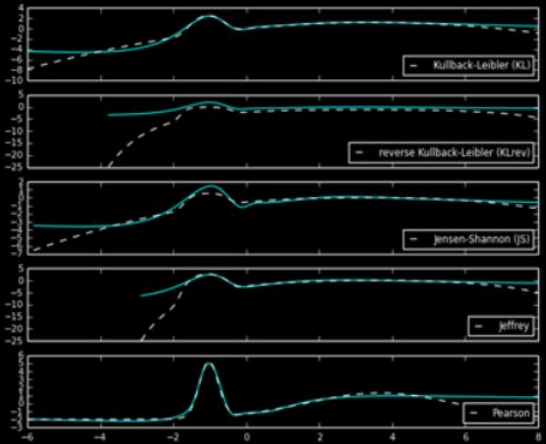
$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

DEMO

Synthetic 1D Univariate



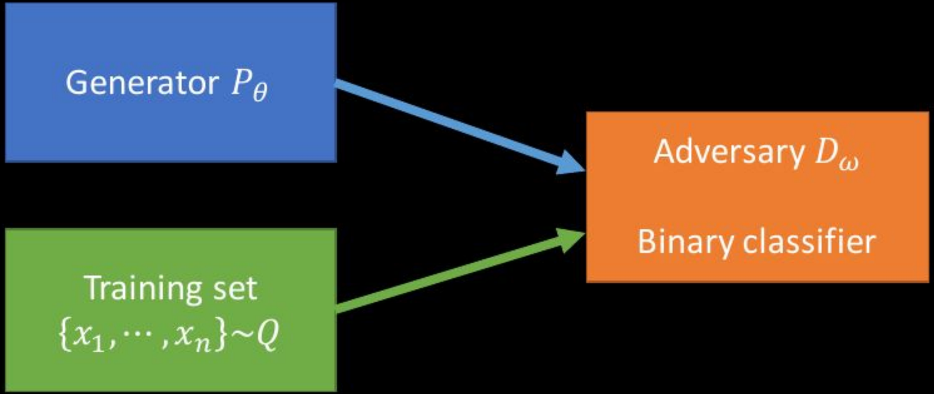




	KL	KL-rev	JS	Jeffrey	Pearson
$D_f(P\|\|Q_{\theta^*})$	0.2831	0.2480	0.1280	0.5705	0.6457
$F(\hat{\omega}, \hat{\theta})$	0.2801	0.2415	0.1226	0.5151	0.6379
μ^*	1.0100	1.5782	1.3070	1.3218	0.5737
$\hat{\mu}$	1.0335	1.5624	1.2854	1.2295	0.6157
σ^*	1.8308	1.6319	1.7542	1.7034	1.9274
$\hat{\sigma}$	1.8236	1.6403	1.7659	1.8087	1.9031

Nowozin slides

GAN Training Objective [Goodfellow et al., 2014]



$$\min_{\theta} \max_{\omega} \mathbb{E}_{x\sim Q}[\log D_\omega(x)] + \mathbb{E}_{x\sim P_\theta}[\log(1 - D_\omega(x))]$$

- Generator tries to fool discriminator (i.e. generate realistic samples)
- Discriminator tries to distinguish fake from real samples
- Saddle-point problem

Nowozin slides

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:

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

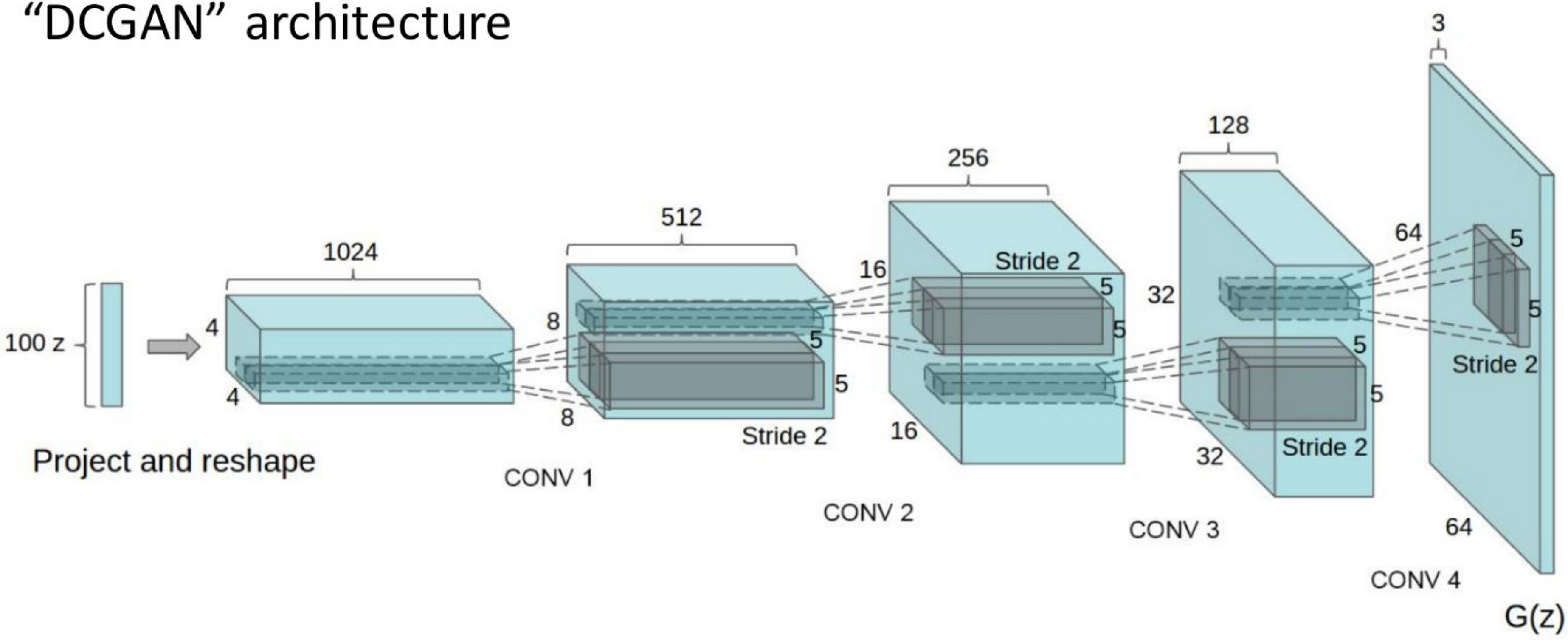
end for

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

Goodfellow 2014

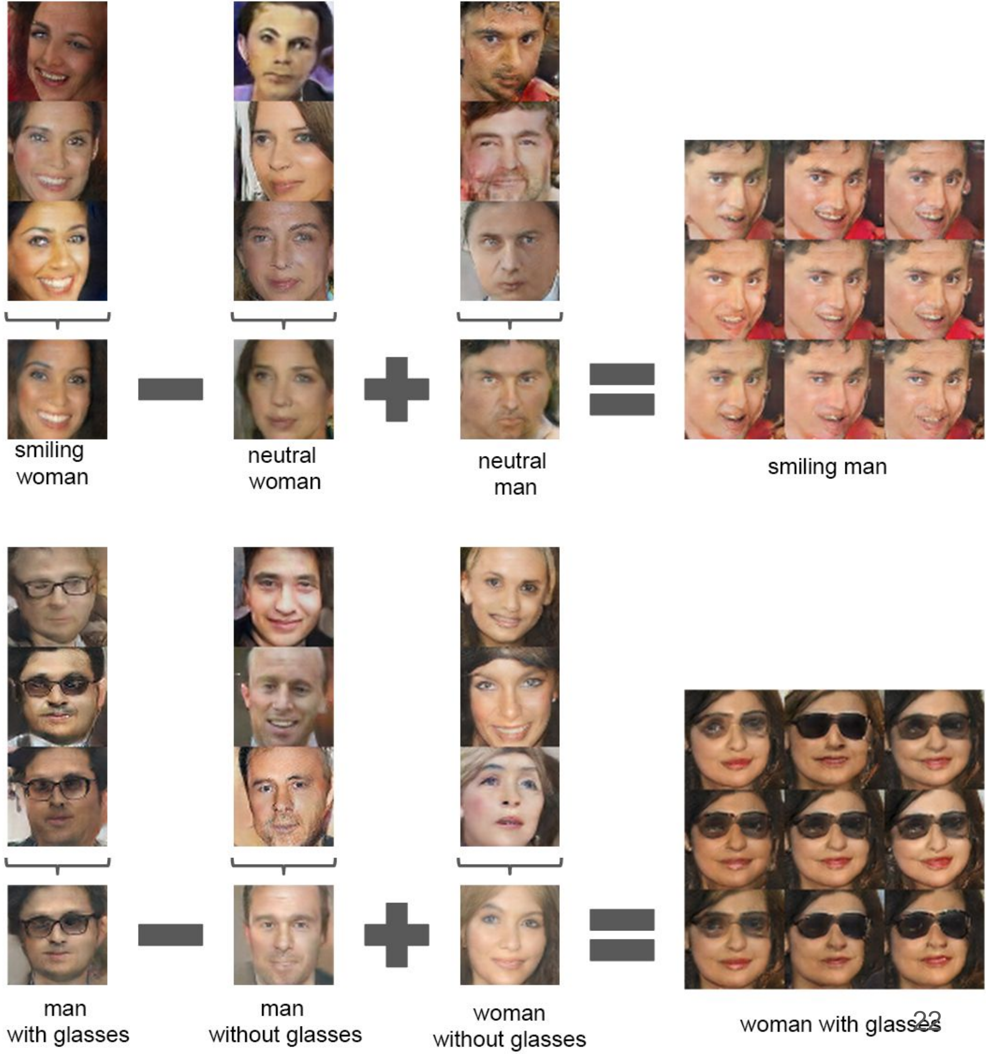
DEMO

“DCGAN” architecture



Groundtruth MNIST GAN DCGAN (ours)

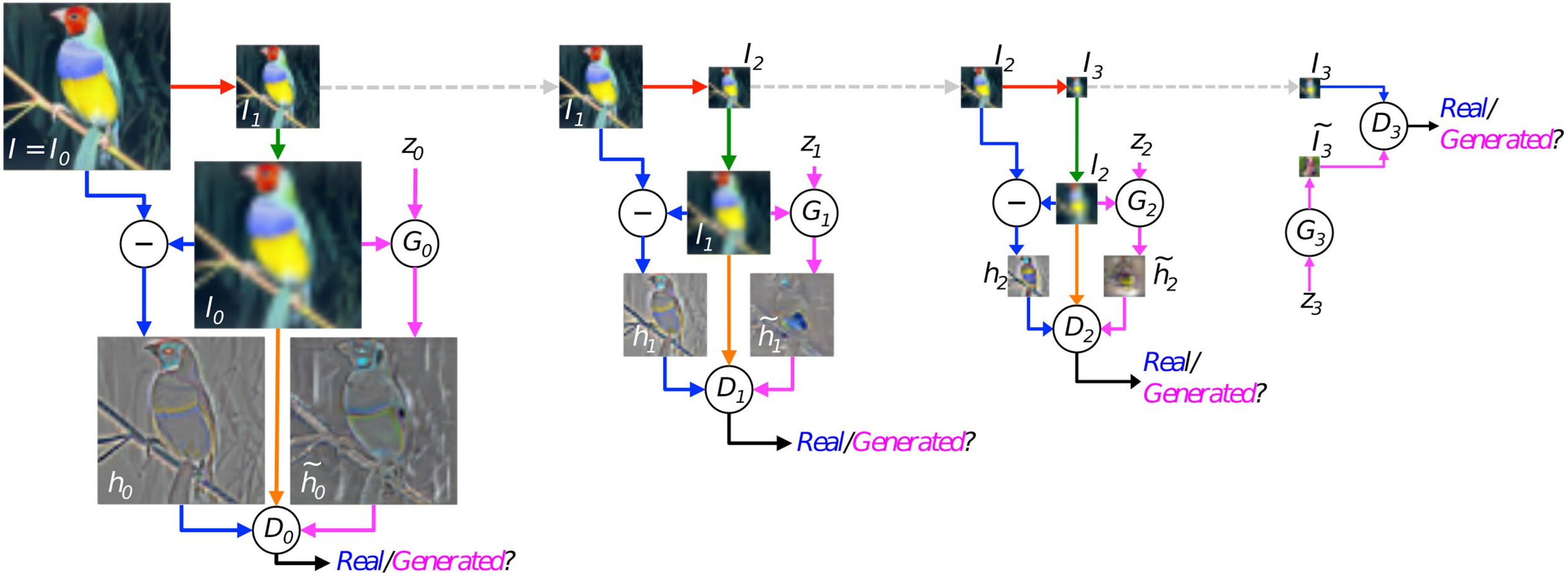
Arithmetics in latent space:



Face rotation is linear interpolation between latent variables:

Generated album covers:

LAPGAN (Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks)



Problems with GANs:

- Instability in training (small gradients)
- Mode collapse
- Qualitative problems: counting, perspective, global structure

Tricks for training: https://github.com/soumith/ganhacks

1. Normalize the inputs
2. A modified loss function
3. Use a spherical Z
4. BatchNorm
5. Avoid Sparse Gradients. ReLU, MaxPool
6. … and more, see link.

Problems with Counting



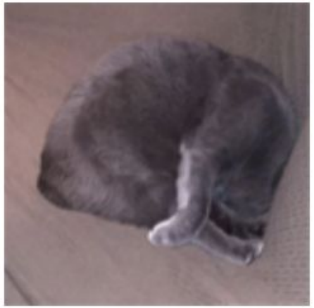



(Goodfellow 2016)

Problems with Perspective







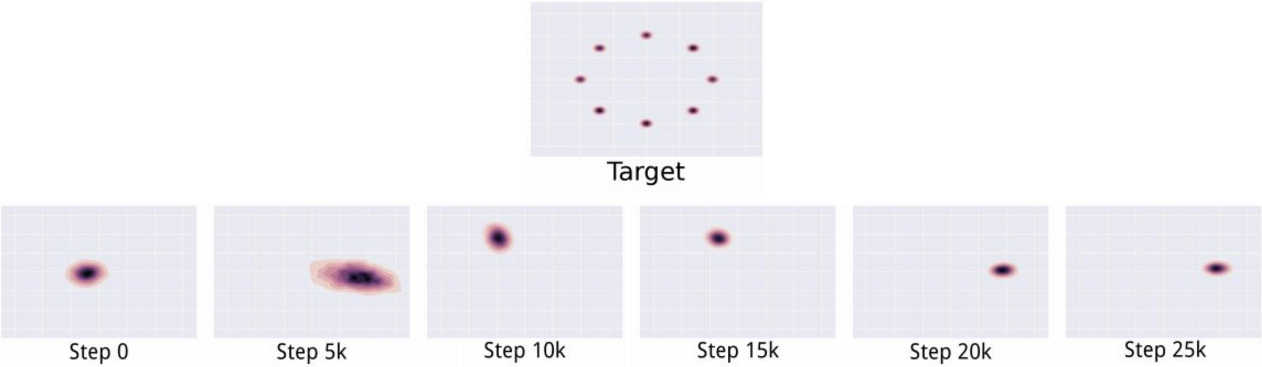
Problems with Global Structure

(Goodfellow 2016)

Mode Collapse

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

- D in inner loop: convergence to correct distribution
- G in inner loop: place all mass on most likely point

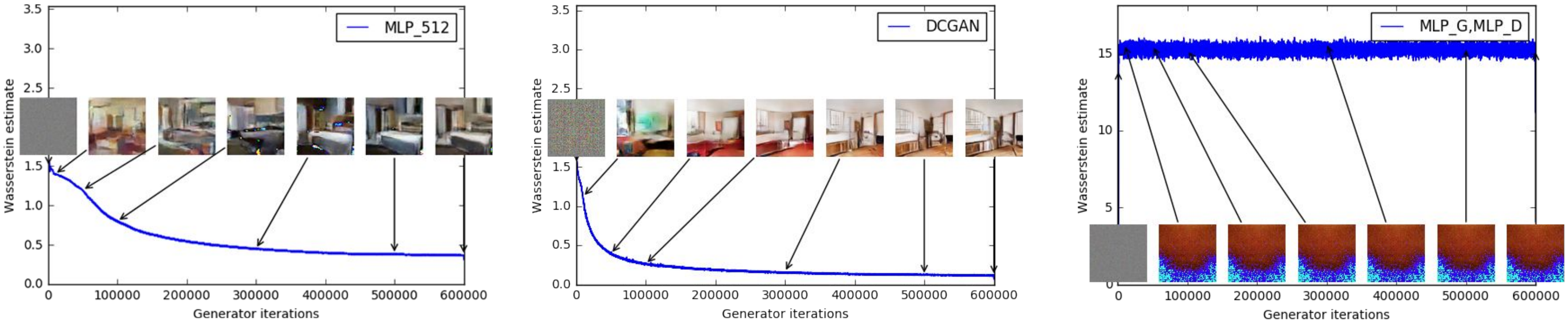


(Metz et al 2016)

(Goodfellow 2016)

Wasserstein GAN

Generator sample quality correlates with discriminator loss



Wasserstein metric

$$W_p(\mu,\nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p},$$

- Also known as Earth mover's distance
- Moving “pile of dirt” analogy
- Comparing discrete distributions, e.g. the color histograms of two digital images

Example 1 (Learning parallel lines). Let $Z \sim U[0,1]$ the uniform distribution on the unit interval. Let $\mathbb{P}_0$ be the distribution of $(0, Z) \in \mathbb{R}^2$ (a 0 on the x-axis and the random variable Z on the y-axis), uniform on a straight vertical line passing through the origin. Now let $g_\theta(z) = (\theta, z)$ with θ a single real parameter. It is easy to see that in this case,

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$,
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ . \end{cases}$

When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t \in \mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL, or TV divergences.

Wasserstein metric

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x - y\| \,\big]$$

W is equivalent to

(Kantorovich-Rubinstein duality)

$$W(P_r, P_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim P_r}[f(x)] - \mathbb{E}_{x\sim P_\theta}[f(x)]$$

K-Lipschitz function

$$d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2)$$

Supremum still intractable, but now it's easier to approximate:

Parametrized function family $\{f_w\}_{w \in \mathcal{W}}$,

w are weights, *W* are all possible weights, let's suppose they are K-Lipschitz

$$\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim P_r}[f_w(x)] - \mathbb{E}_{x \sim P_\theta}[f_w(x)] \leq \sup_{\|f\|_L \leq K} \mathbb{E}_{x \sim P_r}[f(x)] - \mathbb{E}_{x \sim P_\theta}[f(x)] = K \cdot W(P_r, P_\theta)$$

K-Lipschitz critic? Weight clipping!

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.

Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**

  for $t = 0, ..., n_{\text{critic}}$ **do**

    Sample $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$ a batch from the real data.

    Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.

    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))\right]$

    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$

    $w \leftarrow \text{clip}(w, -c, c)$

  end for

  Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.

  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))$

  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$

end while

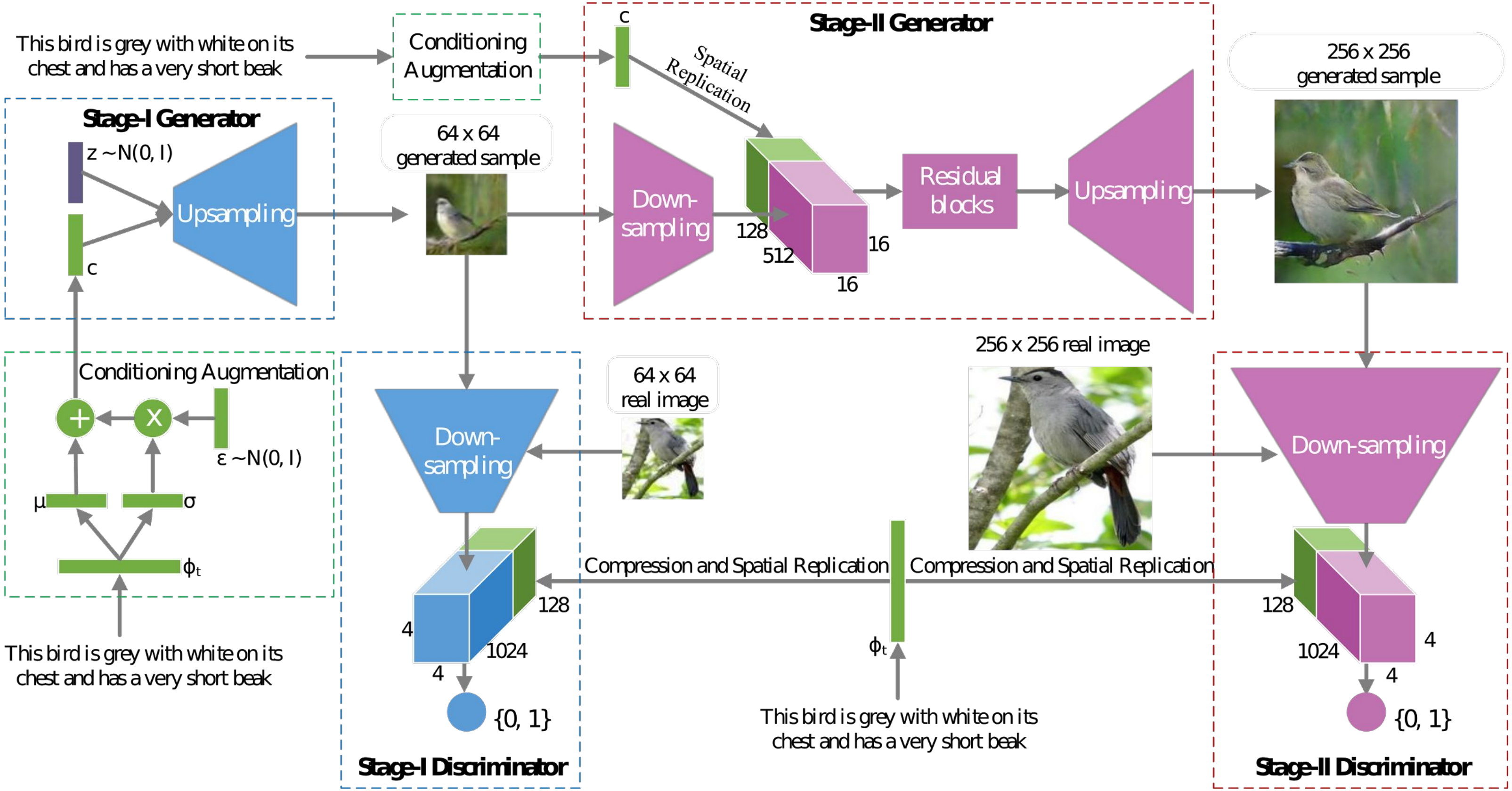
Value of c?

- low c trains more reliably, but high c trains faster when it does work

Read-through: http://www.alexirpan.com/2017/02/22/wasserstein-gan.html

Original paper: https://arxiv.org/abs/1701.07875

StackGAN



This bird is grey with white on its chest and has a very short beak
Conditioning Augmentation
Stage-II Generator
c
Spatial Replication
256 x 256 generated sample
Stage-I Generator
z ~N(0, I)
Upsampling
c
64 x 64 generated sample
Down-sampling
128
512
16
16
Residual blocks
Upsampling
Conditioning Augmentation
ε ~N(0, I)
μ
σ
ϕt
This bird is grey with white on its chest and has a very short beak
Down-sampling
64 x 64 real image
256 x 256 real image
Down-sampling
Compression and Spatial Replication
Compression and Spatial Replication
128
4
1024
4
{0, 1}
Stage-I Discriminator
ϕt
This bird is grey with white on its chest and has a very short beak
128
1024
4
4
{0, 1}
Stage-II Discriminator

A small yellow bird with a black crown and a short black pointed beak

Stage-I

Stage-II

A white bird with a black crown and yellow beak

Stage-I

Stage-II

This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments

Stage-I

Stage-II

This flower has long thin yellow petals and a lot of yellow anthers in the center

Stage-I

Stage-II

Nearest neighbours:

Images generated from text in test sets

Five nearest neighbors from training sets

Bird linear interpolation in latent space:

The bird is completly red → The bird is completly yellow

This bird is completely red with black wings and pointy beak →
this small blue bird has a short pointy beak and brown on its wings

This bird is completely red with black wings and pointy beak →
The bird has a yellow breast with grey features and a small beak

Small mode collapse

This small blue bird has a short pointy beak and brown on its wings

This bird is completely red with black wings and pointy beak

A small sized bird that has a cream belly and a short pointed bill

A small bird with a black head and wings and features grey wings

Failures

Text description						
This particular bird has a brown body and brown bill	Grey bird with black flat beak with grey and white big wings	Bird has brown body feathers, brown breast feathers, and brown beak	The medium sized bird has a dark grey color, a black downward curved beak, and long wings	Colored bill with a white ring around it on the upper part near the bill	This bird has a dark brown overall body color, with a small white patch around the base of the bill	This medium sized bird is primarily black and has a large wingspan and a long black bill with a strip of white at the beginning of it

Stage-I images

Stage-II images

Failures

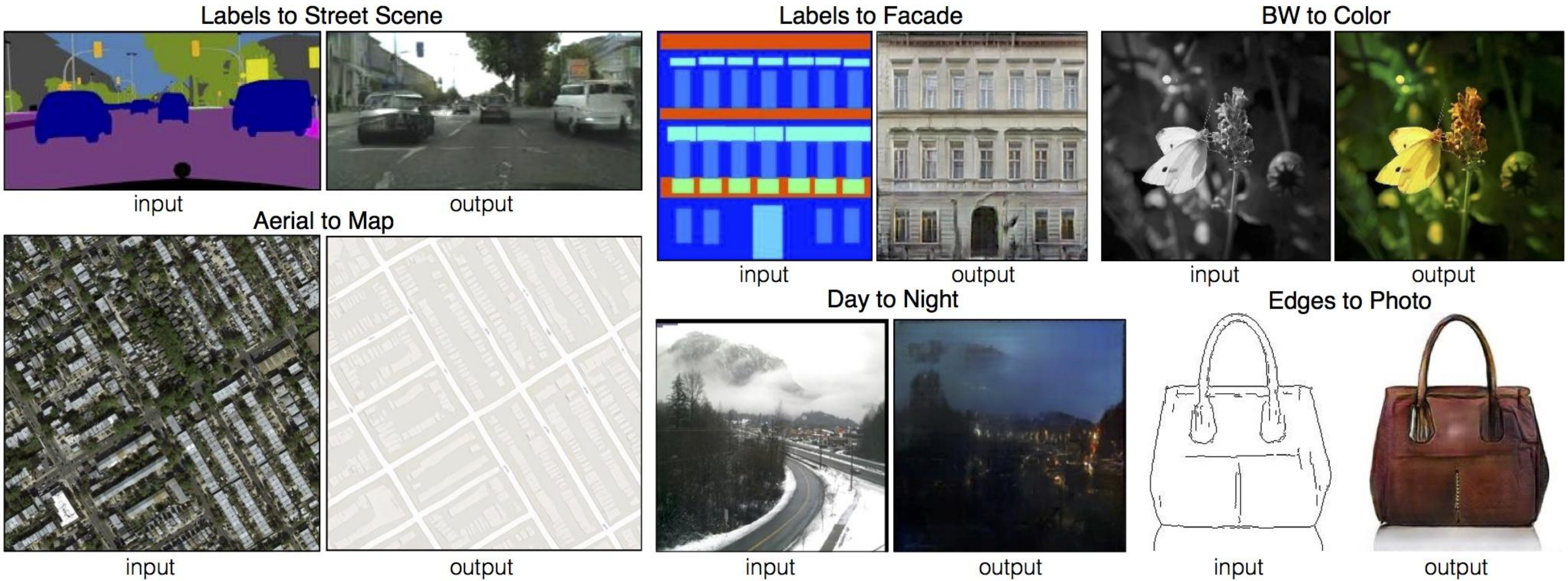
Oxford-102 failure cases:

Text description	The petals of this flower are white with a large stigma	A unique yellow flower with no visible pistils protruding from the center	This flower is pink and yellow in color, with petals that are oddly shaped	This is a light colored flower with many different petals on a green stem	This flower is yellow and green in color, with petals that are ruffled	The flower have large petals that are pink with yellow on some of the petals	A flower that has white petals with some tones of yellow and green filaments

Stage-I images

Stage-II images

Pix2pix (Image-to-Image Translation with Conditional Adversarial Nets)



Sources (images, slides)

- Workshop NIPS 2016:
  https://sites.google.com/site/nips2016adversarial/
- Workshop FB group:
  https://www.facebook.com/groups/675606912596390/
- Goodfellow slides & GAN paper:
  http://www.iangoodfellow.com/slides/2016-12-04-NIPS.pdf
  https://arxiv.org/abs/1701.00160
- StackGAN:
  https://arxiv.org/abs/1612.03242
- Pix2pix (Image-to-Image Translation with Conditional Adversarial Nets)
  https://phillipi.github.io/pix2pix/
- Wasserstein GAN
  https://arxiv.org/abs/1701.07875
- LAPGAN
  https://arxiv.org/abs/1506.05751
- DCGAN (unsupervised representation learning with deep convolutional generative adversarial networks)
  https://arxiv.org/abs/1511.06434
- Nowozin, slides from NIPS workshop
  http://www.nowozin.net/sebastian/blog/nips-2016-generative-adversarial-training-workshop-talk.html
- Salakhutdinov slides
  https://t.co/sxJzyovEvq

Thank You

For Your Attention

Slides at: http://lectures.ai/gan_mlmu

http://twitter.com/michal_sustr