GENERATIVE AND DENSITY MODELS

Traditional Approaches

- parametric density estimation
- mixture models
- Bayesian networks
- causal models
- Boltzmann machine
- restricted Boltzmann machine
- Energy Based Models

Generative and Density Models

Important Applications:

- speech generation, game content generation: conditional generation of content based on some inputs
- image and video enhancement: upscaling, sharpening, interpolation, ray tracing, etc.

Largely Unfulfilled Potential:

- unsupervised learning for transfer learning
- training dataset generation and augmentation
- domain transfer (winter->summer, etc.)

FLOW MODELS

Flow Models Idea

- model some complicated $p(\xi)$
- given $x \sim p(x) = \mathcal{N}(0, \mathbf{1})$
- learn an invertible function $x = f_\theta(\xi)$

Flow Models - Mapping Densities

Example:

- linear mapping of a standard normal density $\mathcal{N}(0, 1)$ to a general normal density $\mathcal{N}(\mu, \Sigma)$
- requires changing the normalization factor b/c linear map is not volume preserving

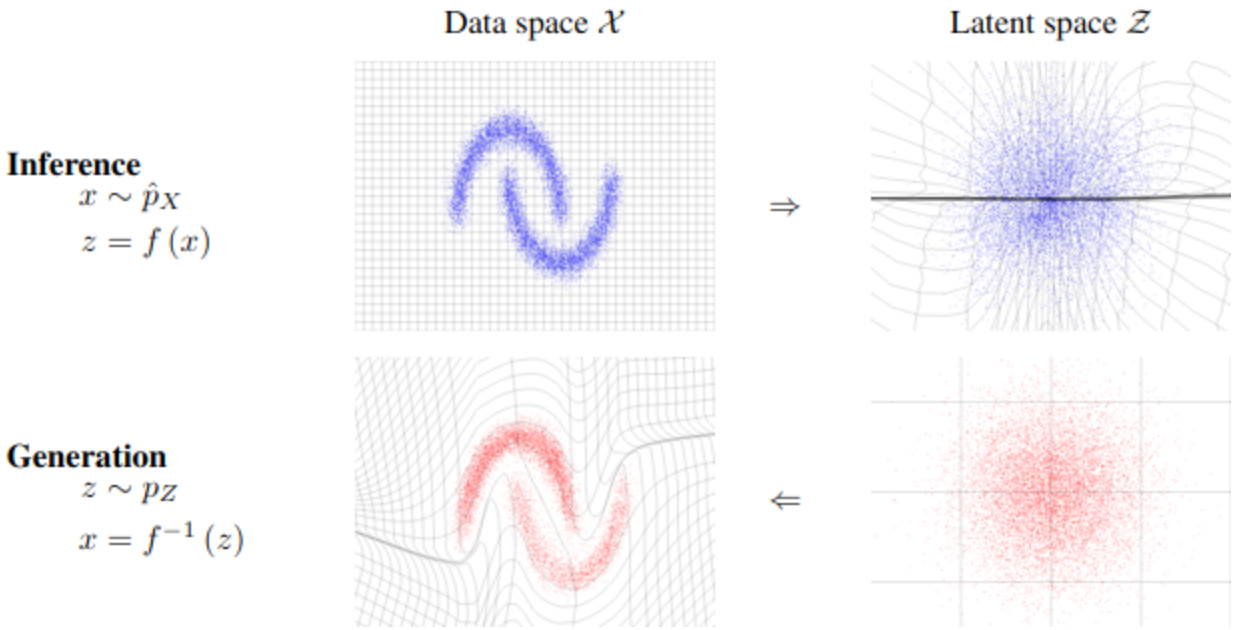
Generalization:

- the local linear relationship between x and ξ is given by the Jacobian J_{f_θ}
- we account for the volume change using the determinant of the Jacobian
- $p(x) = p(\xi) \cdot \det(J_{f_\theta}(\xi))$

Updating the loss function:

- $\arg\min_\theta \mathbb{E}_x[-\log p(x)] = \arg\min_\theta \mathbb{E}_x[-\log p(f_\theta(x)) - \log\det(J_{f_\theta}(\xi))]$

Flow Models - Illustration

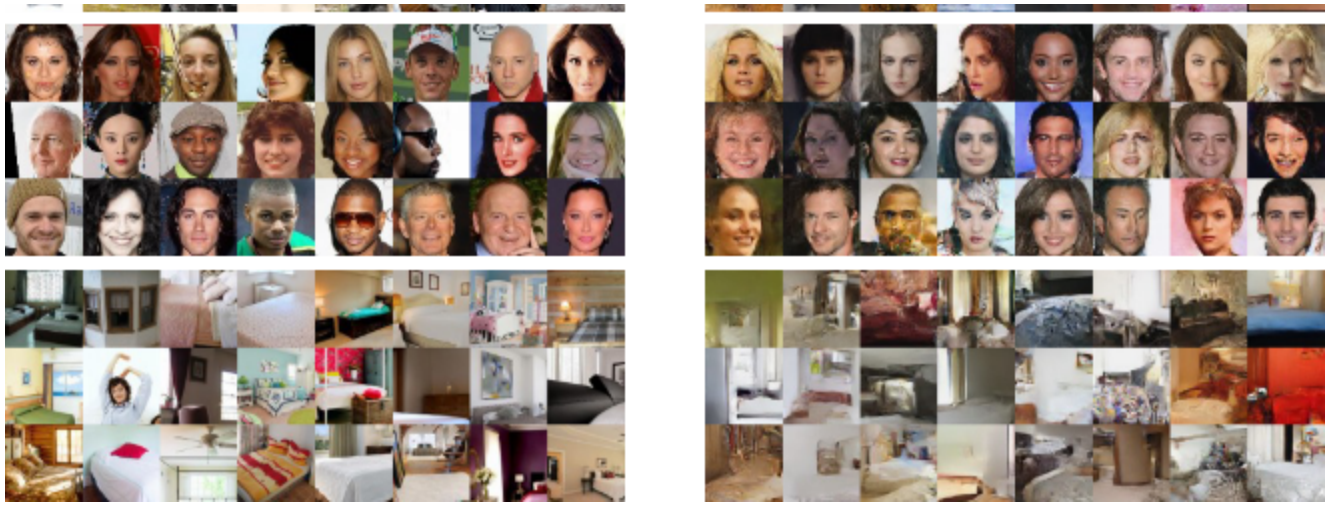


Dinh et al. ICLR 2017

Flow Models - Technicalities

- mappings compose: if $f = g \circ h$
  - then $p(x) = p(\xi) \cdot \det(J_g(h(\xi))) \det(J_h(\xi))$
- regular linear layers are too costly to invert, use affine coupling layers instead (NICE/RealNVP)
  - split input to layer in two halves $x = (x_l, x_h)$
  - $y_l = x_l$
  - $y_h = x_h \odot f_\theta(x_l) + g_\theta(x_l)$ (elementwise affine)
  - the Jacobian is triangular, so its determinant is the product of the diagonal elements
- need to "dequantize" images
- other options: mixtures of logistics, continuous time flows, ...

Flow Models - Examples



Dinh et al. ICLR 2017

VAE

Variational Autoencoder

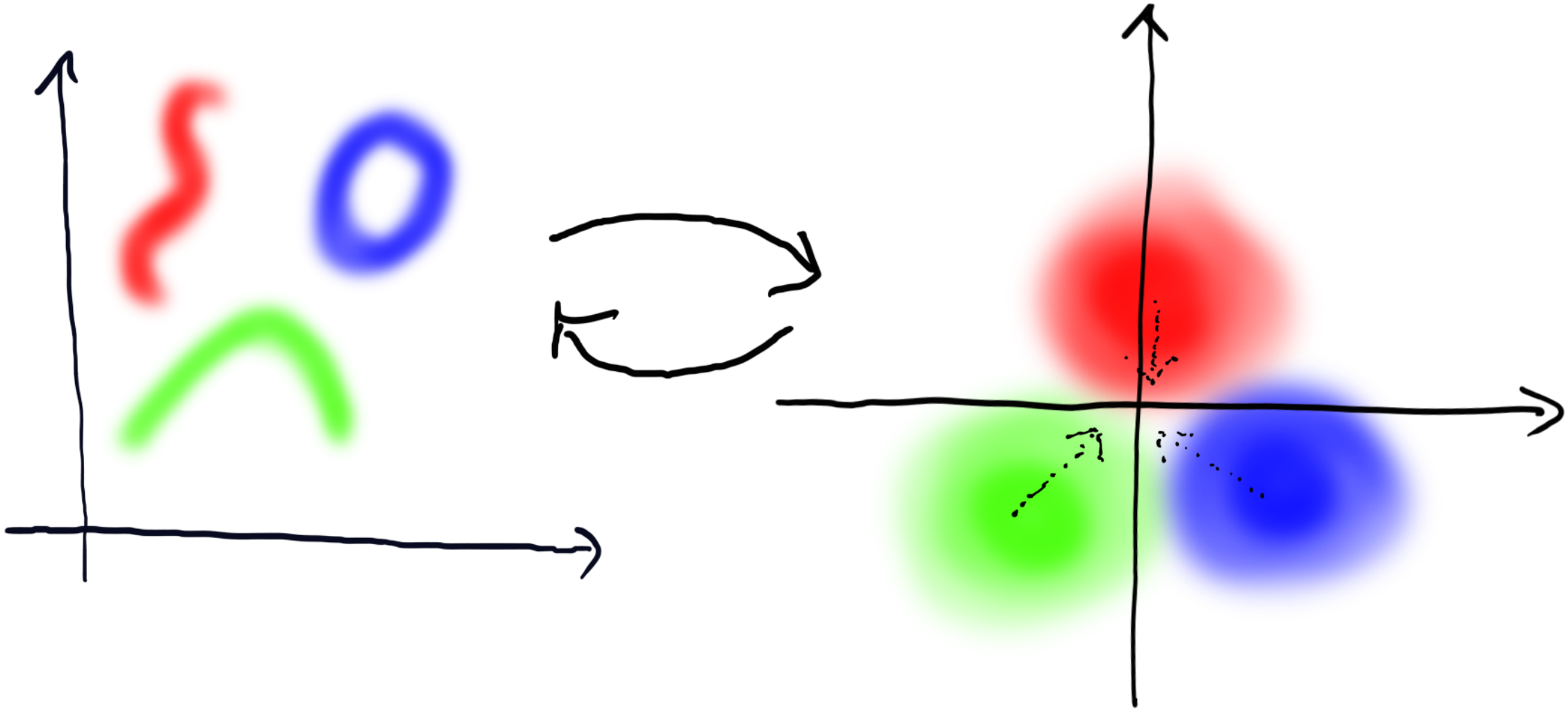
Problem using Autoencoders for Generative Models:

- tends not to generate useful latent space representations
- small changes in latent space not guaranteed to correspond to small changes in output

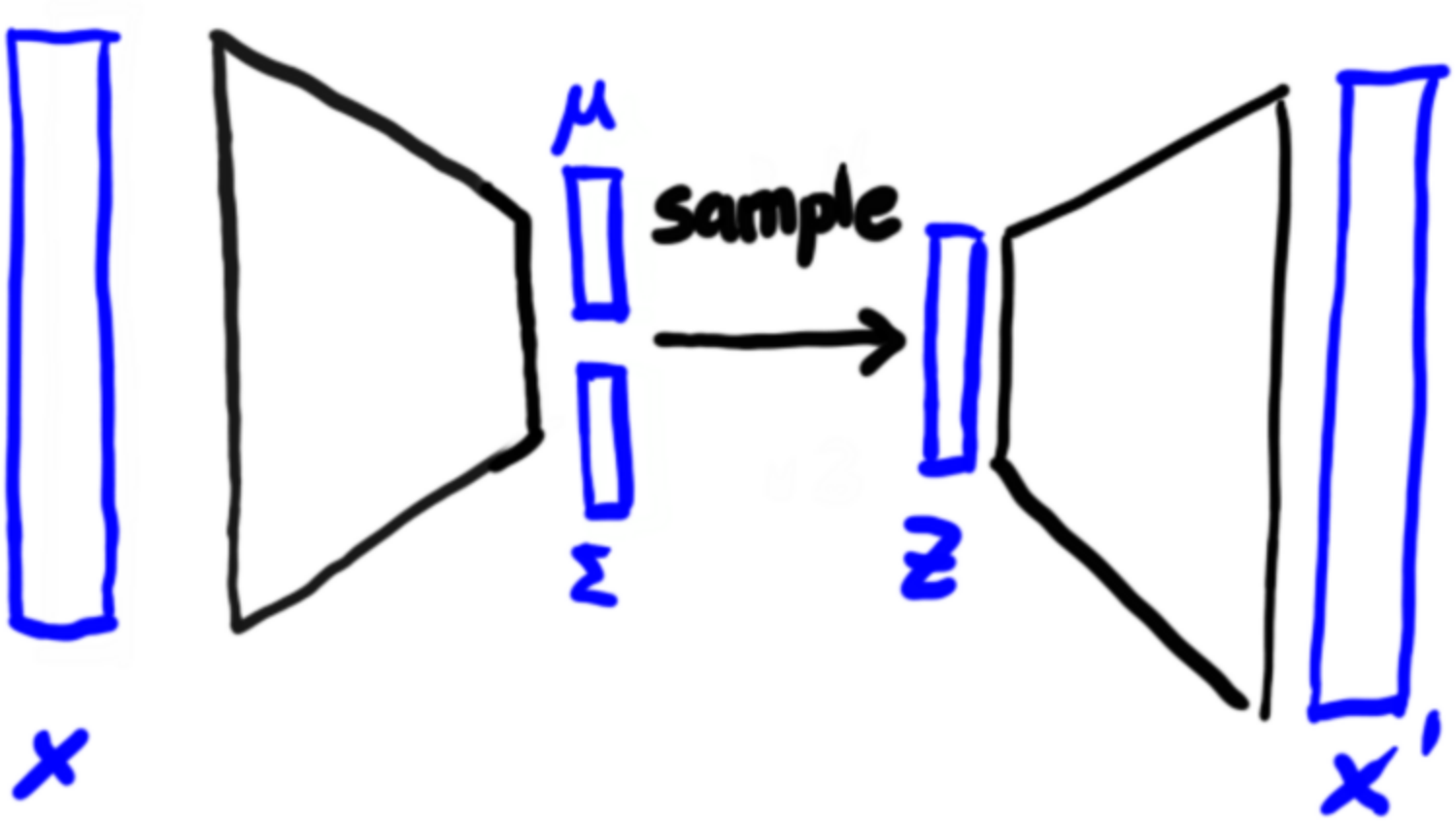
Solution:

- map each input vector x to a conditional distribution over y
- specifically
  - output parameters for a parametric distribution $p_\phi(y|x)$
  - choose $p_\phi(y|x) = \mathcal{N}(y, \mu_x, \Sigma_x)$
  - move $p_\phi(y|x)$ towards $\mathcal{N}(0, \mathbf{1})$
- reconstruct x by sampling from that conditional distribution

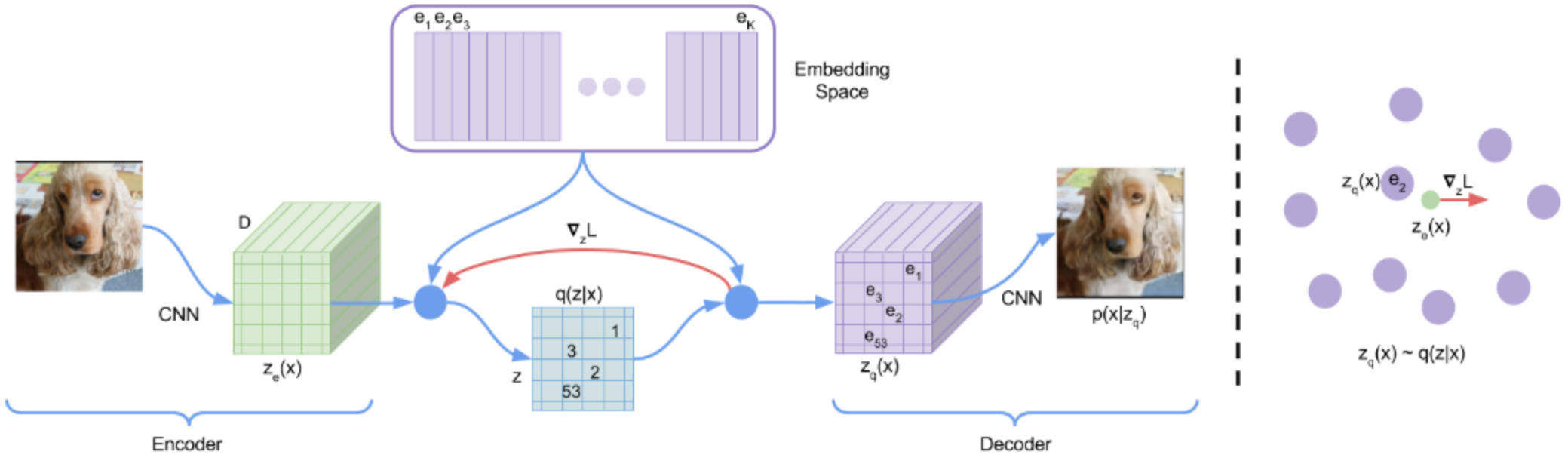
VAE Objective



VAE Structure



VQ-VAE



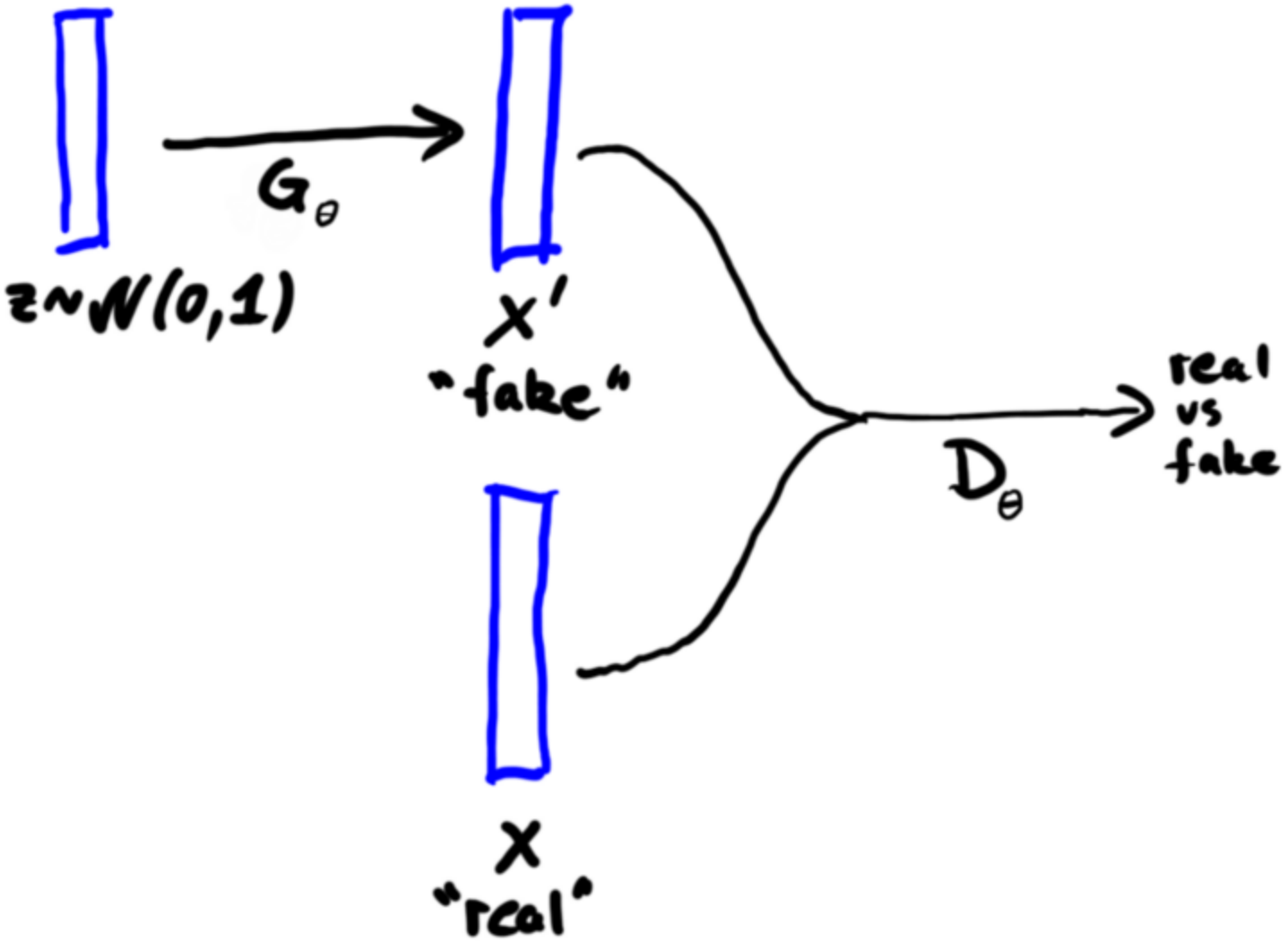
VQ-VAE uses a discrete latent space with a "language model", giving rise to a rather different kind of model; van den Oord et al., Arxiv 2018

GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Models

- sampling-only models
- sample from an underlying distribution $z \sim p(z)$
- compute output as $f_\theta(z)$, inducing some $p_\theta(x)$
- for other models, we tried to make $p_\theta(x)$ close to the data
- with GANs, we cannot access $p_\theta(x)$, only sample from it
- a kind of "game" in which a generator tries to fool a discriminator

GAN

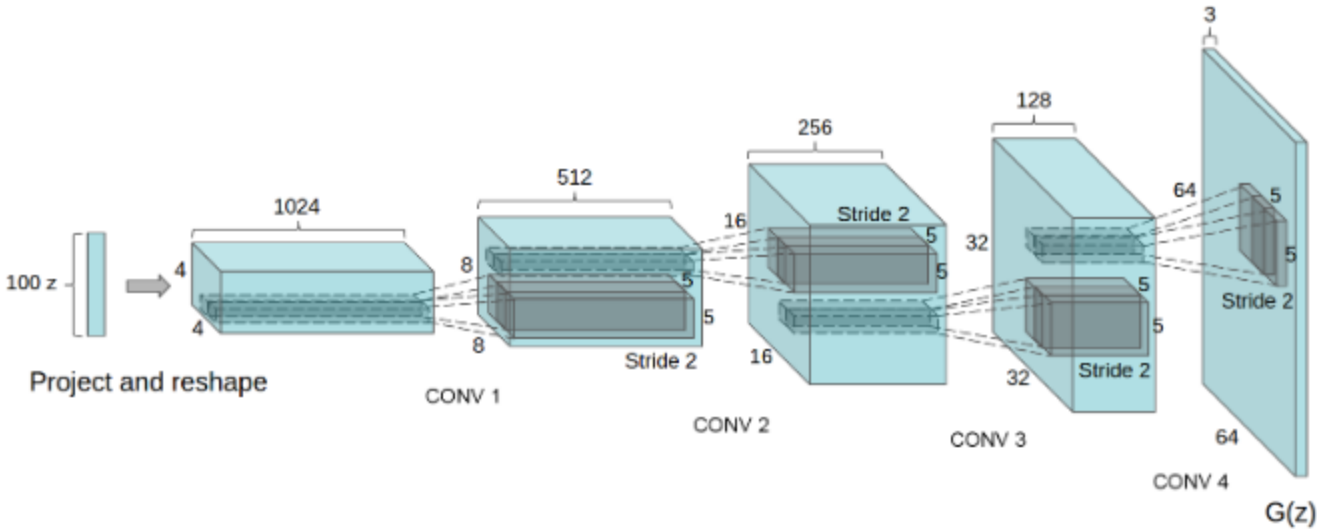


Major Difficulties Implementing GAN

- low signal from a single global discriminator
  - solution: use patchwise/pixel-wise discrimination
- discriminator easily saturates and generates no usable deltas
  - solution: carefully choose discriminator architecture and limited training
- mode collapse: if $p(x)$ is multimodal, model may sample a single mode
  - solution: use Wasserstein GAN (roughly: discriminators w/bounded Lipschitz coefficients)

Original GAN Results

Deep Convolutional GAN (DCGAN)



Radford et al., 2016; Arxiv 1511.06434

DCGAN - Results

Radford et al., 2016; Arxiv 1511.06434

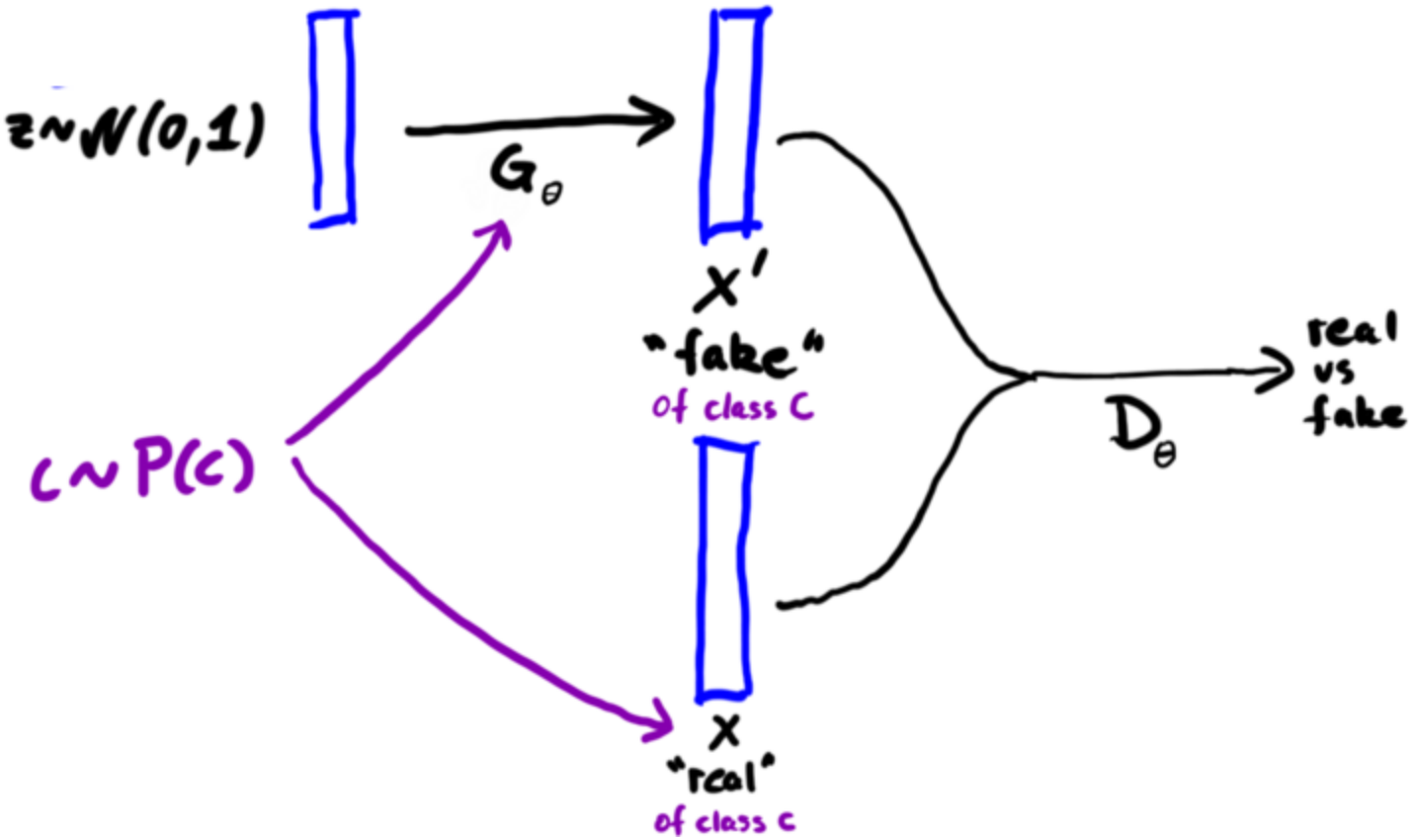
DCGAN - Representation Learning / Features

Table 1: CIFAR-10 classification results using our pre-trained model. Our DCGAN is not pre-trained on CIFAR-10, but on Imagenet-1k, and the features are used to classify CIFAR-10 images.

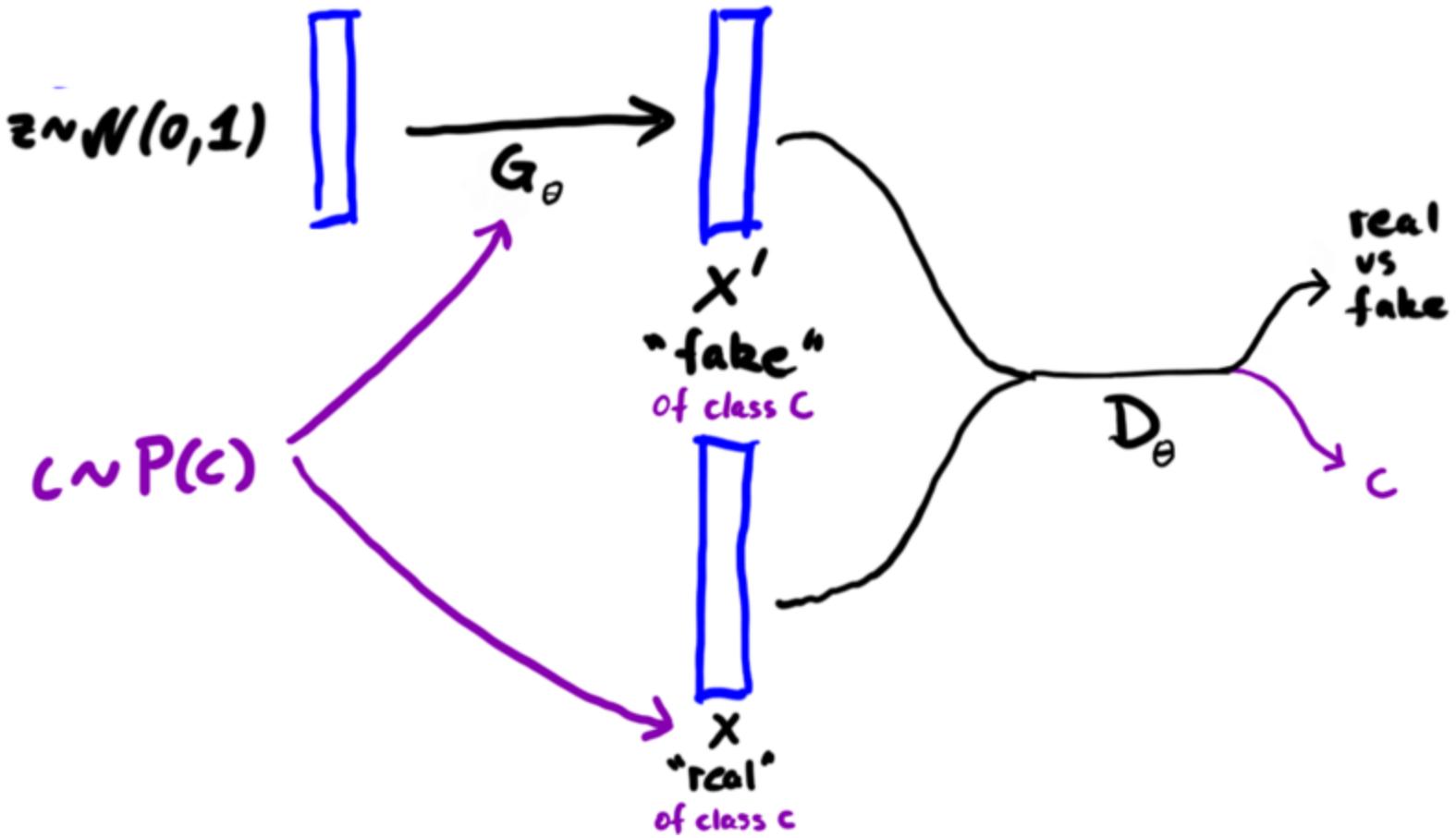
Model	Accuracy	Accuracy (400 per class)	max # of features units
1 Layer K-means	80.6%	63.7% (±0.7%)	4800
3 Layer K-means Learned RF	82.0%	70.7% (±0.7%)	3200
View Invariant K-means	81.9%	72.6% (±0.7%)	6400
Exemplar CNN	84.3%	77.4% (±0.2%)	1024
DCGAN (ours) + L2-SVM	82.8%	73.8% (±0.4%)	512

Radford et al., 2016; Arxiv 1511.06434

Conditional GAN

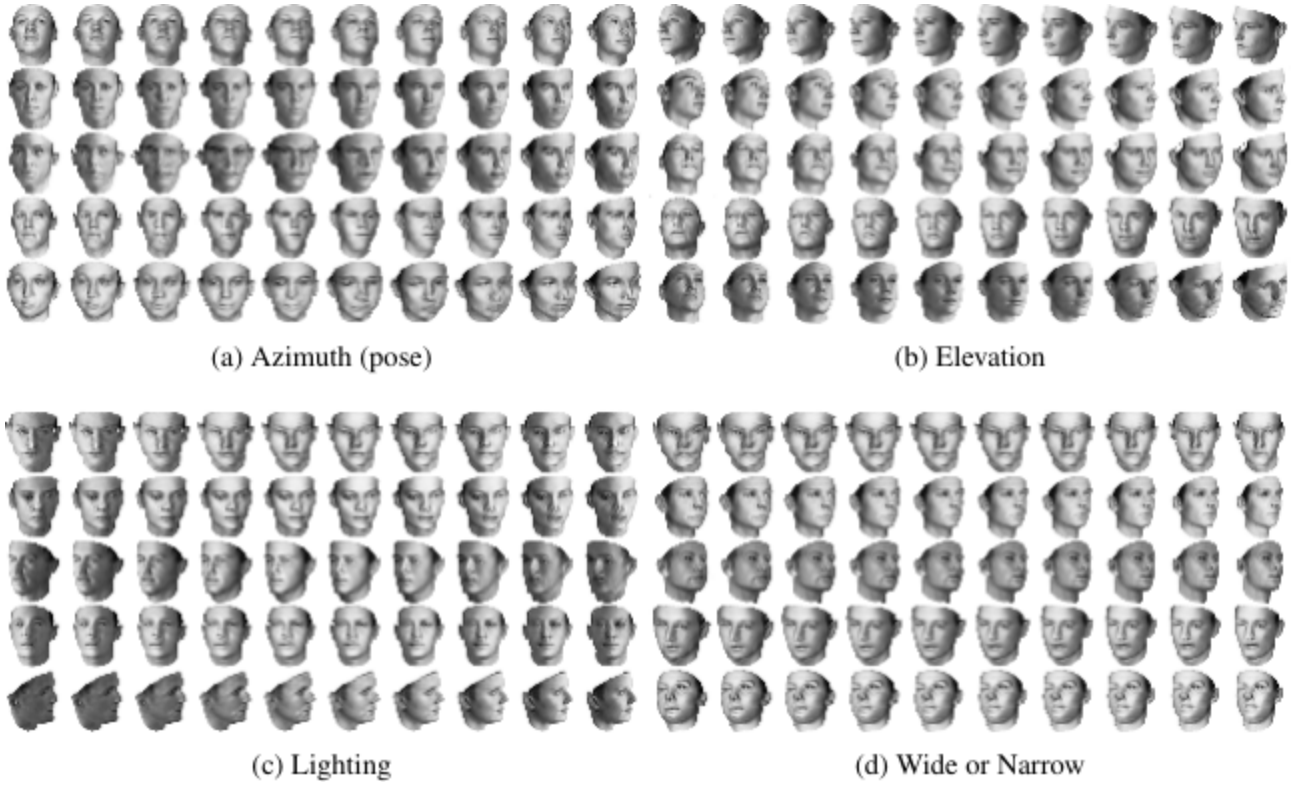


InfoGAN



Chen et al., 2016; arXiv:1606.03657.

InfoGAN - Disentangling Latent Codes



Chen et al., 2016; arXiv:1606.03657.

BigGAN Results

Large bag of tricks. Brock et al. 2018.

DISTRIBUTION ALIGNMENT

Problem

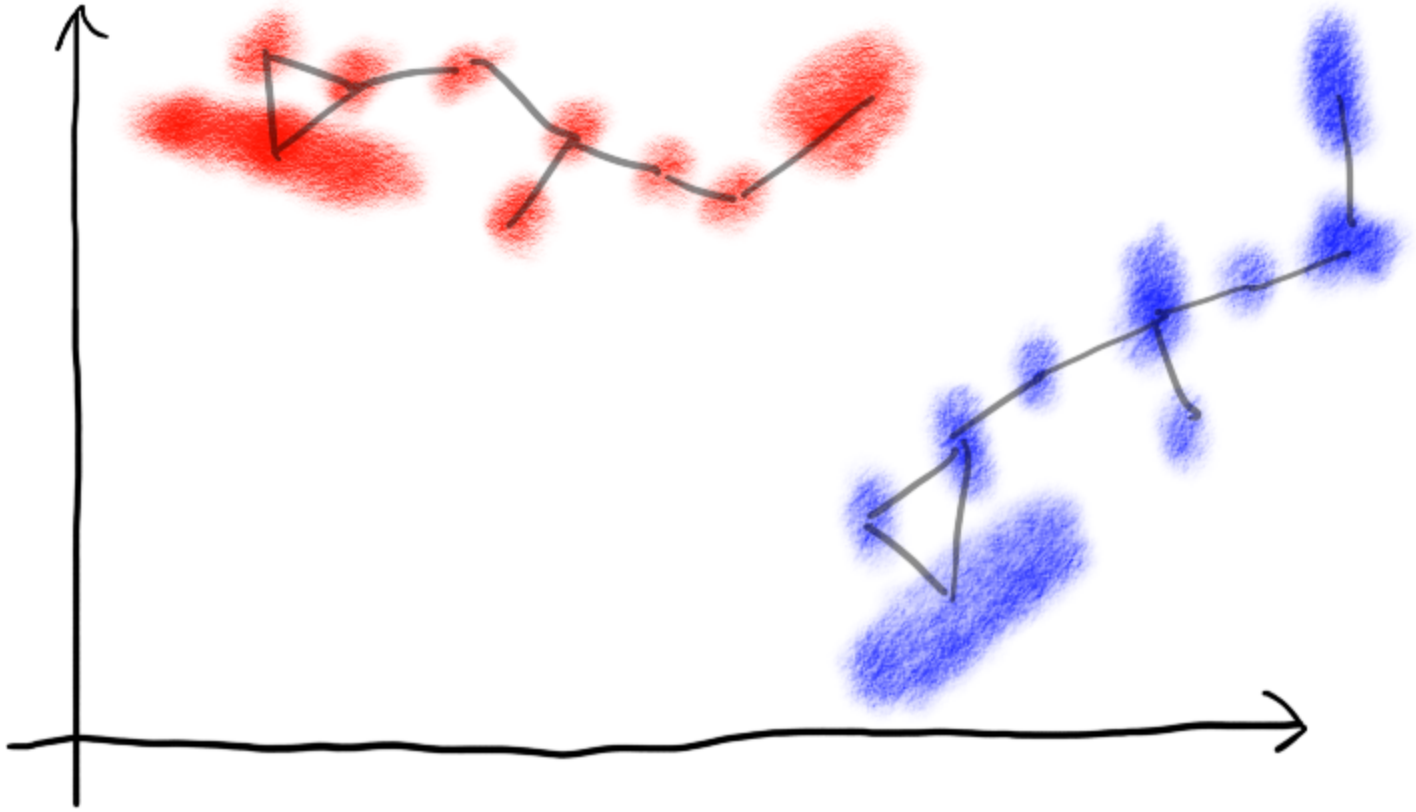
Assume we have two different but related distributions $p(x)$ and $p'(x)$ but no corresponding samples.

Examples:

- daytime vs nightime images
- outdoors scenes in sunshine vs rain
- photographic images vs oil paintings
- clean binary document image vs photographic capture

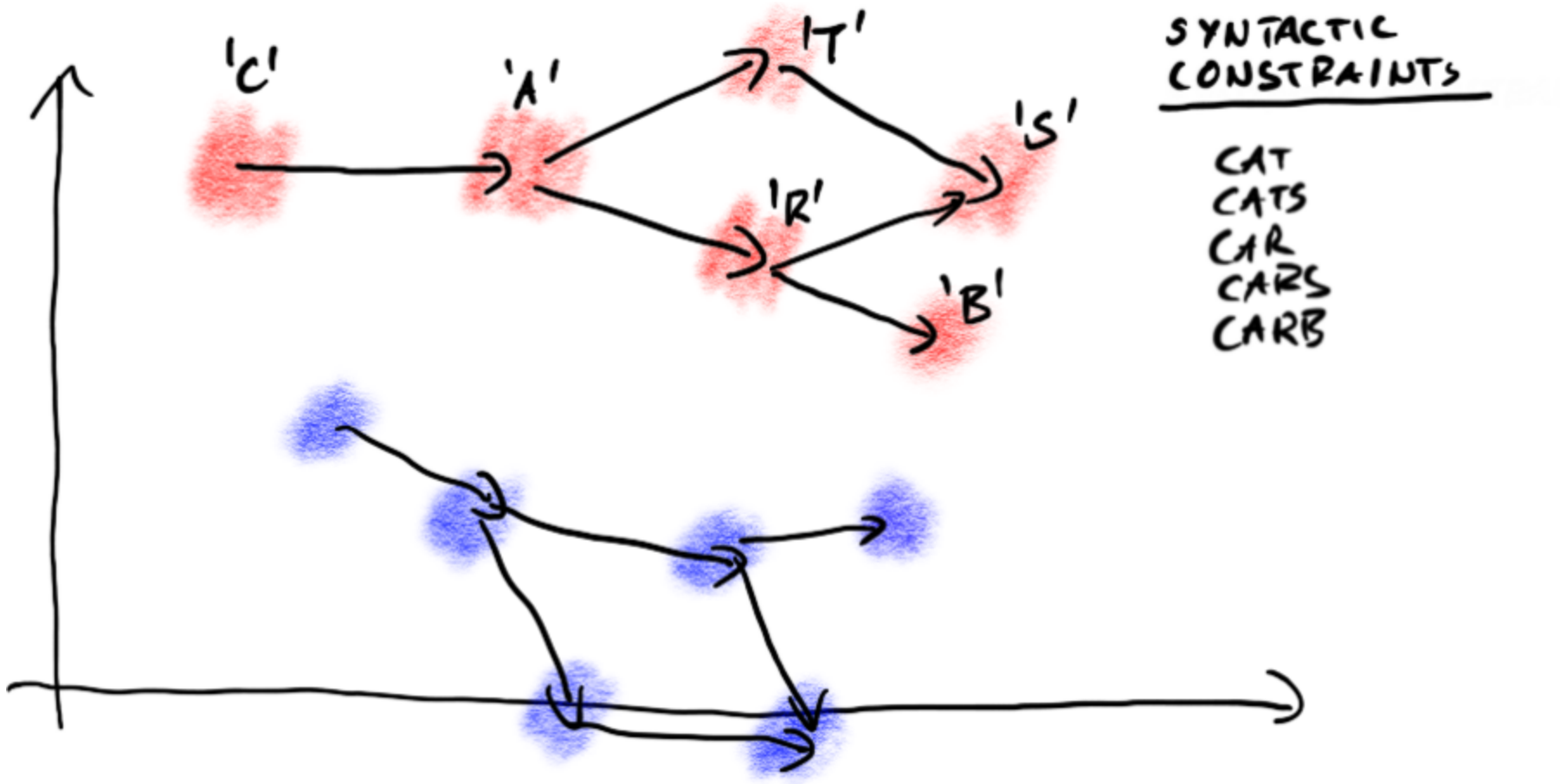
Can we map between these two domains given just independent samples from each? Sometimes.

Alignment of Mixture Densities (no other info)



- distribution alignment is possible for distributions with a lot of structure
- ... if the "general shape" of the distribution is preserved

Alignment of Mixture Densities (with sequences)



- syntactic constraints (e.g. sequences of samples) help with distribution alignment
- this is what EM training for OCR, speech, etc. tends to do

Uses of Distribution Alignment for Training

E.g. AV images: summer and winter, lots of summer training data, little winter training data

Approach 1:

- train model on summer + translate winter to summer

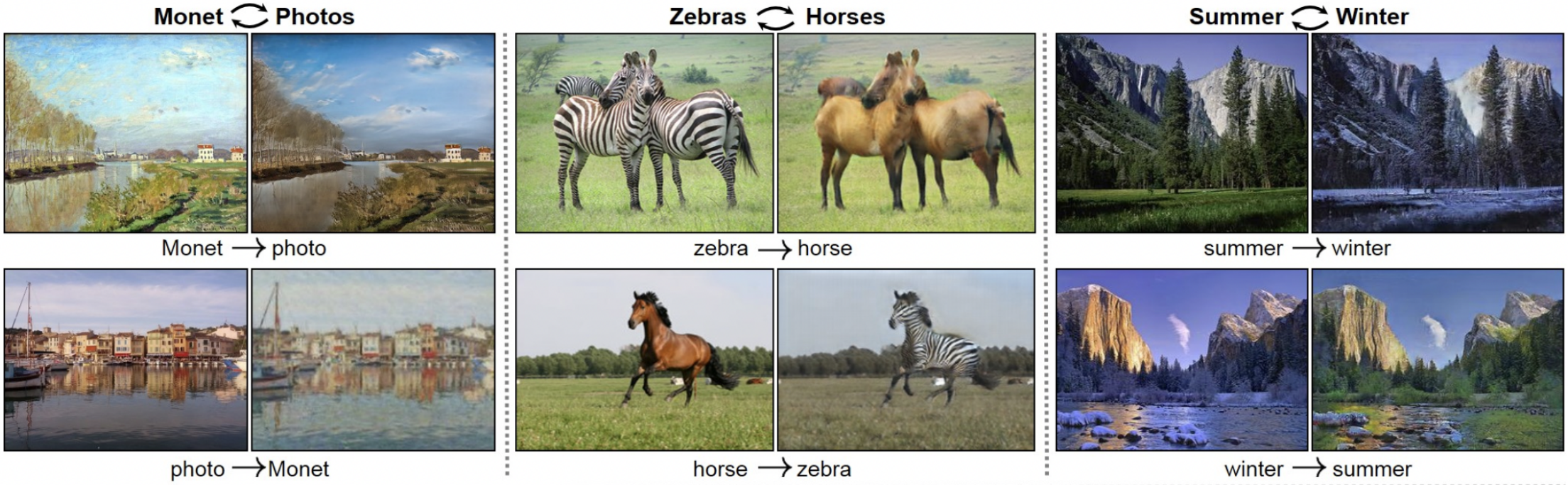
Approach 2:

- use distribution alignment to generate lots of winter images from summer images, then augment the training set with that
- train model on summer images and artificially generated winter images

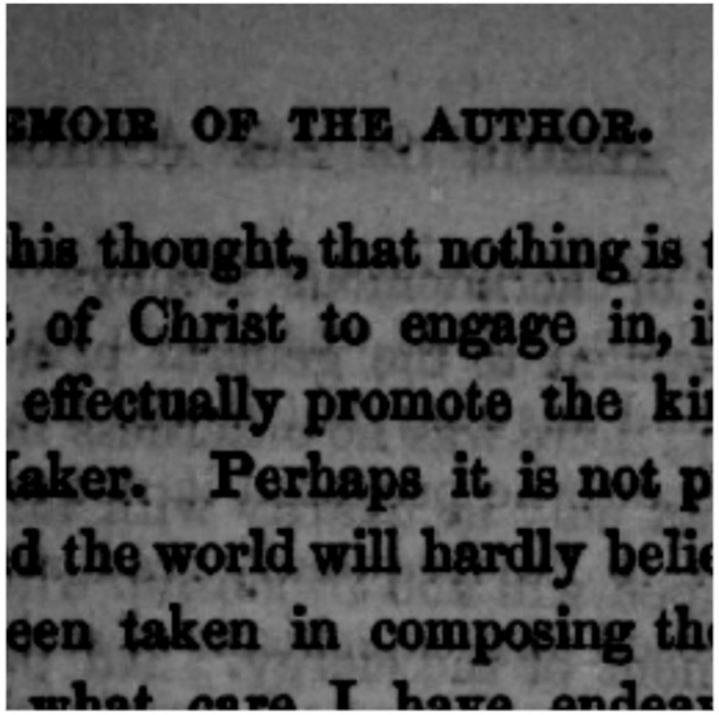
CycleGAN

- two generators $G' : x \rightarrow x'$, $G : x' \rightarrow x$
- two discriminators: D' (discriminates real/fake x'), D (discriminates real/fake x)
- train G, D and G', D' like a regular GAN
- require that $G' \circ G$ and $G \circ G'$ behave like identities

CycleGAN Examples

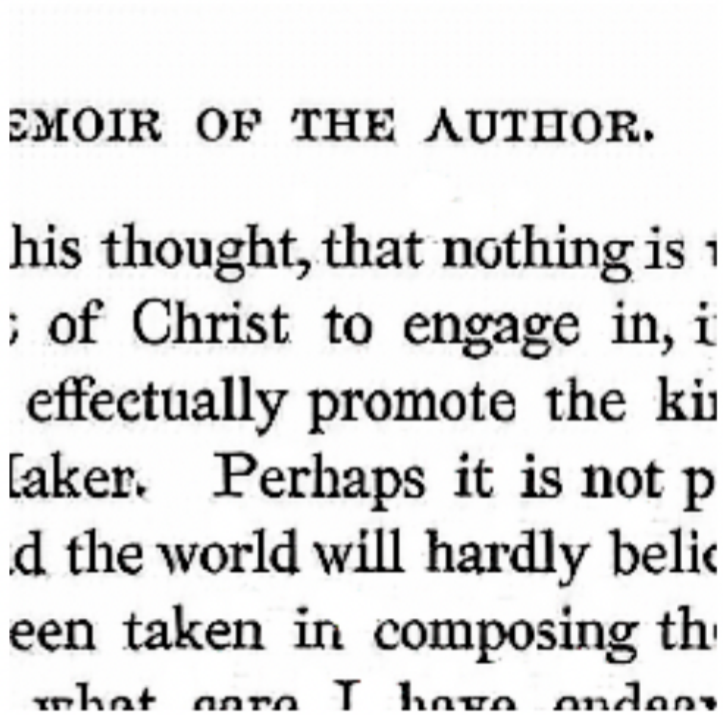


CycleGAN Examples



g_{AB}

g_{BA}



d_A d_B

Ongoing and Future Research

- extensions to text, speech, music, video, etc.
- condition on language (of course, already lots of news coverage)
- engineering work to remove artifacts and make output indistinguishable from real
- create forensic tools that can detect fakes
- combine transformers with each of the different generative approaches
- use non-generative unsupervised pretraining to help improve generative models

Limitations

It is an open question whether generative models will ever be useful unsupervised models for transfer learning to discriminative or other generative models.

High quality generative output requires modeling a large amount of detail that is irrelevant to most tasks.