Amsterdam UMC
University Medical Centers

Deep Learning II: Interpretability & Adversarial Methods

8DM40 Machine Learning in Medical Imaging and Biology

Jelmer Wolterink

02-10-2019

Deep Learning II

Me

-Postdoctoral researcher @ Amsterdam UMC – Location AMC

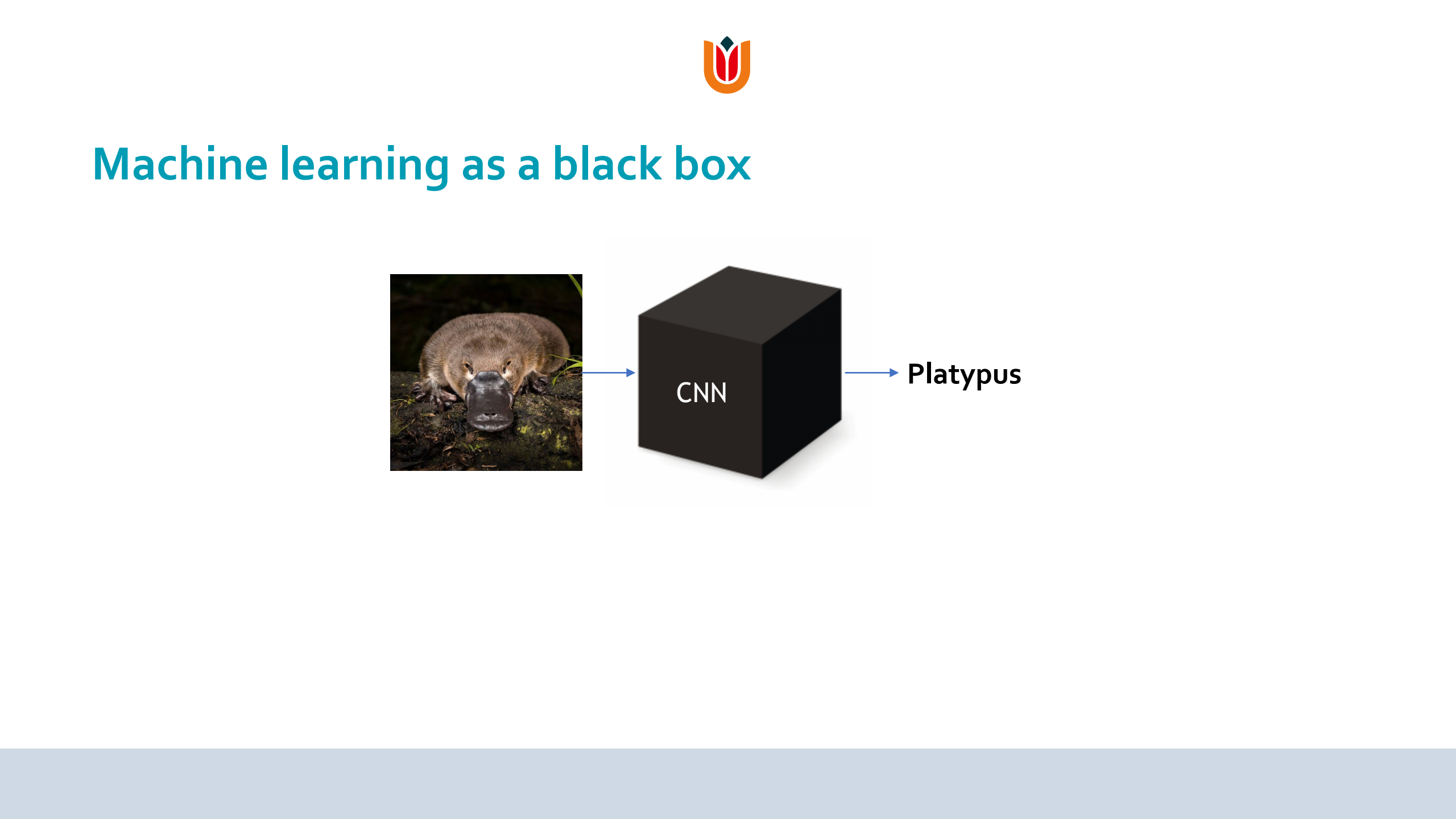
-Deep learning for cardiovascular image analysis (CT, MR)

Today

1. Advanced neural network architectures

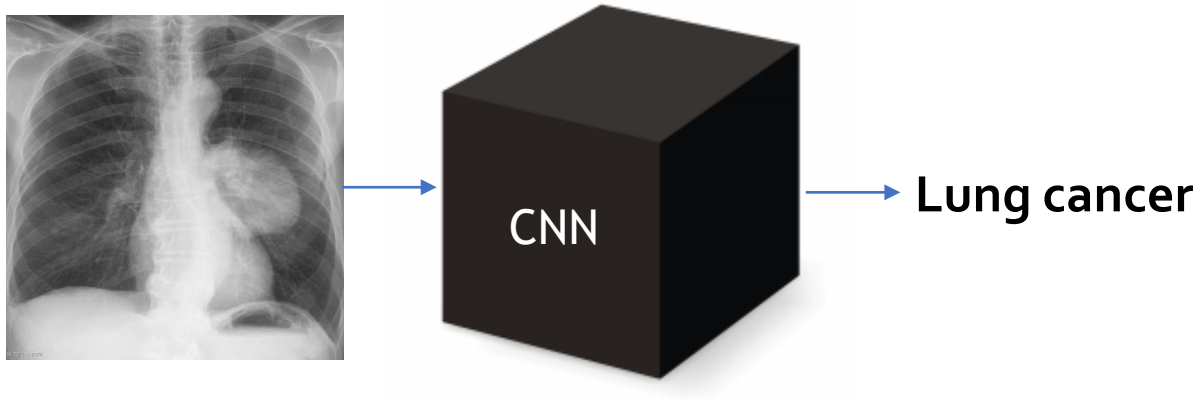
2. **Interpretability and generative adversarial networks**

3. Practical assignment in Keras



Machine learning as a black box

Machine learning as a black box



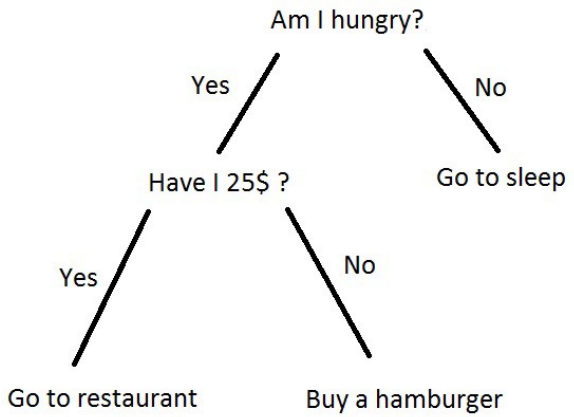
Interpretability: Can we predict what the model will do if the input changes?

Explainability: Do we know exactly what the model is doing?

Uncertainty: How sure is the network of its prediction?

Important questions for ethical use and liability

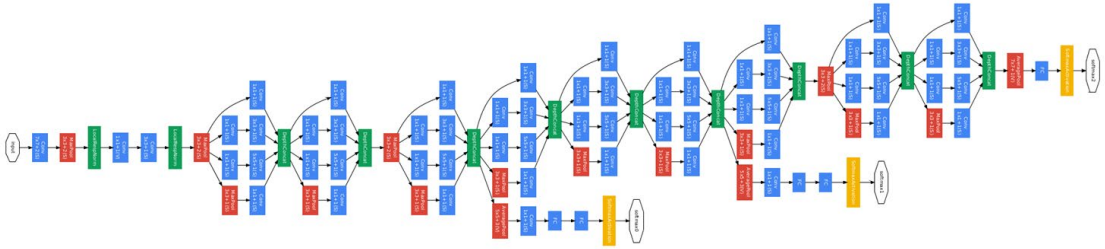
Complexity/performance trade-off



Decision tree

Explainable/interpretable

Performance



CNN

Explainable/interpretable

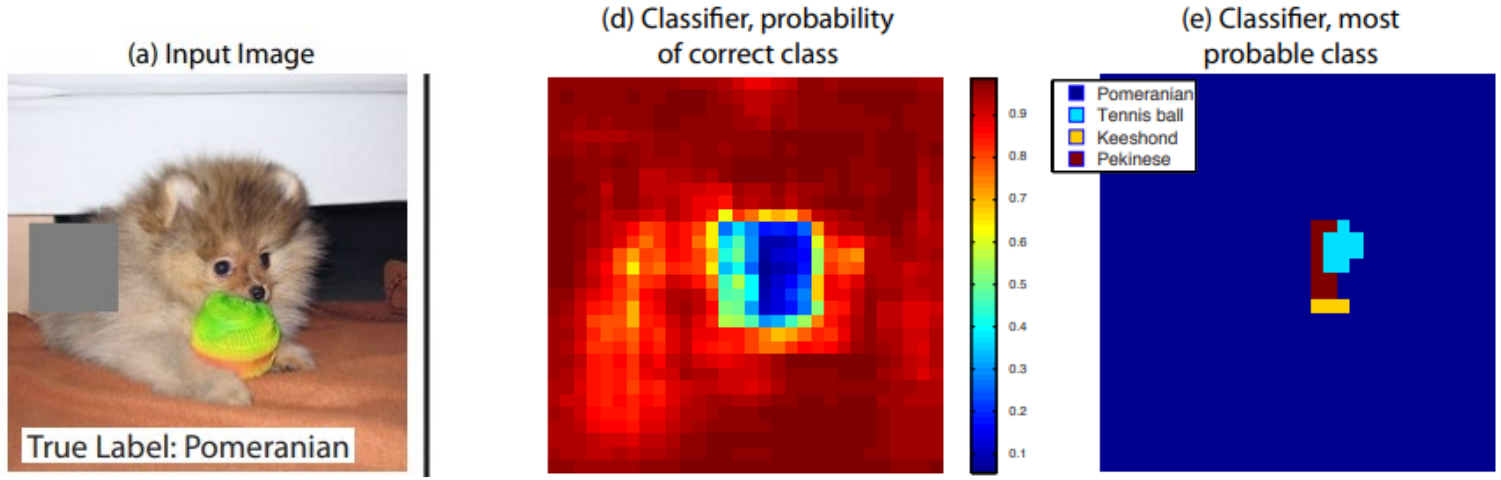
Performance

Interpreting CNNs: saliency maps

- Millions of parameters, impossible to know in detail what each parameter does
- Substitute: what part of the input is associated with a CNN feature or output?
- Popular approaches
  - Occlusion
  - Deconvolution
  - Class activation mapping

Occlusion

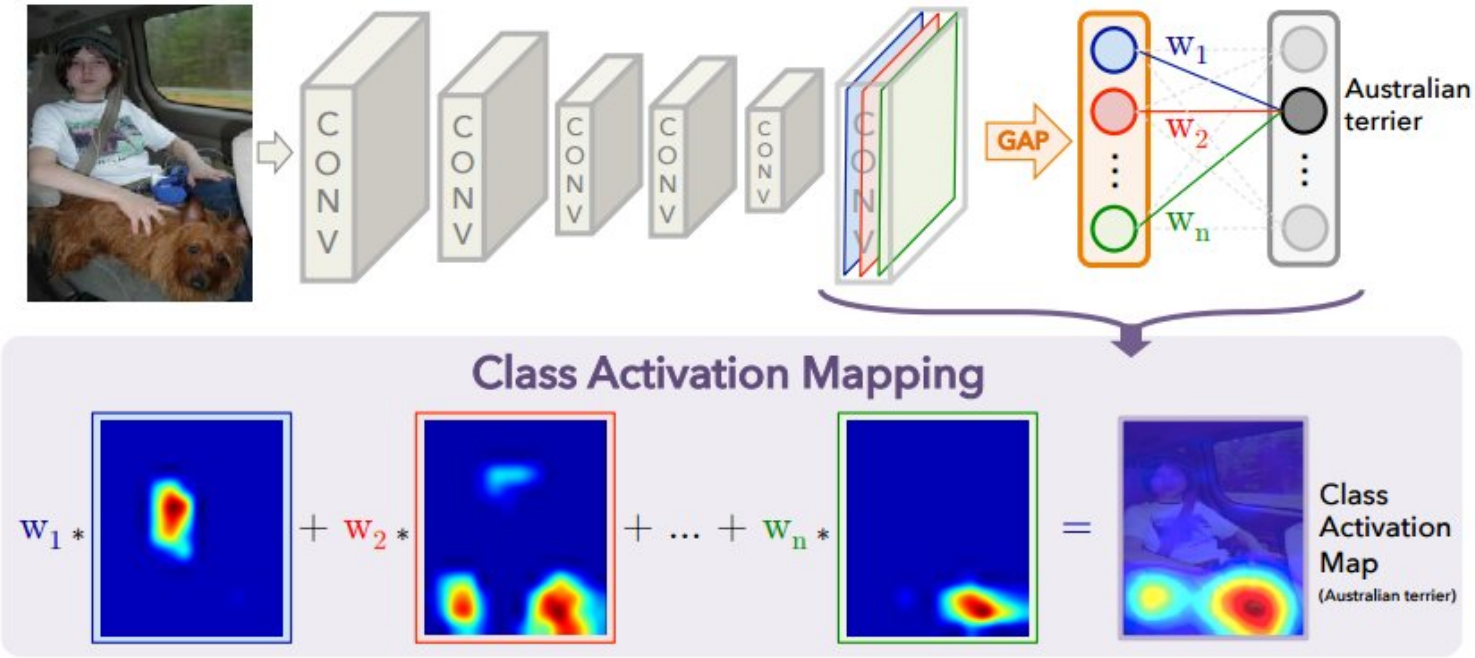
1. Move grey square over image (a)
2. See how classifier output for correct class changes (d)



Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In ECCV (pp. 818-833). Springer, Cham.

Class activation maps (CAM)

- Global average pooling: simply compute average of whole feature map
- Feature maps represent heat maps, weighting gives evidence

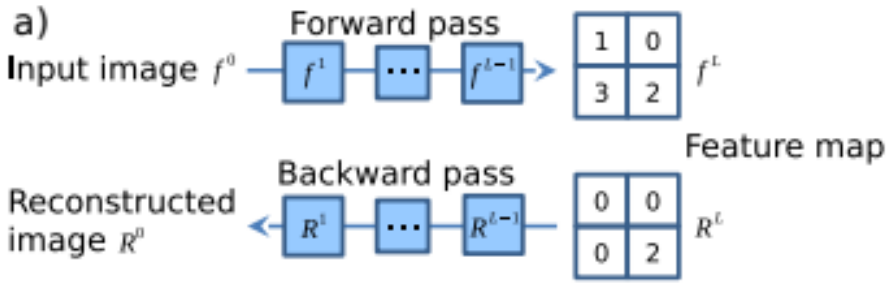


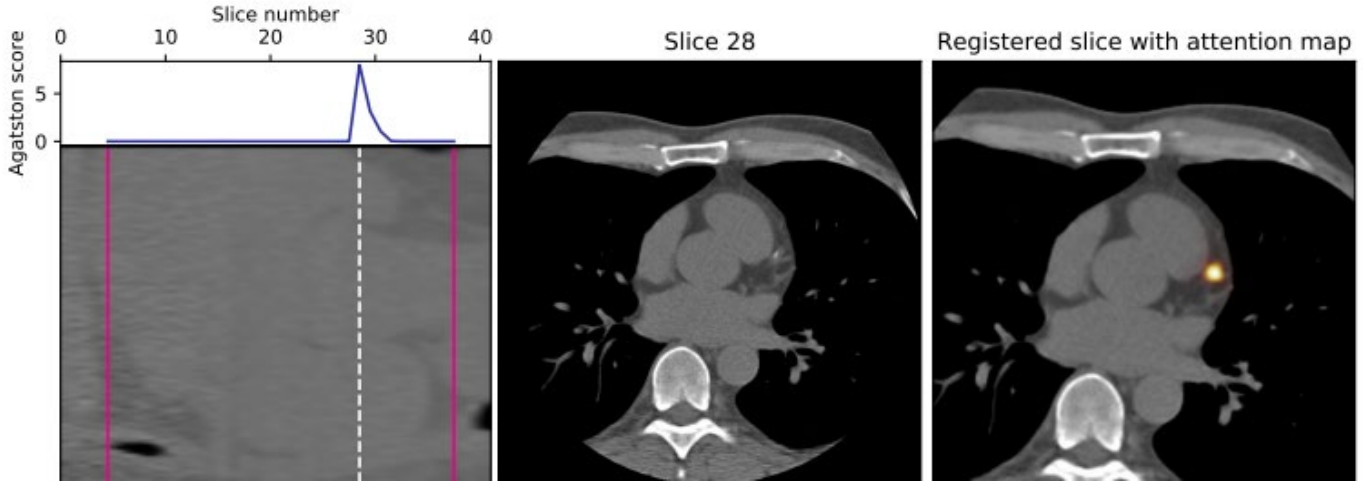
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. CVPR (pp. 2921-2929).

Deconvolution

Inversion: find input corresponding to a particular feature

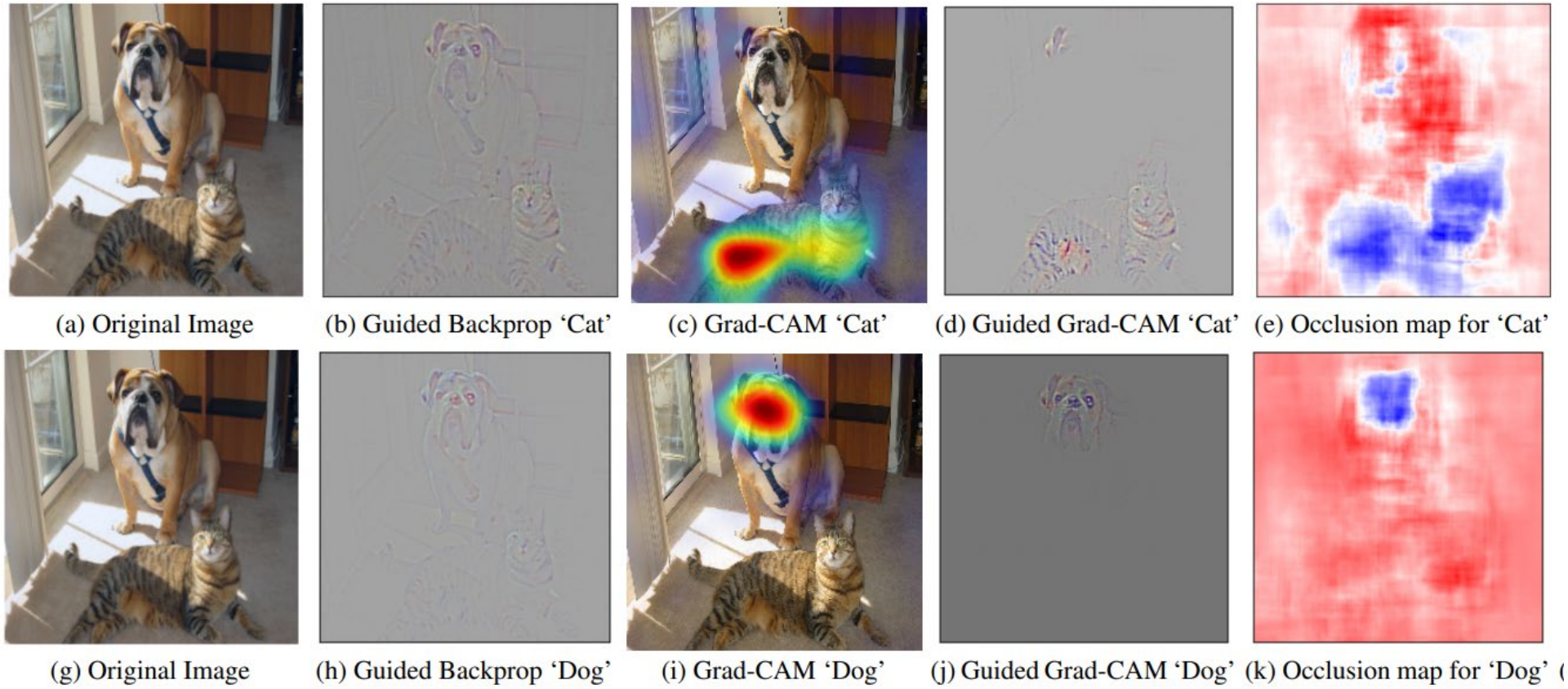
1. Push input through CNN
2. Fix value for one 'neuron', set all others to zero
3. Perform inverse of convolution, pooling, activation functions





Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In ECCV (pp. 818-833). Springer, Cham.
de Vos, B. D., Wolterink, J. M., Leiner, T., de Jong, P. A., Lessmann, N., & Išgum, I. (2019). Direct automatic coronary calcium scoring in cardiac and chest CT. IEEE transactions on medical imaging.
Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.

Comparison



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat'

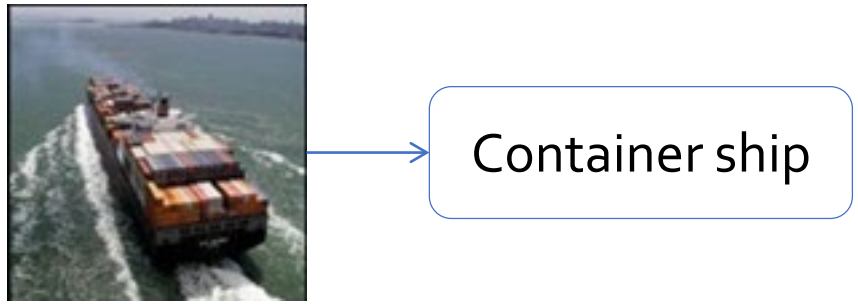
(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In ICCV (pp. 618-626).

Discriminative vs. generative models

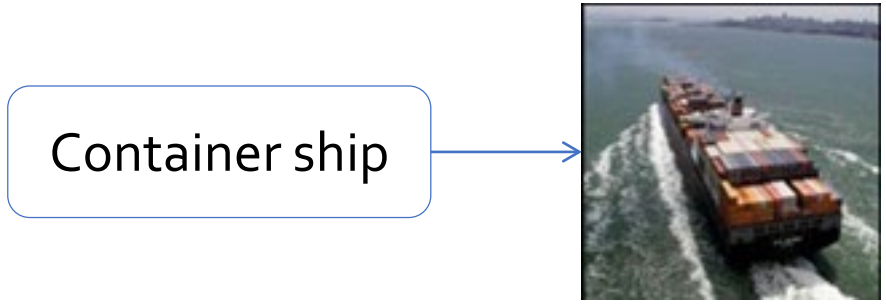
Discriminative models

- Go from high-dimensional sample to low-dimensional prediction
- E.g. from an image to the label of that image



Generative models

- Go from low-dimensional input to high-dimensional sample
- E.g. from an image label to an image sample



But why would we do this?

- **Unsupervised** learning: we don't need labels for our training data
- Learning more about your data set
- Synthesizing new data for discriminative models
- Advanced image manipulation

https://www.christies.com/lotfinder/prints-multiples/edmond-de-belamy-from-la-famille-de-6166184-details.aspx?from=salesummery&intObjectID=6166184&sid=18abf70b-239c-41f7-bf78-99c5a4370bc7

Generative adversarial networks (GANs)

Instead of **one** neural network, we train **two** neural networks

1. A **generator** network generates samples of images
2. A **discriminator** network distinguishes generated samples from real samples

The **generator** network tries to fool the discriminator

Goodfellow et al. *Generative adversarial nets*, NIPS 2014

Generative adversarial networks (GANs)

1. A **generator** network generates samples of images
2. A **discriminator** network distinguishes generated samples from real samples

Generator
Counterfeiter

Discriminator
Detective

Generative adversarial networks (GANs)

1. A **generator** network generates samples of images
2. A **discriminator** network distinguishes generated samples from real samples

Generator
Counterfeiter

Discriminator
Detective

Generative adversarial networks (GANs)

1. A **generator** network generates samples of images
2. A **discriminator** network distinguishes generated samples from real samples

Generator
Counterfeiter

Discriminator
Detective

Generative adversarial networks (GANs)

1. A **generator** network generates samples of images
2. A **discriminator** network distinguishes generated samples from real samples

Generator
Counterfeiter

Discriminator
Detective

Generative adversarial networks (GANs)

A bit more formally…

We are given samples $x_1, \ldots, x_n$ from a data distribution p_{data}

We want to generate new samples similar to p_{data}

We can sample from p_{data} efficiently – examples are available

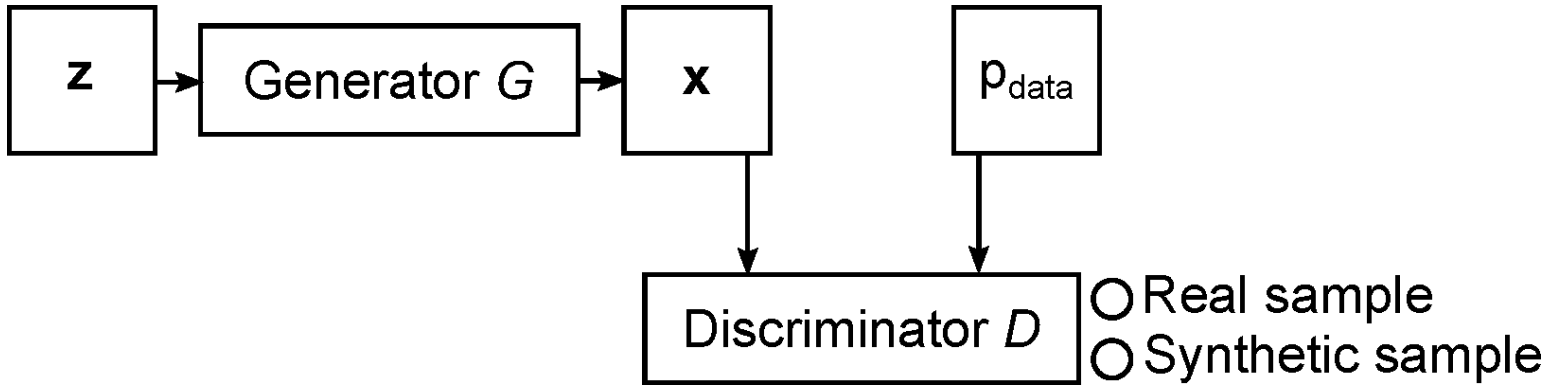
Generative adversarial networks (GANs)

Aim: Find a sample distribution p_{gen} that closely resembles p_{data}

Generator *G* maps points from a normal/uniform distribution *z* to samples *x* in p_{gen}

Discriminator *D* tells samples from p_{data} and p_{gen} apart

Generators *G* and *D* are networks with trainable parameters ϑ_G and ϑ_D



Generative adversarial networks (GANs)

Discriminator *D* maximizes an objective function by assigning a high probability to real samples and a low probability to synthetic samples

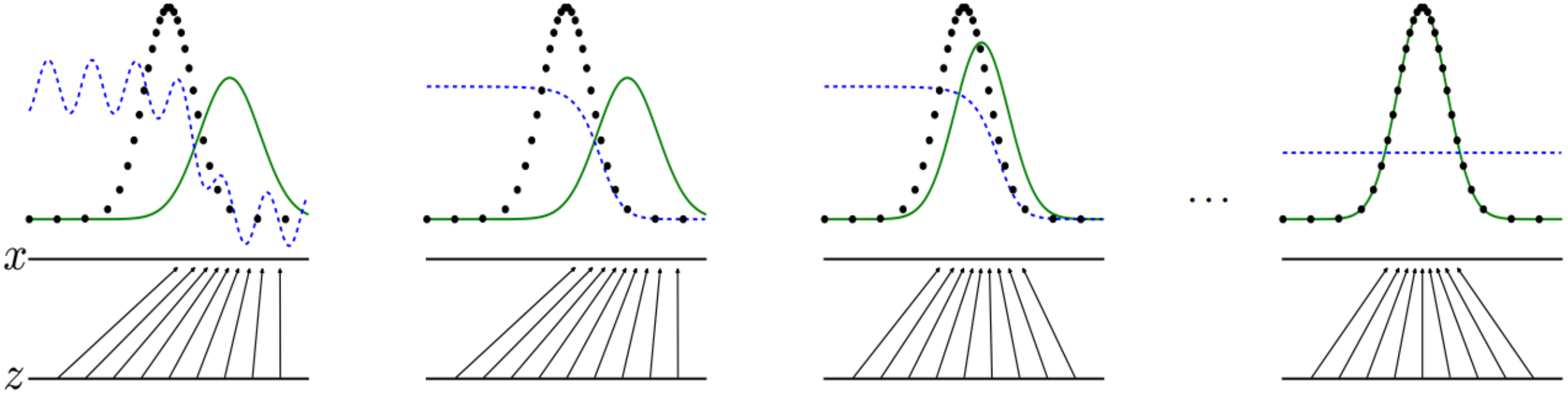
$$V^{(D)}(D,G) = \underset{x \sim p_{data}}{\mathbb{E}} [\log D(x)] + \underset{z \sim p_z}{\mathbb{E}} [\log (1 - D(G(z)))]$$

At the same time, *G* tries to minimize this objective function by generating samples that have a high probability of being real

$$V^{(G)}(D,G) = \underset{z \sim p_z}{\mathbb{E}} [\log (1 - D(G(z)))]$$

Goodfellow et al. *Generative adversarial nets*, NIPS 2014

GANs in theory



Start

Convergence

p_{real} p_{gen} $D(x)$

Demo: https://poloclub.github.io/ganlab/

Goodfellow et al. *Generative adversarial nets*, NIPS 2014

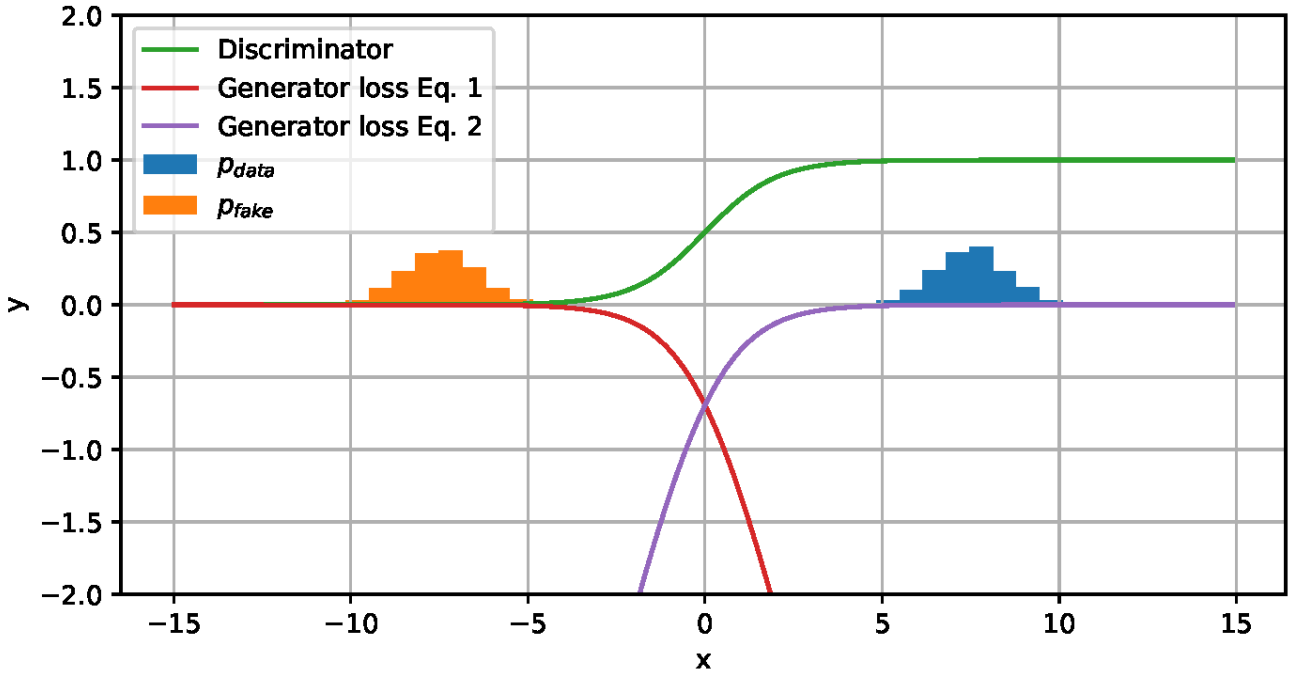
Challenges when training GANs

Poor gradients from discriminator *D* to generator *G*

Especially when the discriminator *D* can easily distinguish fake and real samples

1) $V^{(G)}(D,G) = \underset{z\sim p_z}{\mathbb{E}}\left[\log\left(1 - D(G(z))\right)\right]$

2) $V^{(G)}(D,G) = -\underset{z\sim p_z}{\mathbb{E}}\left[\log\left(D(G(z))\right)\right]$



Challenges when training GANs

1.Poor gradients from discriminator *D* to generator *G*

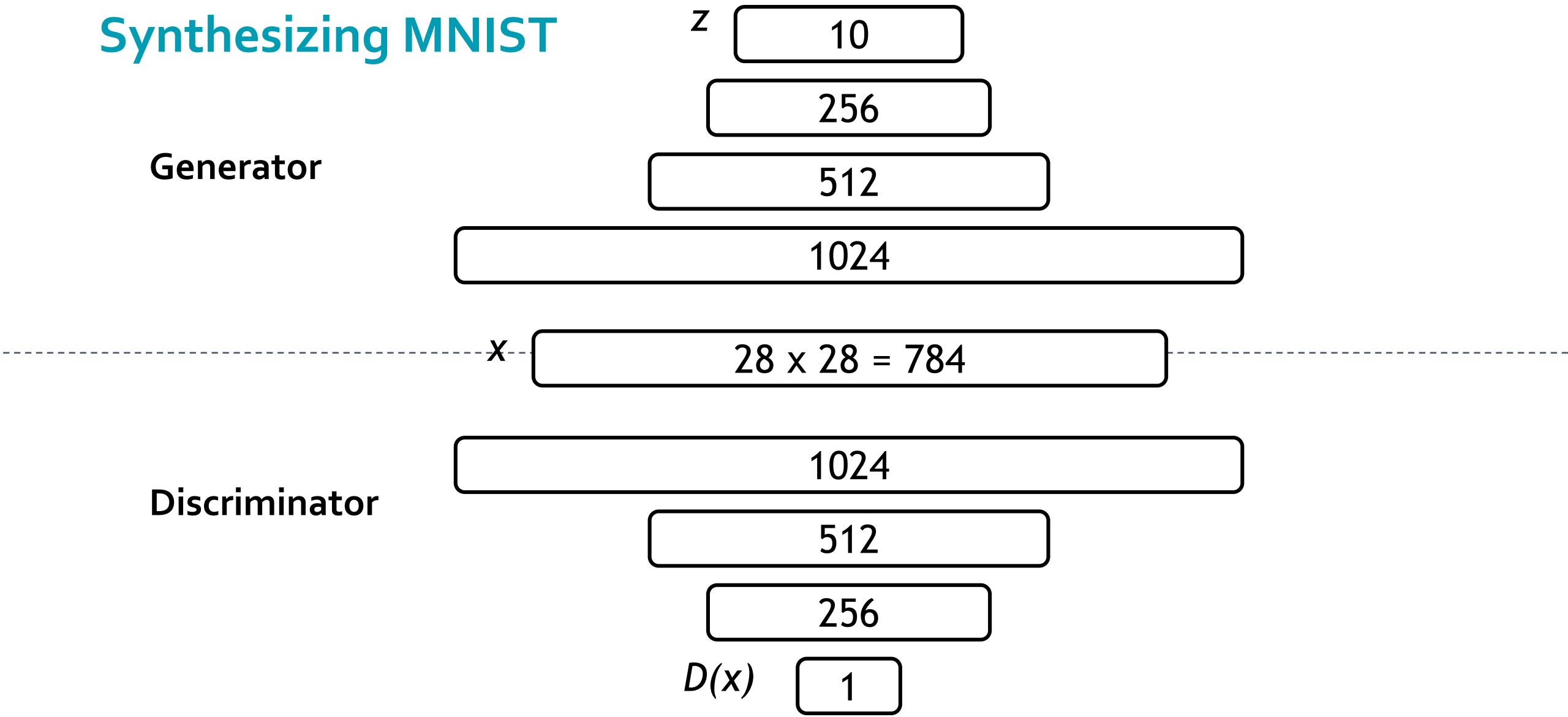
- Especially when the discriminator *D* can easily distinguish fake and real samples

2.Mode collapse

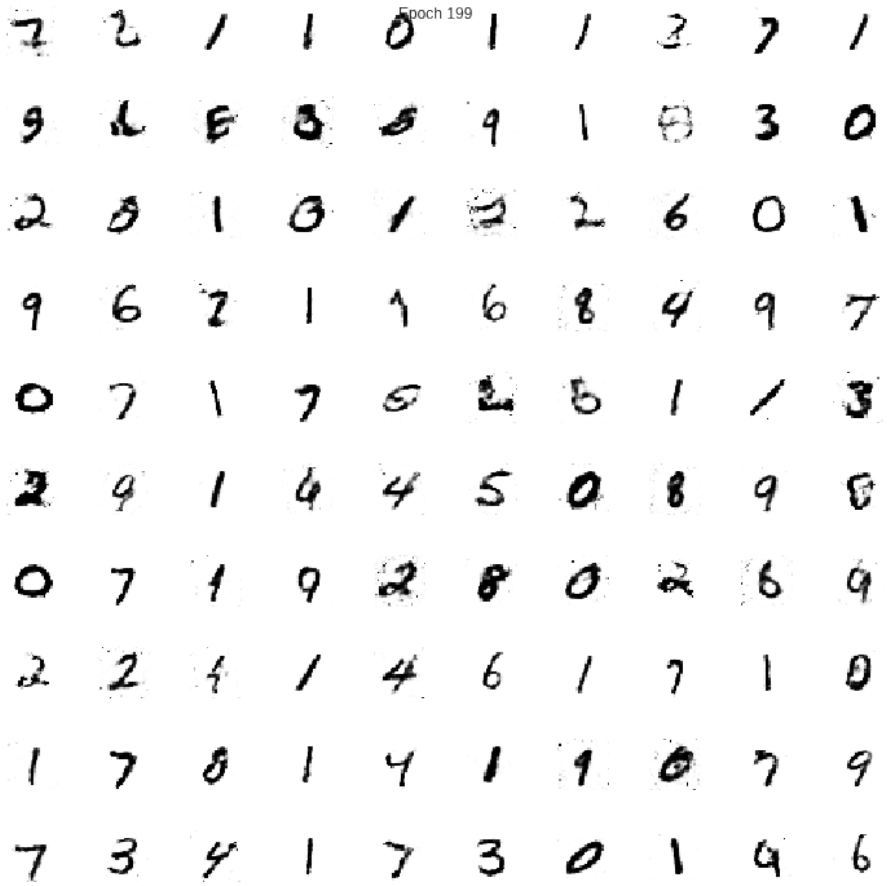
- Generator *G* maps all noise vectors to similar samples *G(z)*
- E.g. generator only synthesizes 1 in MNIST
- Generator and discriminator chase each other from mode to mode

3.Unclear when training is finished

Synthesizing MNIST

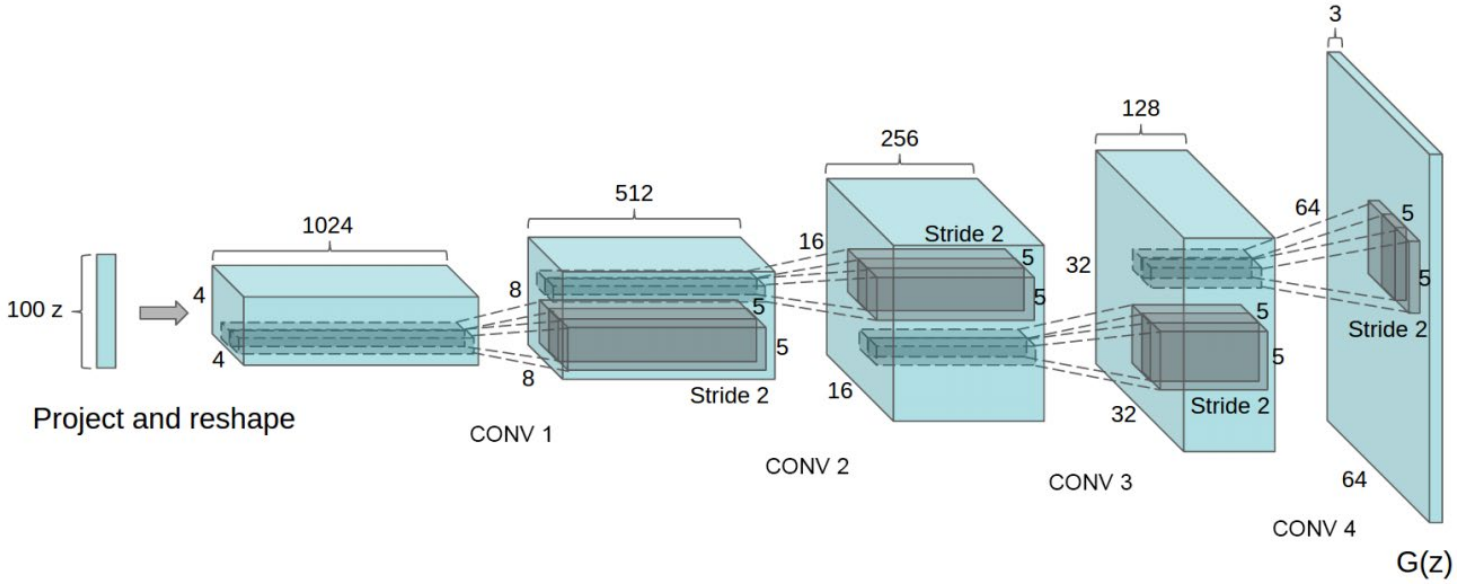


Synthesizing MNIST



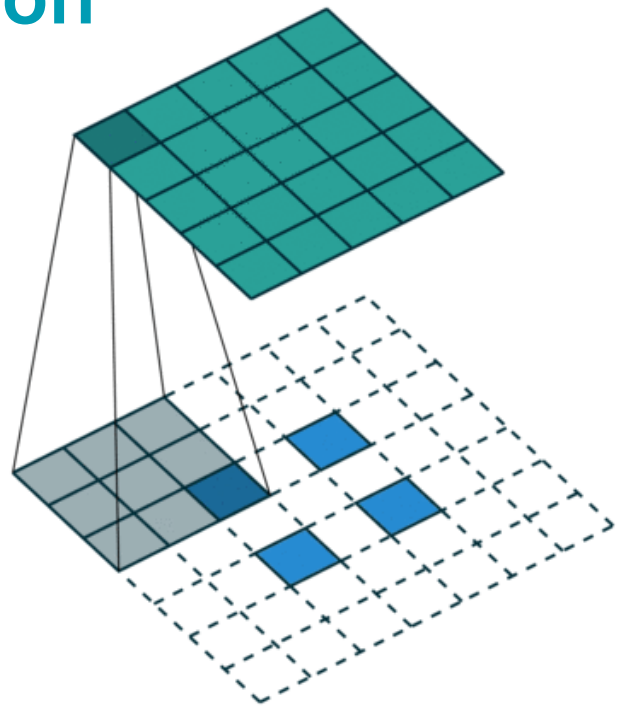
Synthesizing images

- Image synthesis requires convolutional generators and discriminators
- We know how to reduce the image size
- Now we need some way to increase the image size



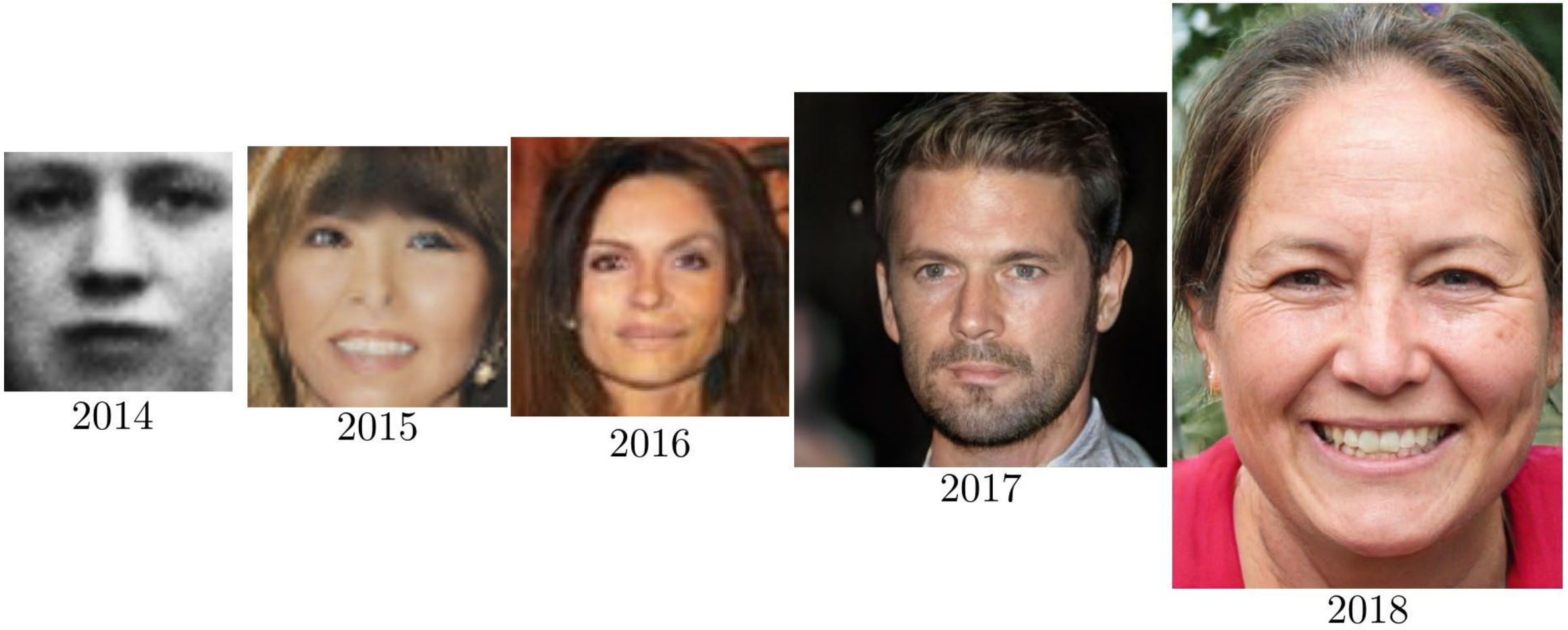
Radford, A., Metz, L., & Chintala, S. (2015). arXiv preprint arXiv:1511.06434.

Transposed convolution



- Deconvolution
- Fractionally-strided convolution
- Transposed convolution

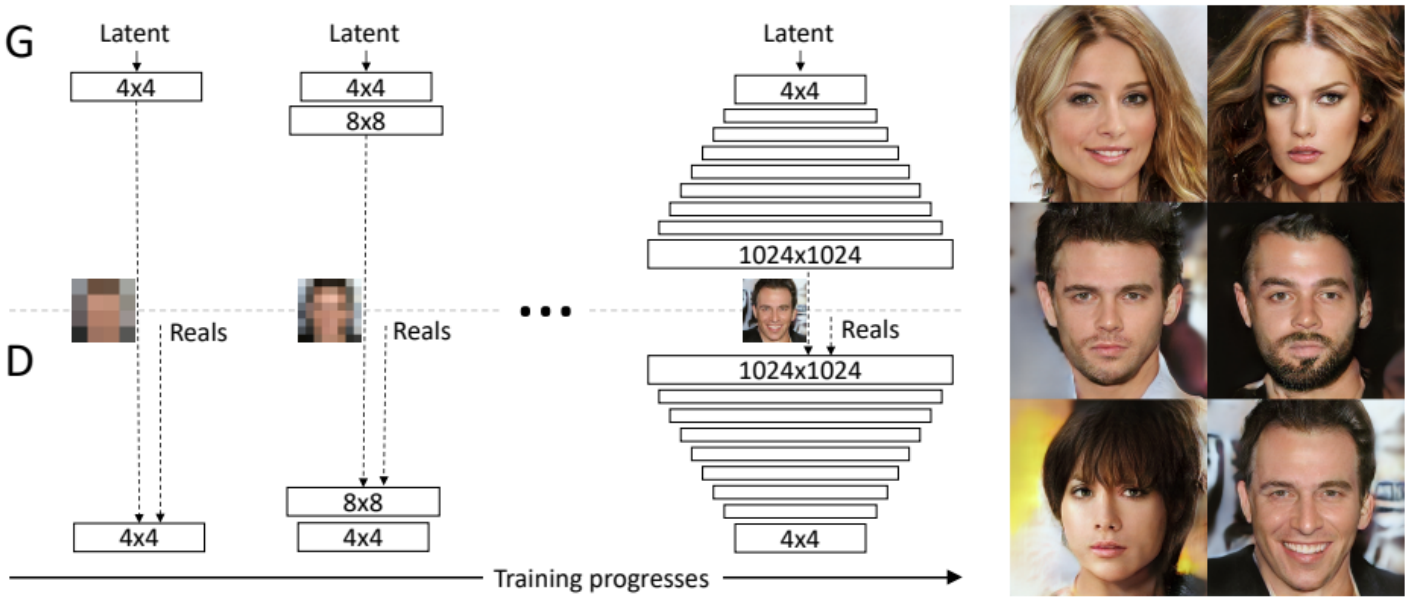
http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html



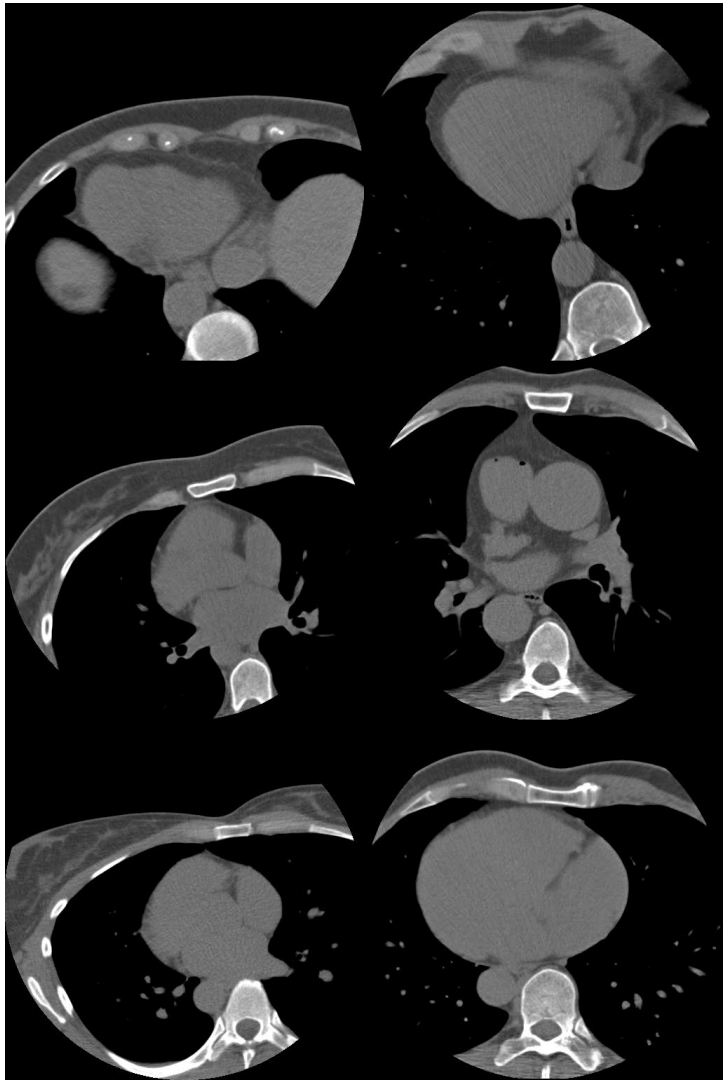
https://twitter.com/goodfellow_ian/status/1084973596236144640?s=19
https://arxiv.org/abs/1406.2661, https://arxiv.org/abs/1511.06434, https://arxiv.org/abs/1606.07536, https://arxiv.org/abs/1710.10196, https://arxiv.org/abs/1812.04948

Progressive GAN

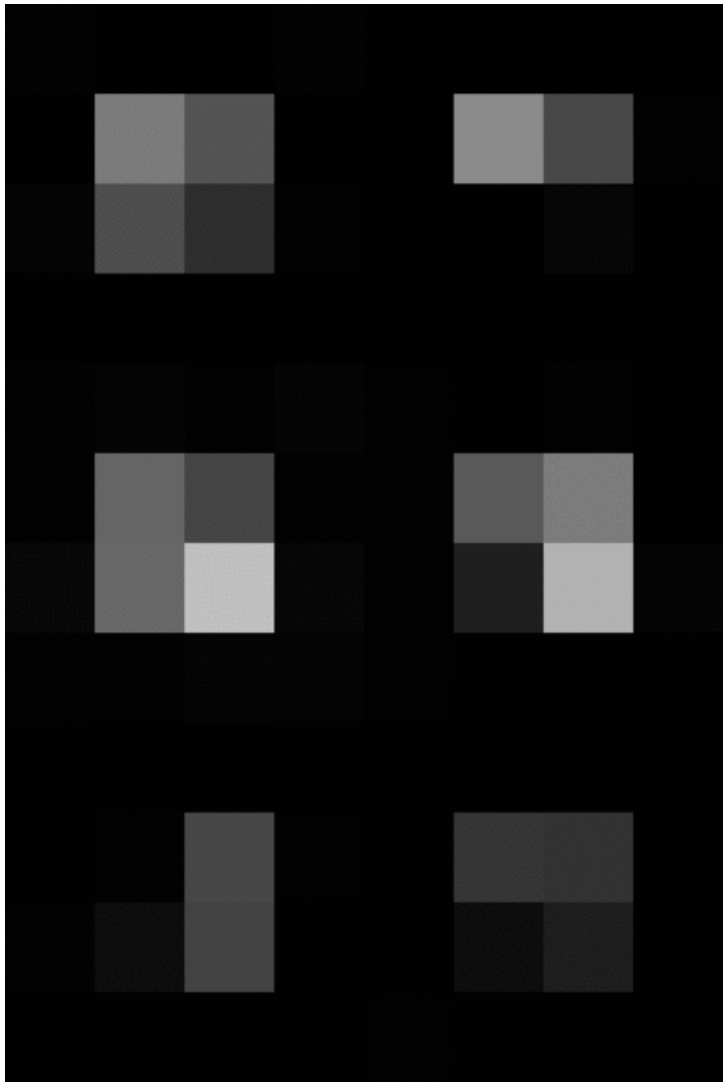
- Goal: high-resolution image synthesis (1024 x 1024 pixels)
- Problem: optimizing large and deep networks can lead to stability issues
- Trick: periodically increase resolution in *G* and *D*



Real



Fake



Karras et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, ICLR 2018

StyleGAN



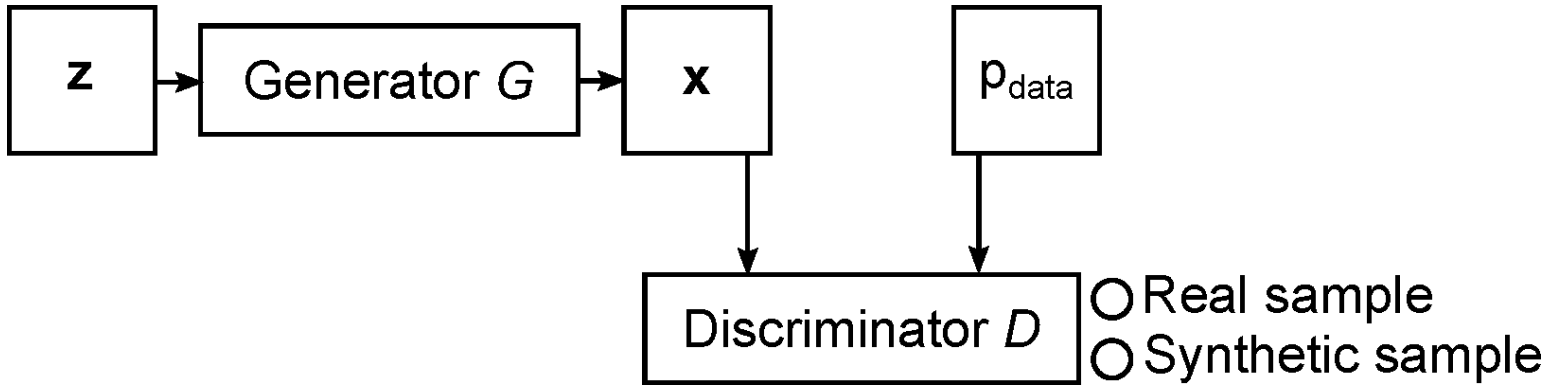


Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948.
https://thispersondoesnotexist.com/

Spaces

- GANs can help us learn more about our data
- Different latent space points correspond to different sample space points
- The latent space is structured
- Interpolation in the latent space leads to smooth transitions

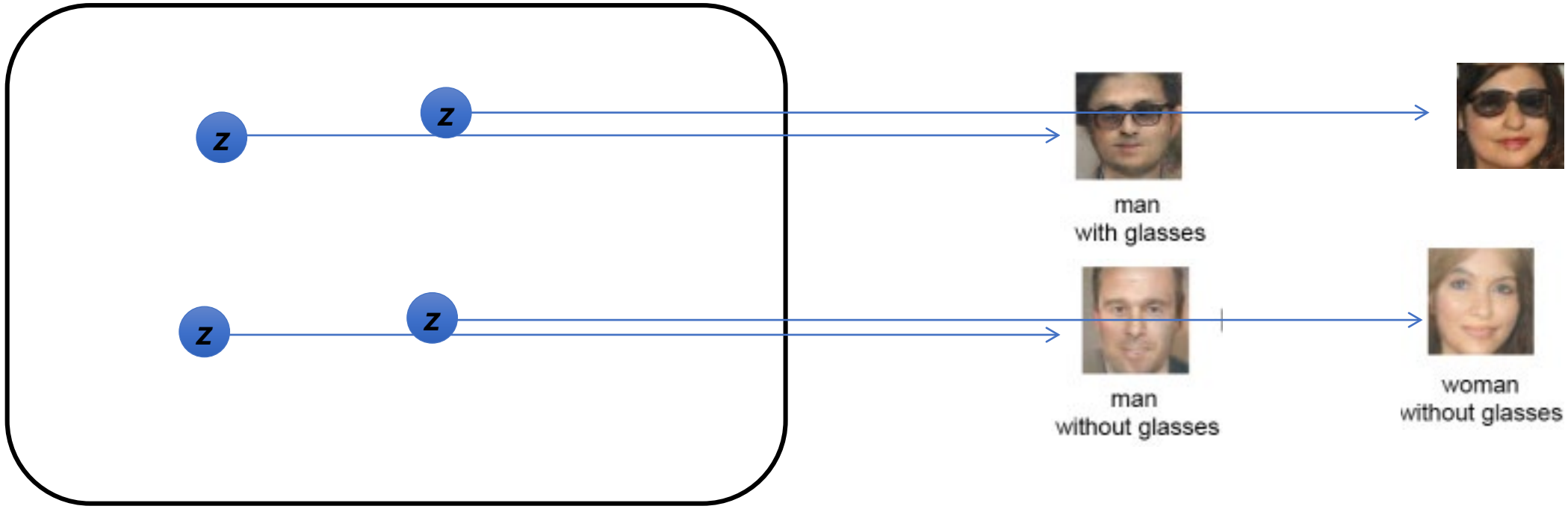
The latent space **The sample space**



Spaces

The latent space

The sample space

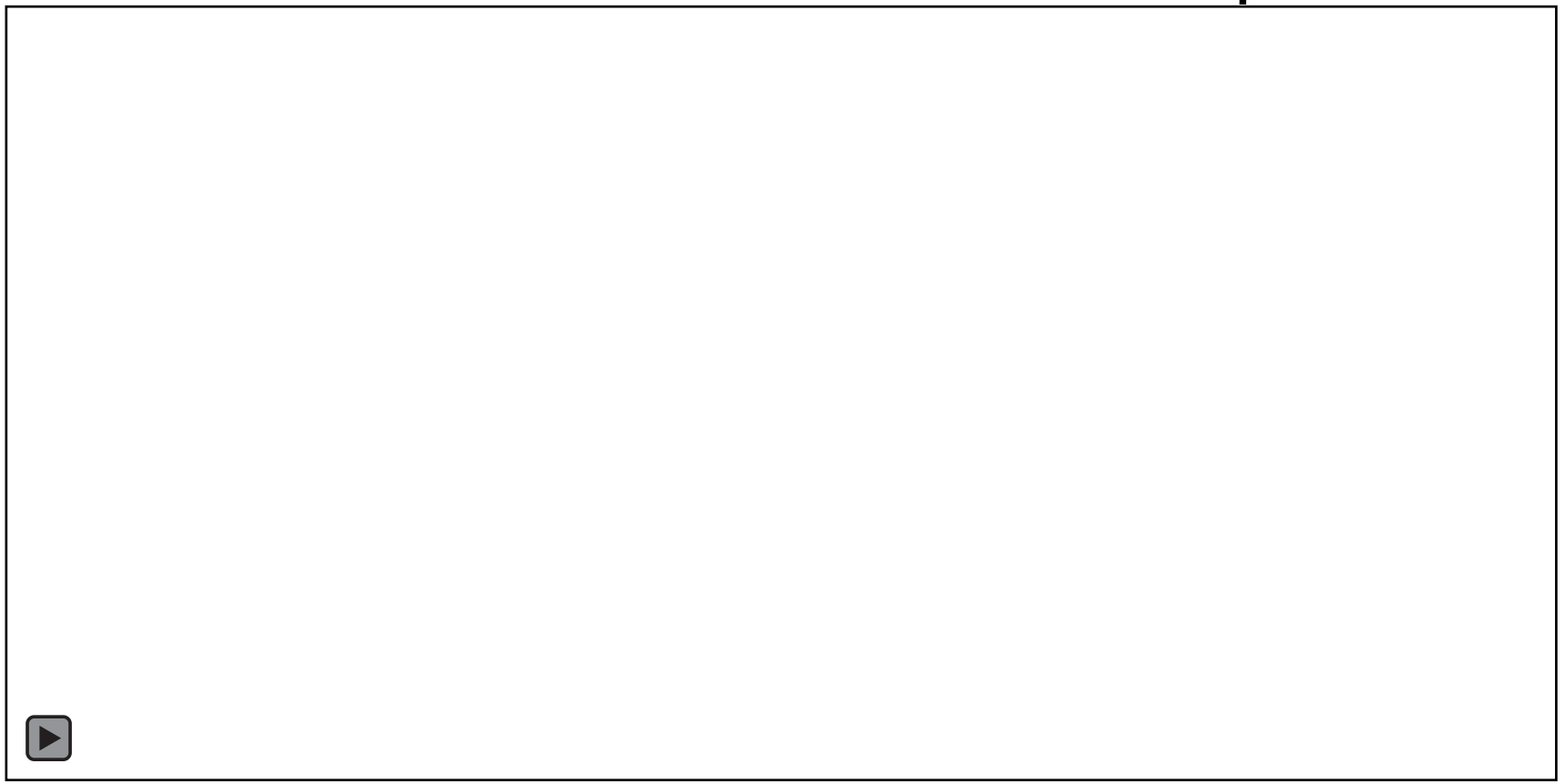


Radford, A., Metz, L., & Chintala, S. (2015). arXiv preprint arXiv:1511.06434.

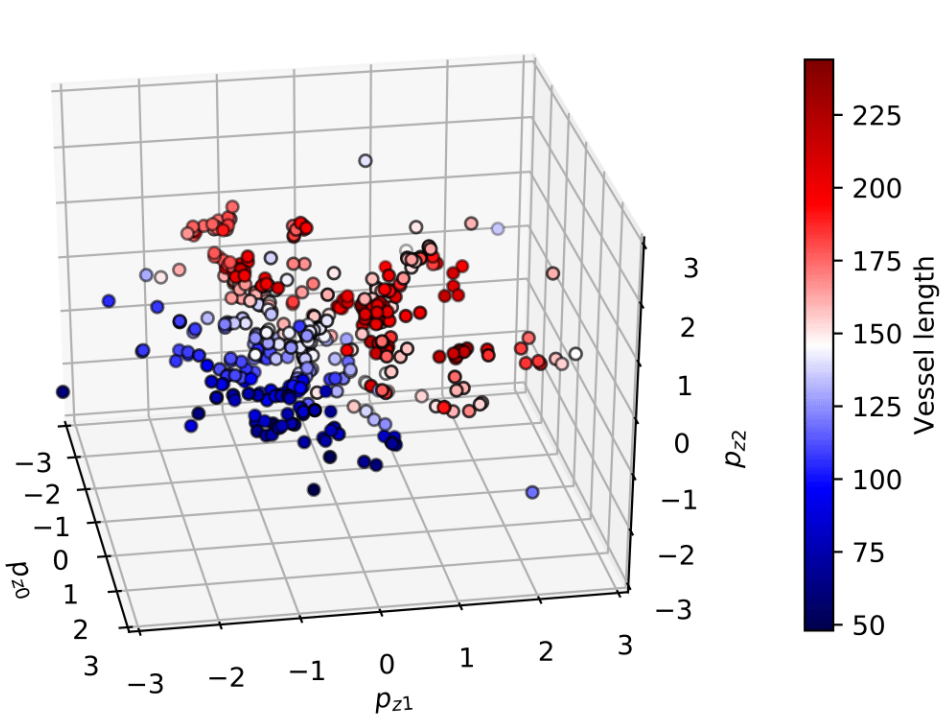
Example: coronary artery synthesis

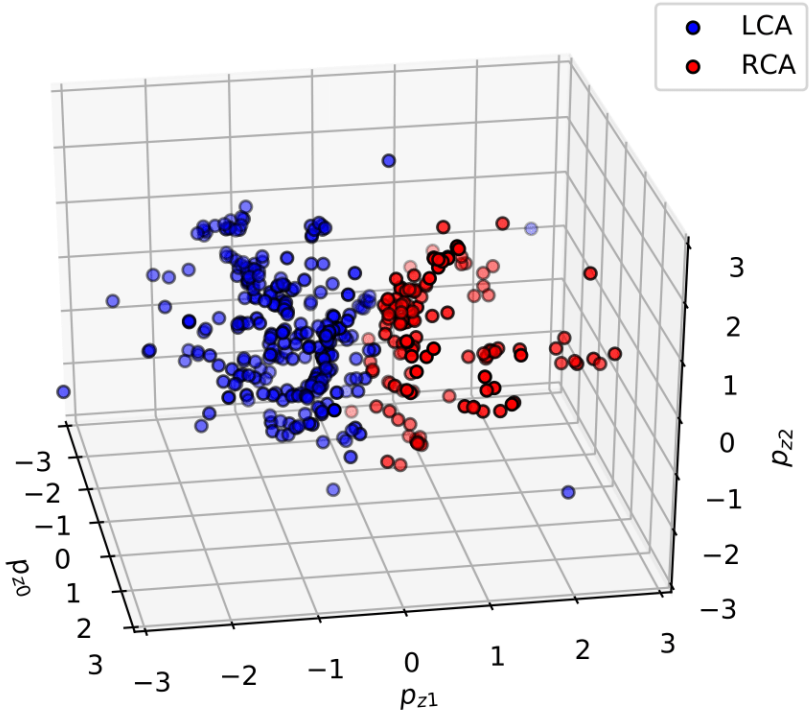
The sample space

The latent space



Example: coronary artery synthesis

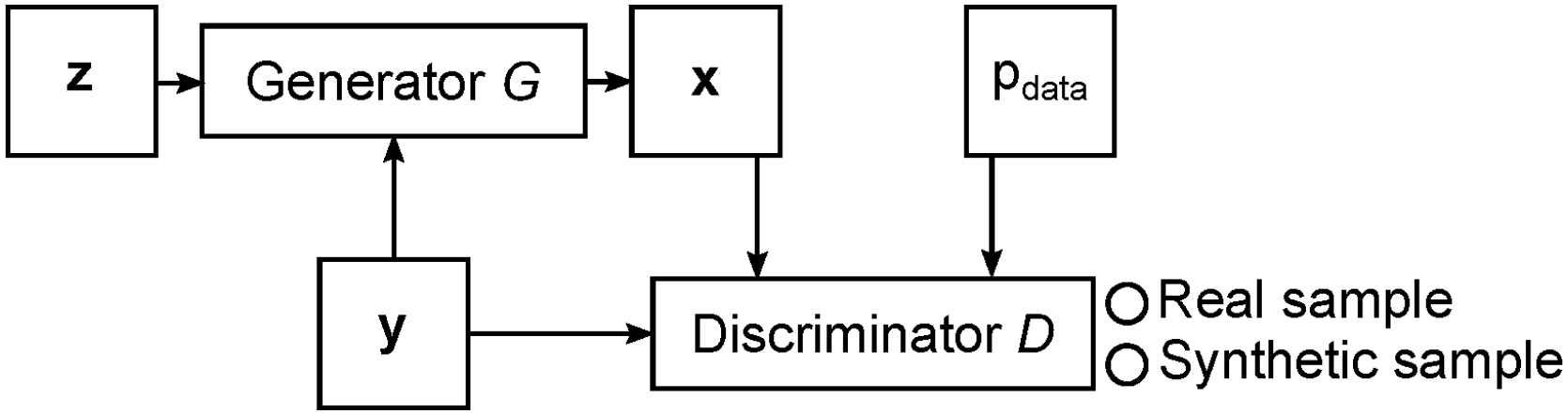




Wolterink, J. M., Leiner, T., & Isgum, I. (2018). Blood Vessel Geometry Synthesis using Generative Adversarial Networks. arXiv preprint arXiv:1804.04381.

Conditional GAN (cGAN)

- GAN: difficult to control what is generated
- Condition GAN on extra input *y*

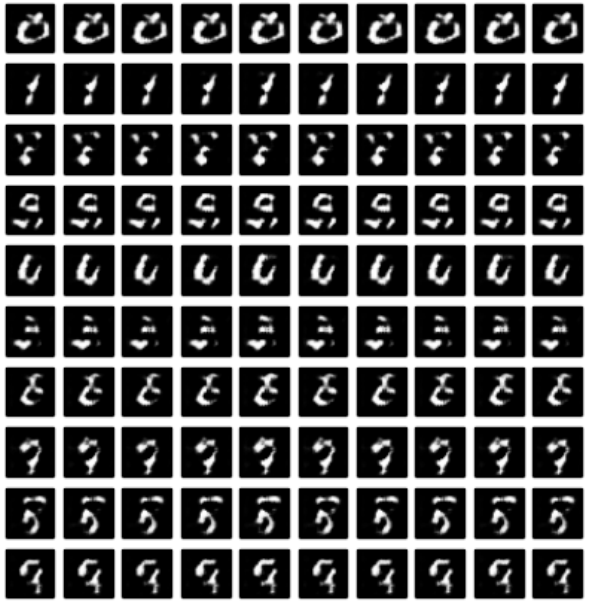


Mirza et al. *Conditional generative adversarial nets*, 2014

Conditional GAN (cGAN)

The generator and discriminator optimize an objective function based on the predictions for real samples and synthetic samples, given extra information

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



Epoch 1

Mirza et al. *Conditional generative adversarial nets*, 2014

BigGAN

- State of the art in conditional GAN training
- Trained on ImageNet (~million images)
- Get any class you want

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.

In medical imaging

We typically have some image to condition on, 'image-to-image' applications

- Segmentation
- Synthesis
- Image quality enhancement

It's not always easy to determine a loss function that captures what we want

- Noise-free segmentations?
- Sharp images?

Use a conditional GAN where *y* is the input image

Image-to-image translation

Domain A: input image *y*

Domain B: output image *x*

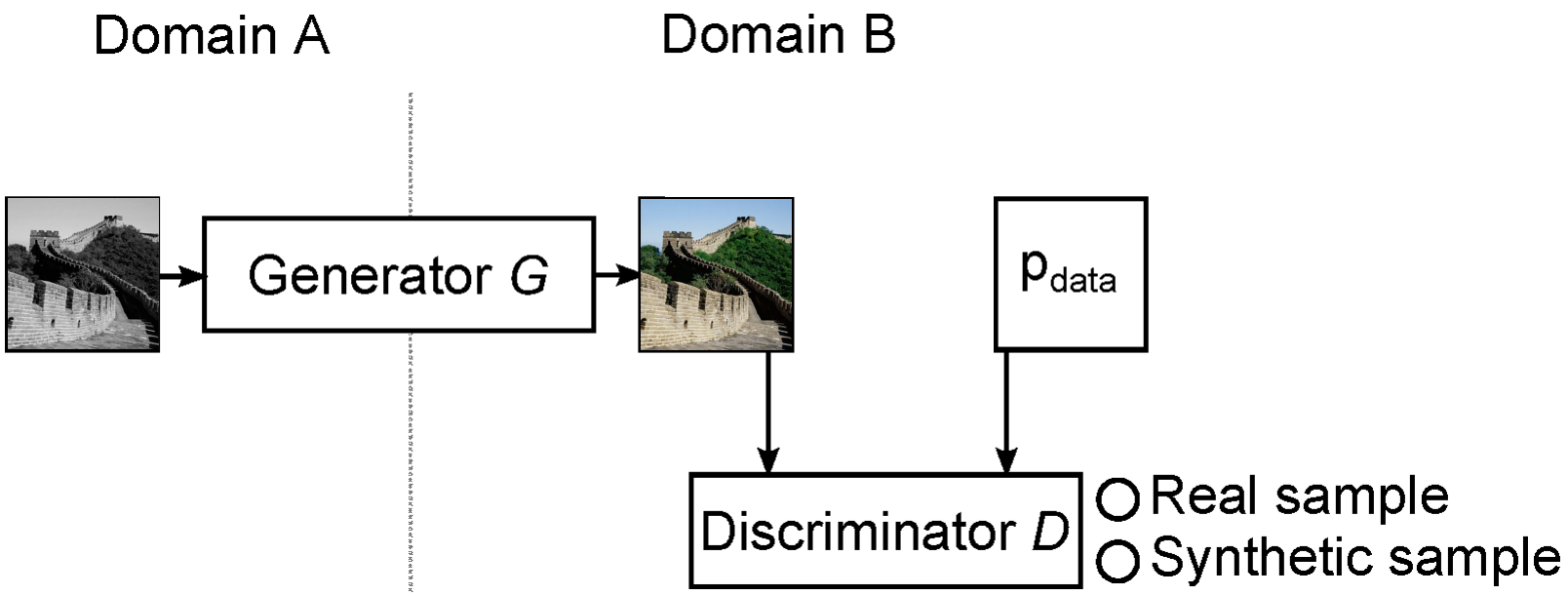
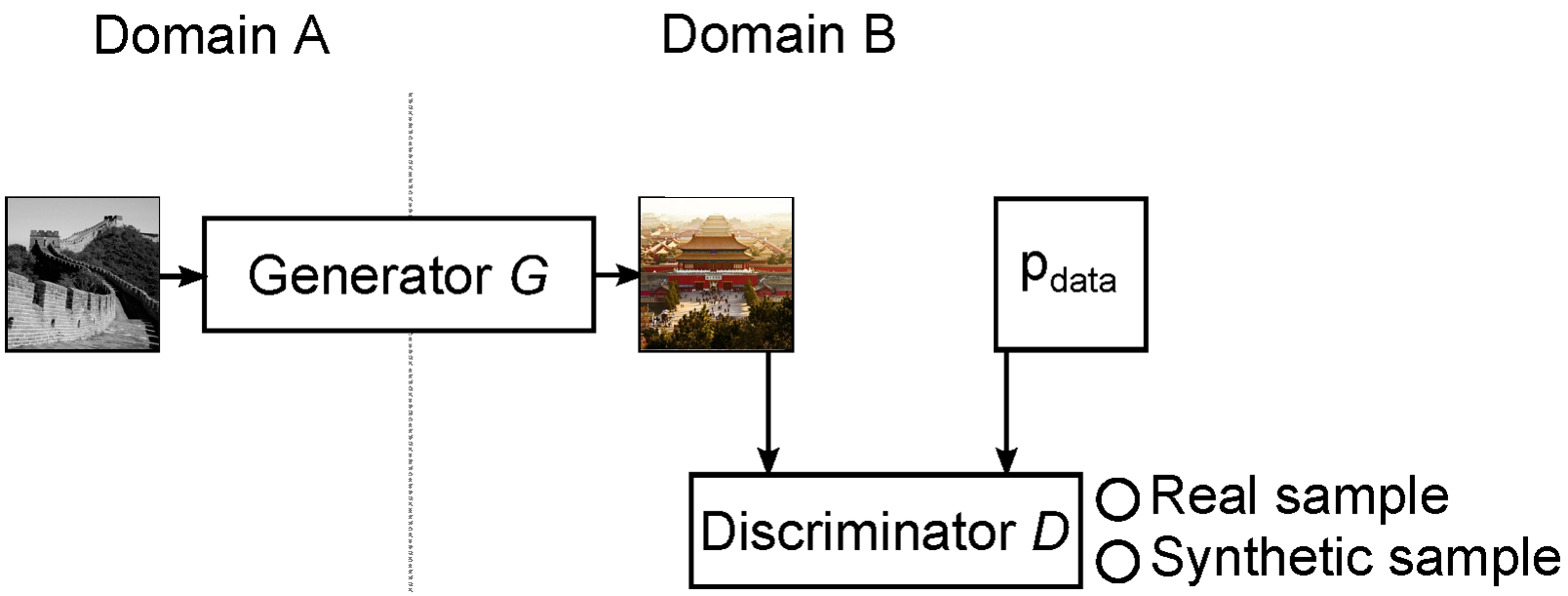


Image-to-image translation

The output image *x* could be very realistic but 'wrong'

Regularize to make sure that it corresponds to the input image *y*



Paired vs. unpaired training data

Paired

Domain A

Domain B

Unpaired

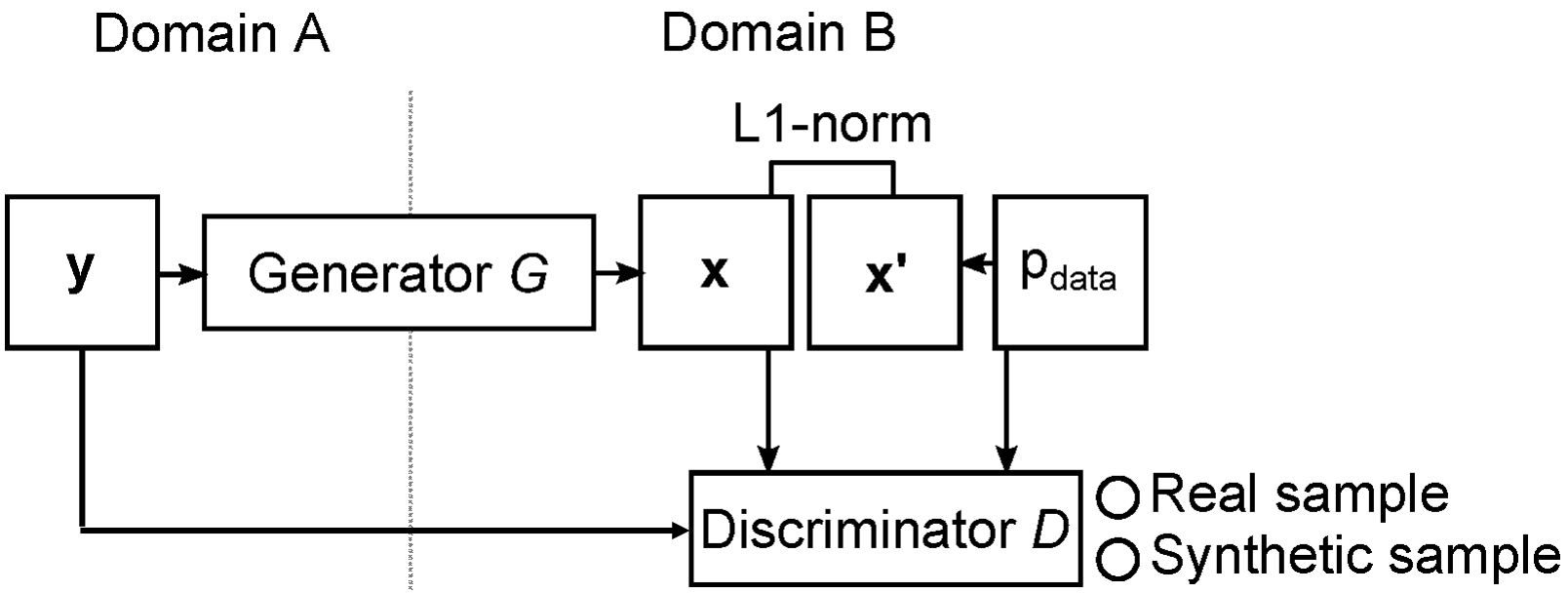
Domain A

Domain B

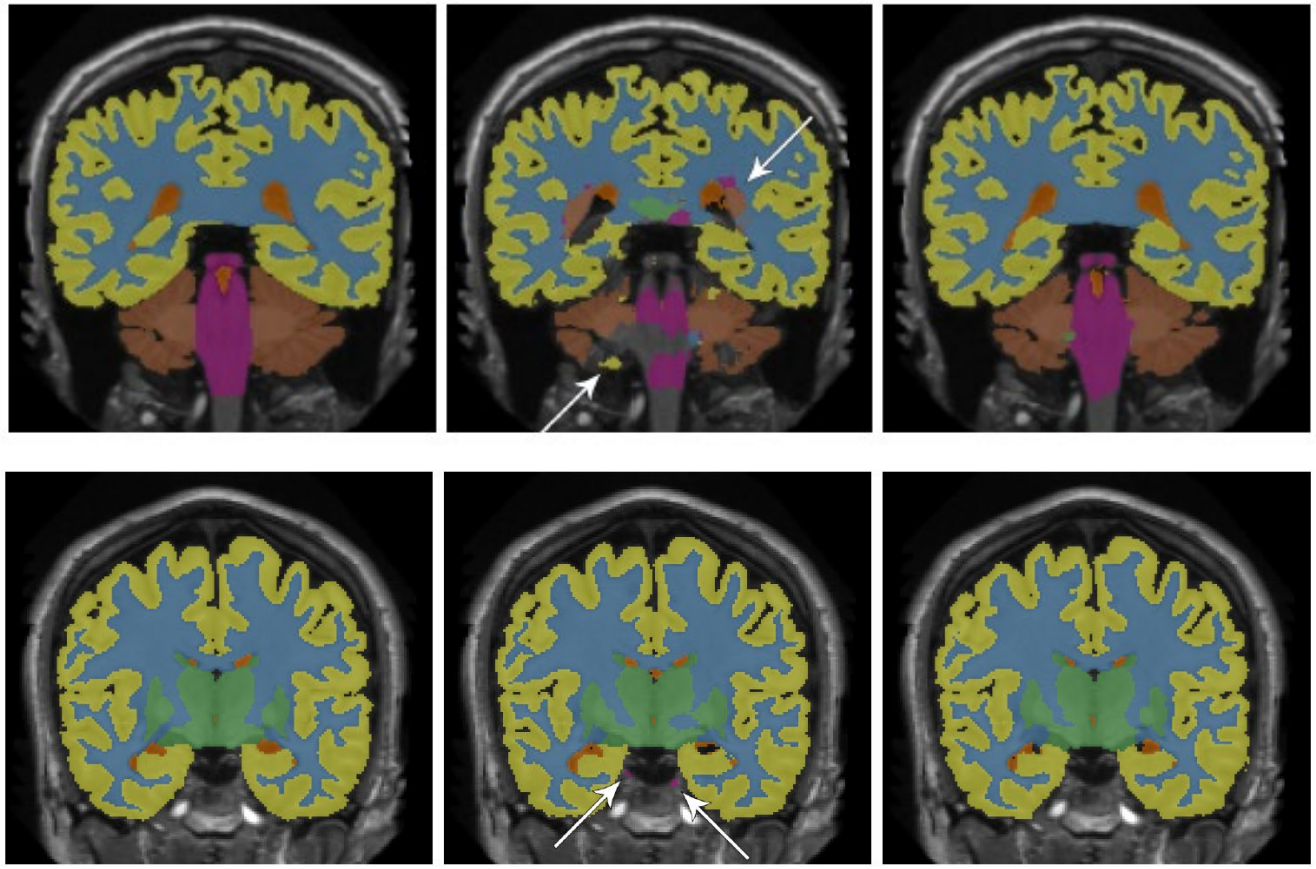
Paired image-to-image translation

The discriminator and generator optimize an objective function based on predictions for real and synthetic images in domain B, given input images in domain A while minimizing the difference between real and synthetic images in domain B

$$\min_G \max_D V(D,G) = \mathop{\mathbb{E}}_{x,y\sim p_{data_{B,A}}}[\log D(x,y)] + \mathop{\mathbb{E}}_{y\sim p_{data_A}}[\log(1 - D(G(y),y)))] + \lambda \mathop{\mathbb{E}}_{x,y\sim p_{data_{B,A}}}[||x - G(y)||]_1$$



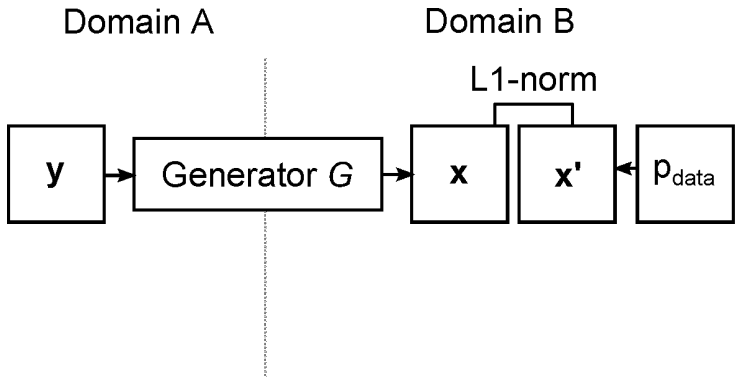
Brain MR segmentation



Ground truth

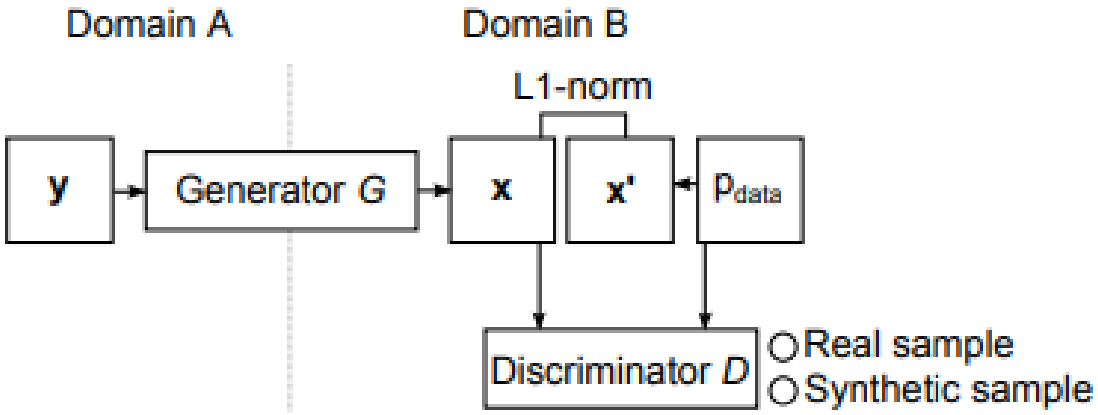
No adversarial training

Adversarial training



Moeskops et al. *Adversarial training and dilated convolutions for brain MRI segmentation*, MICCAI DLMIA Workshop, 2017

Unpaired training: self-regularization



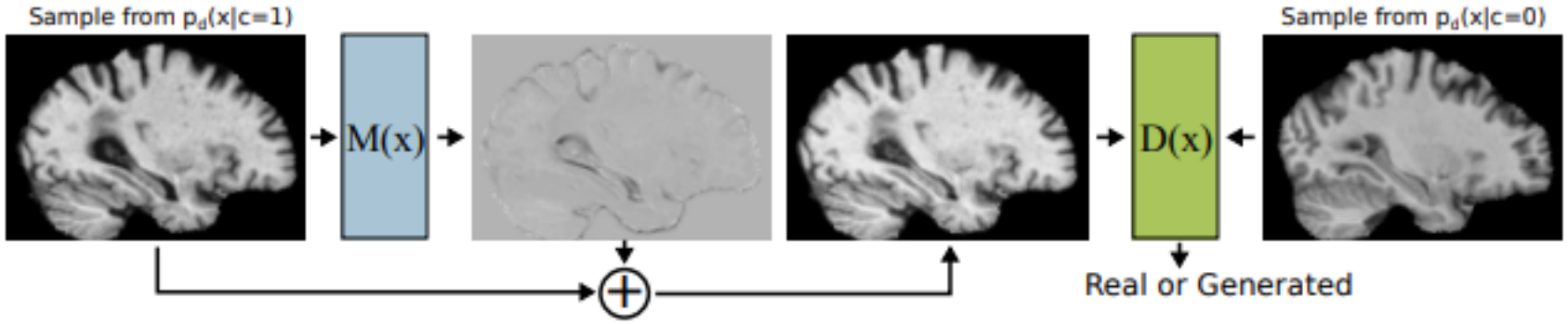
Unpaired training: self-regularization

Domain A

Brain MR images of patients with Alzheimer's

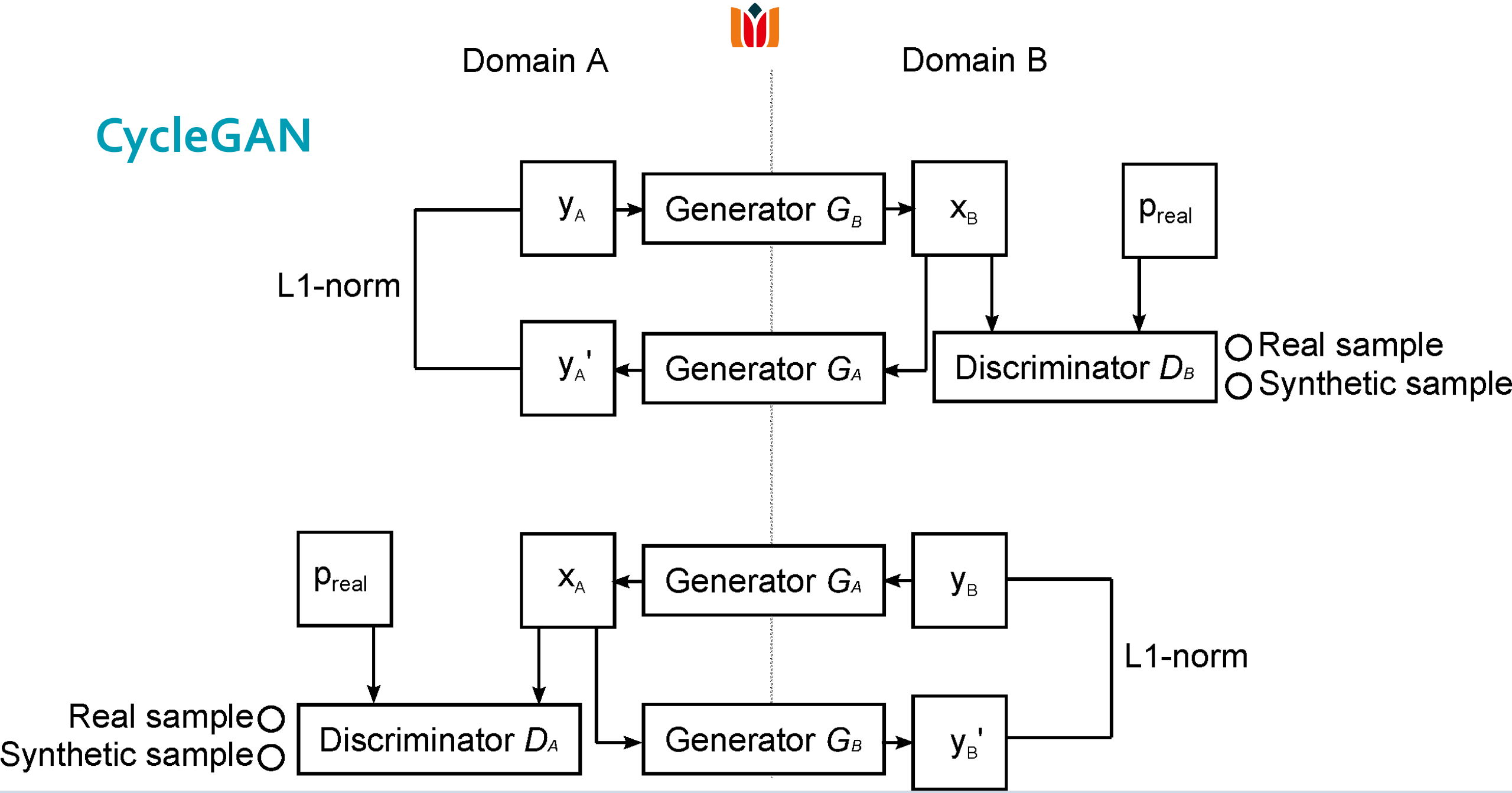
Domain B

Brain MR images of healthy patients



Baumgartner, C. F., Koch, L. M., Can Tezcan, K., Xi Ang, J., & Konukoglu, E. (2018). Visual feature attribution using Wasserstein GANs. In CVPR (pp. 8309-8319).

CycleGAN



Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, ICCV 2017

CycleGAN

GAN objective functions in domain B and domain A and a cycle consistency loss

$$V_B(D_B, G_B) = \mathop{\mathbb{E}}_{x \sim p_{data_B}} [\log D_B(x)] + \mathop{\mathbb{E}}_{y \sim p_{data_A}} [\log (1 - D_B(G_B(y)))]$$

$$V_A(D_A, G_A) = \mathop{\mathbb{E}}_{y \sim p_{data_A}} [\log D_A(y)] + \mathop{\mathbb{E}}_{x \sim p_{data_B}} [\log (1 - D_A(G_A(x)))]$$

$$V_{Cycle}(G_A, G_B) = \mathop{\mathbb{E}}_{y \sim p_{data_A}} [||G_A(G_B(y)) - y||_1] + \mathop{\mathbb{E}}_{x \sim p_{data_B}} [||G_B(G_A(x)) - x||_1]$$

$$\min_{G_A, G_B} \max_{D_A, D_B} V(G_A, G_B, D_A, D_B) = V_B(D_B, G_B) + V_A(D_A, G_A) + \lambda V_{Cycle}(G_A, G_B)$$

Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, ICCV 2017

CycleGAN

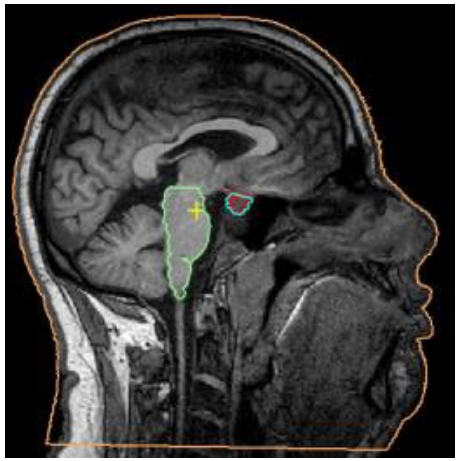
Domain A | **Domain B**

https://github.com/junyanz/CycleGAN

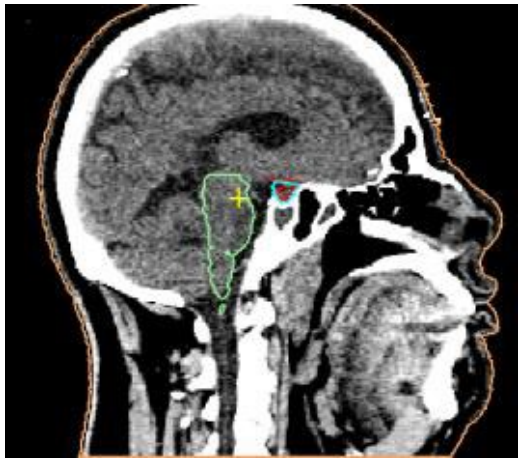
Example: MR to CT synthesis

- Radiotherapy treatment planning requires
  - MR volume
    - Soft tissue contrast
    - Tissue delineation
  - CT volume
    - Electron density
    - Dose calculation

- Acquisition of both volumes leads to
  - Increase in time and money
  - Decrease in patient comfort

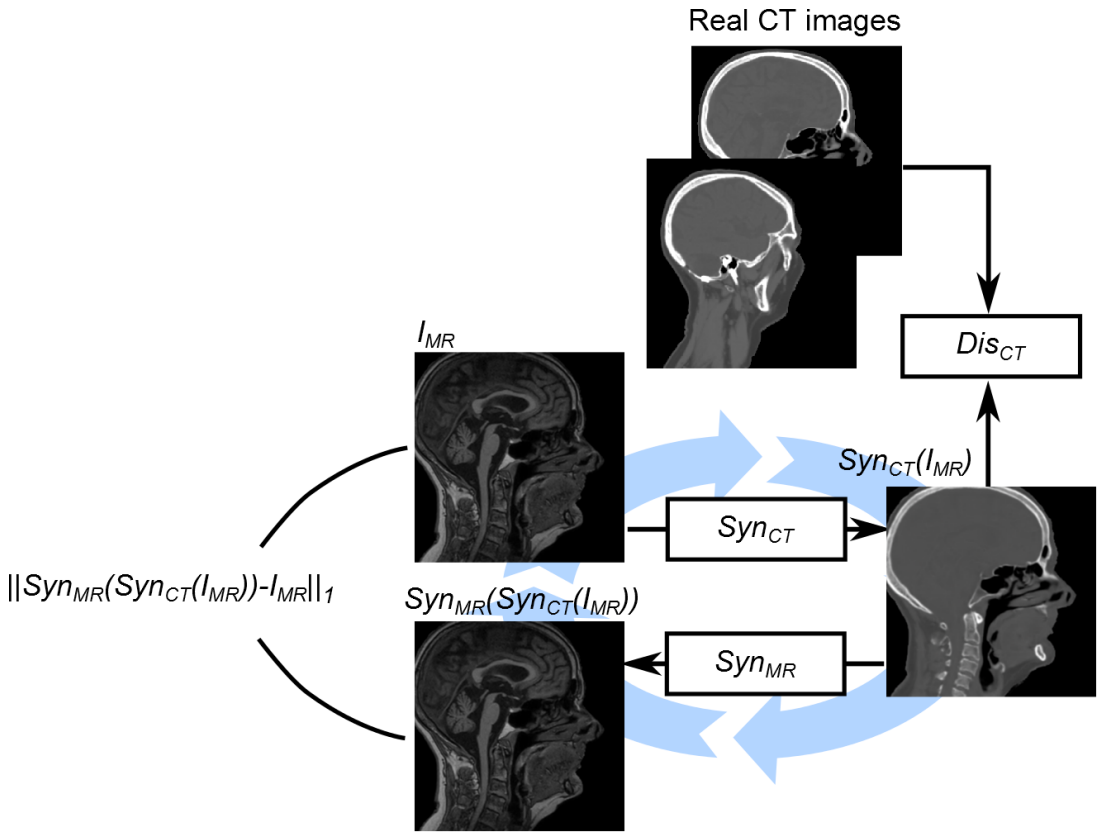


MR

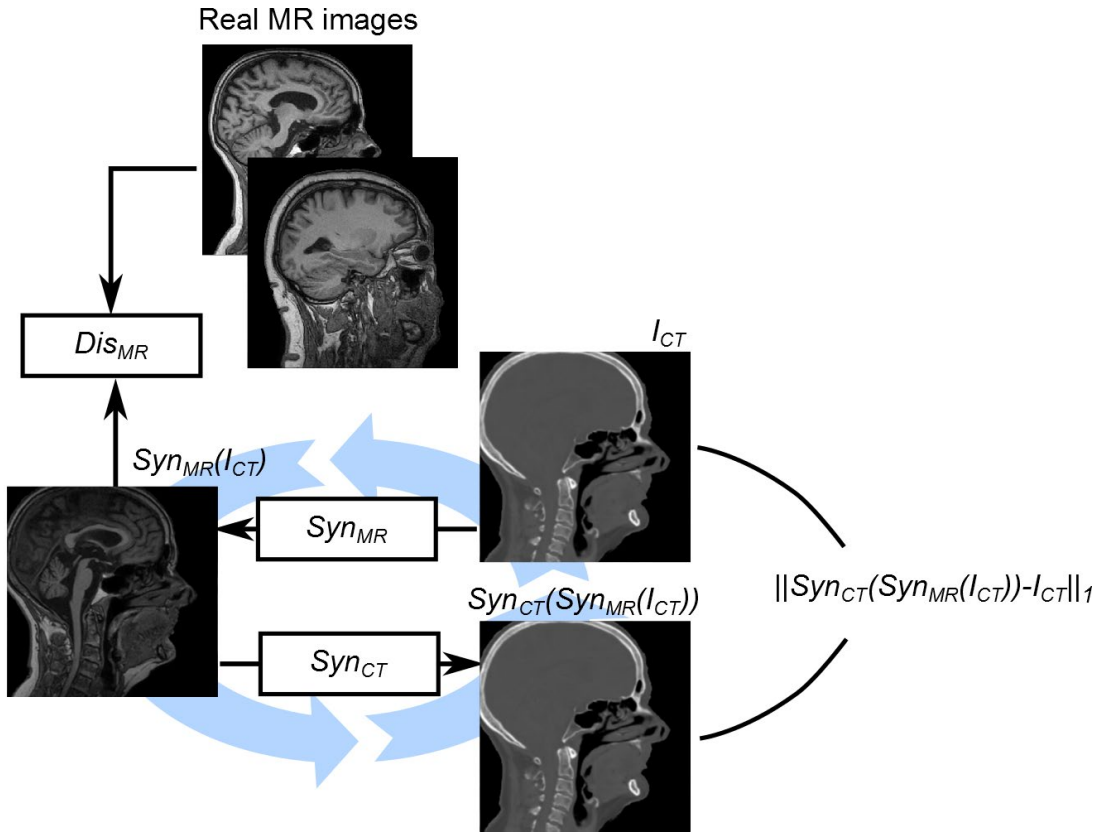


CT

Example: MR to CT synthesis

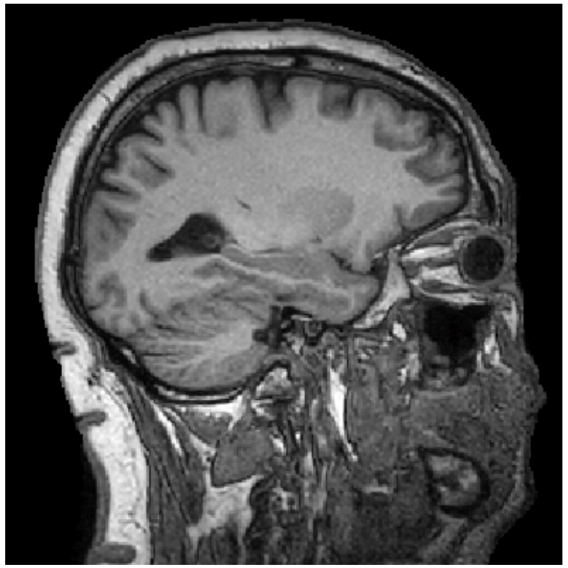


Forward cycle

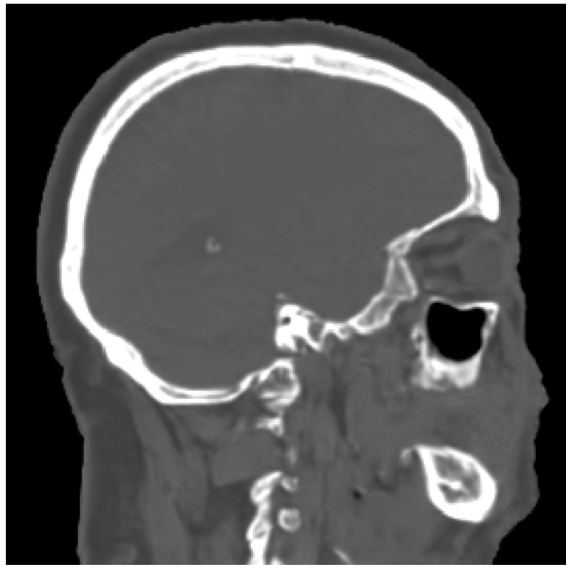


Backward cycle

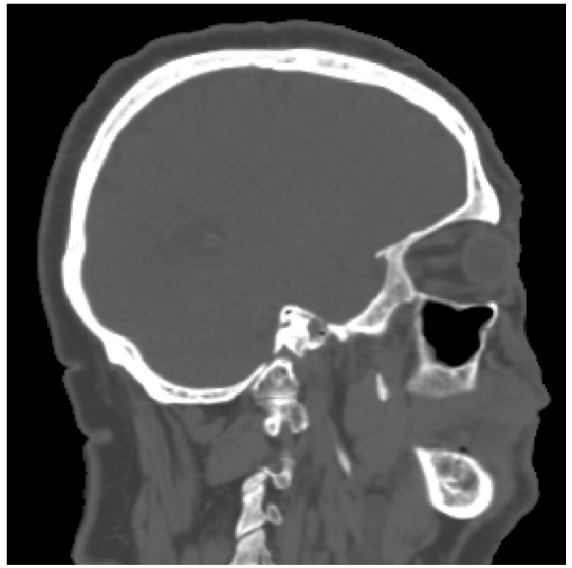
Example: MR to CT synthesis



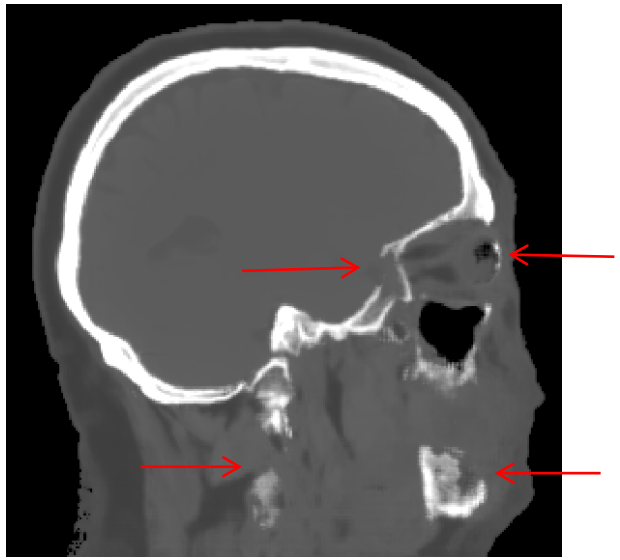
I_{MR}



$Syn_{CT}(I_{MR})$



I_{CT}



Regression

Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, ICCV 2017

Example: MR to CT synthesis

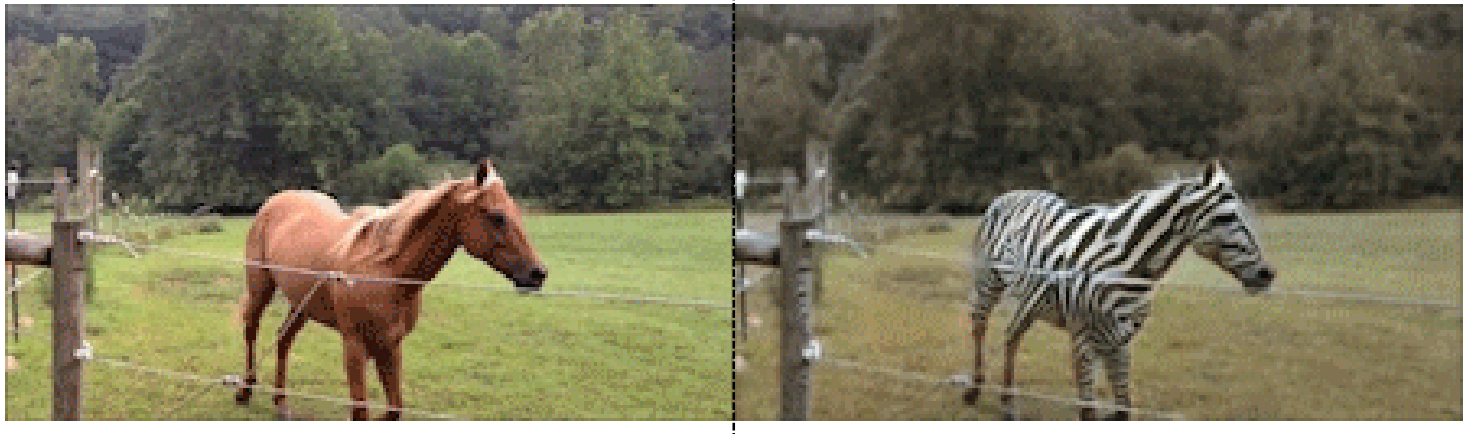
I_{MR} $Syn_{CT}(I_{MR})$ I_{CT} $|Syn_{CT}(I_{MR})-I_{CT}|$

Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, ICCV 2017

CycleGAN

Domain A

Domain B



https://github.com/junyanz/CycleGAN

Summary

- Interpretability + explainability important for practical ML use
- Generative adversarial networks are SOTA for image synthesis
- Conditional GANs synthesize samples with particular characteristics
- Adversarial networks can define loss functions that we cannot
- Many applications in medical image analysis

Practical assignment

Train your own GAN! https://tinyurl.com/capitaselectacolab

- Match a normal distribution
- MNIST digits (unconditional + conditional)
- Histopathology images: https://github.com/basveeling/pcam
- Run with free GPU: in Playground (no saving) or make copy in Google Drive (log in)
- Experiments with BigGAN https://tinyurl.com/y8yuqyqv