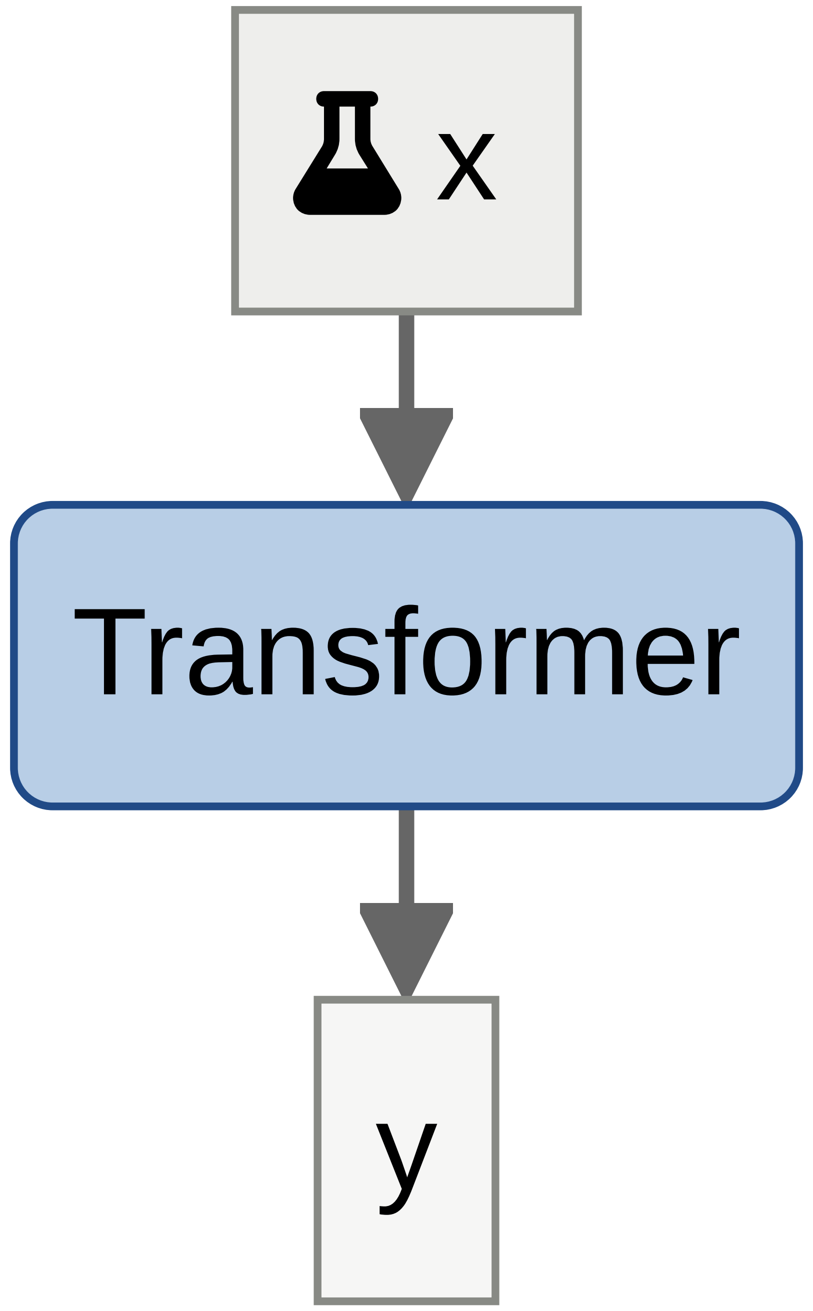
Generative Models

Philipp Krähenbühl, UT Austin

Recap

Model



Data

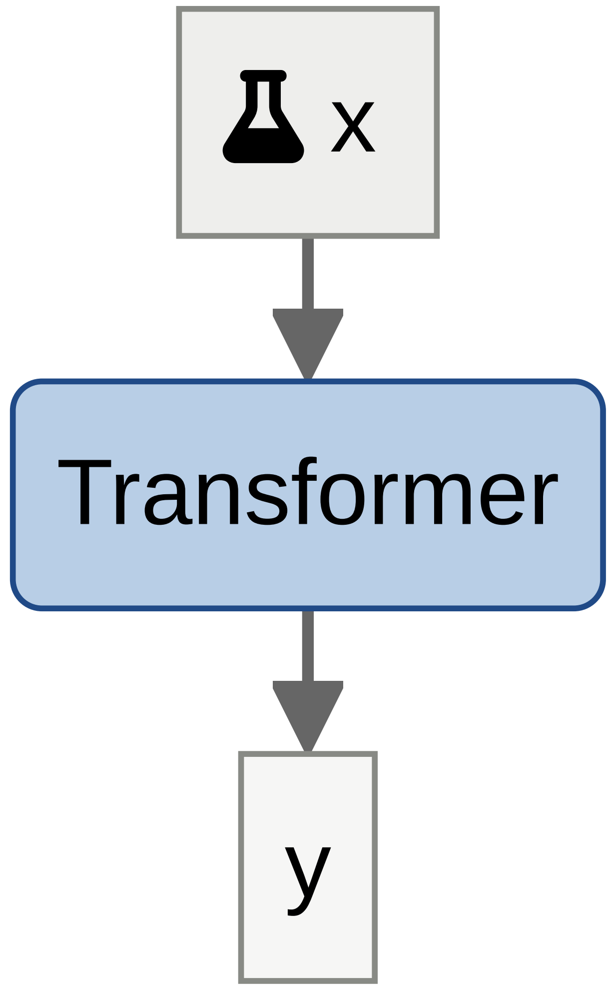
Optimization

```
m = 0
for epoch in range(n):
  for (x, y) in dataset:
    J = ∇l(θ|x,y)
    m = J + momentum * m
    θ = θ - ϵ * m.mT
```

Recap: How to train a network?

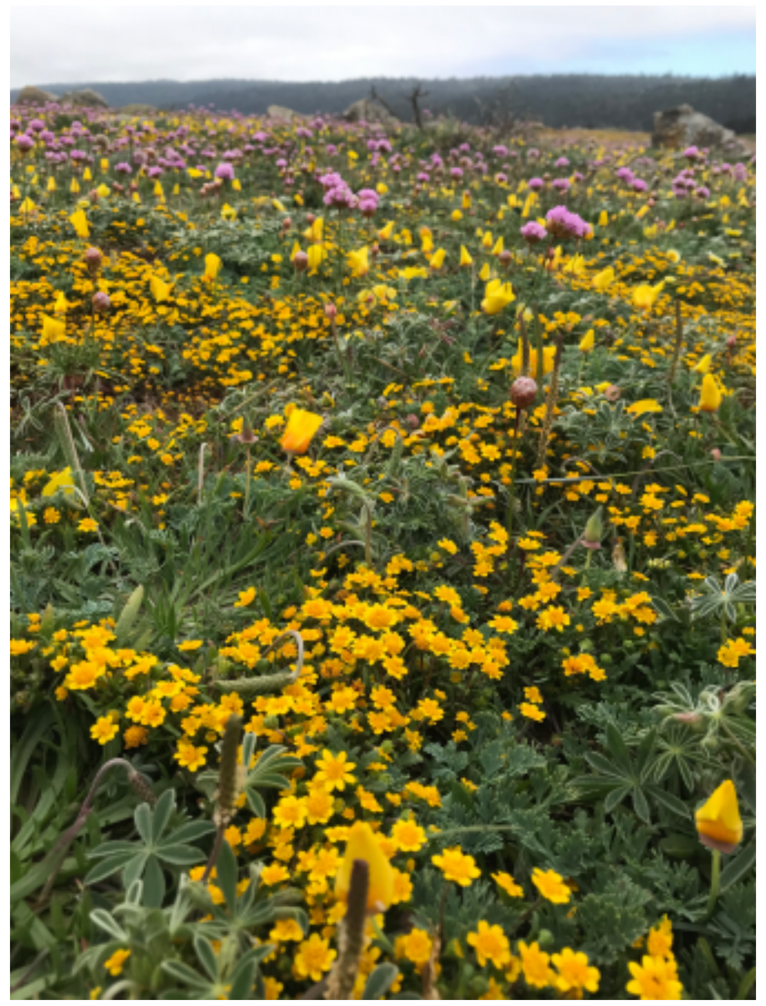
Collect Data

Design / download architecture



Train model

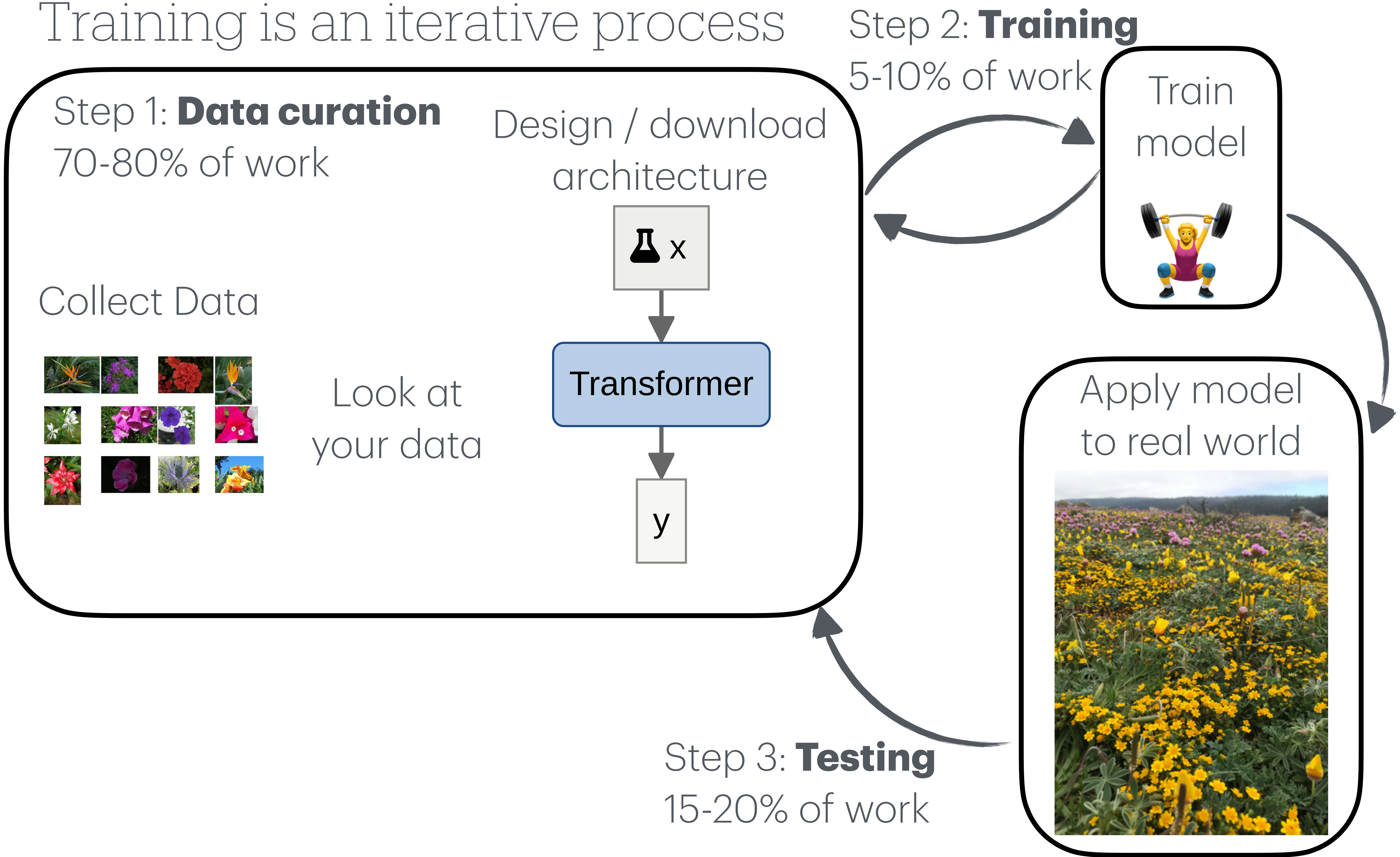
Apply model to real world



This never works !!!

Recap: How to train a network?

Training is an iterative process



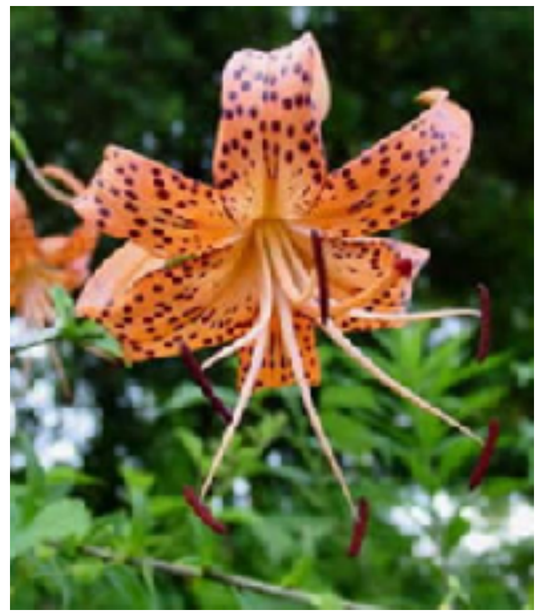
Part I: Done

Data

x_i, y_i

$\ldots$

pink primrose

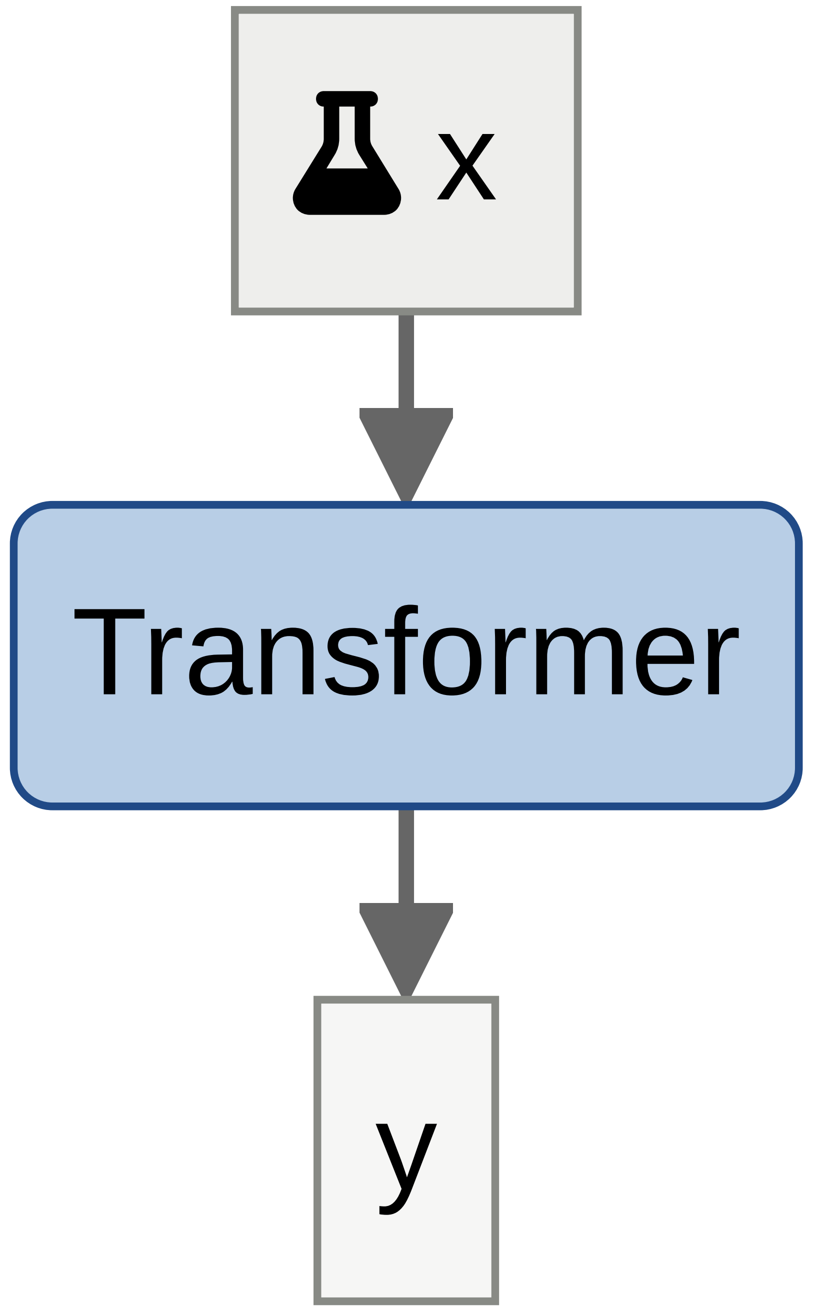


tiger lily

...

Train model

$f: x \rightarrow y$

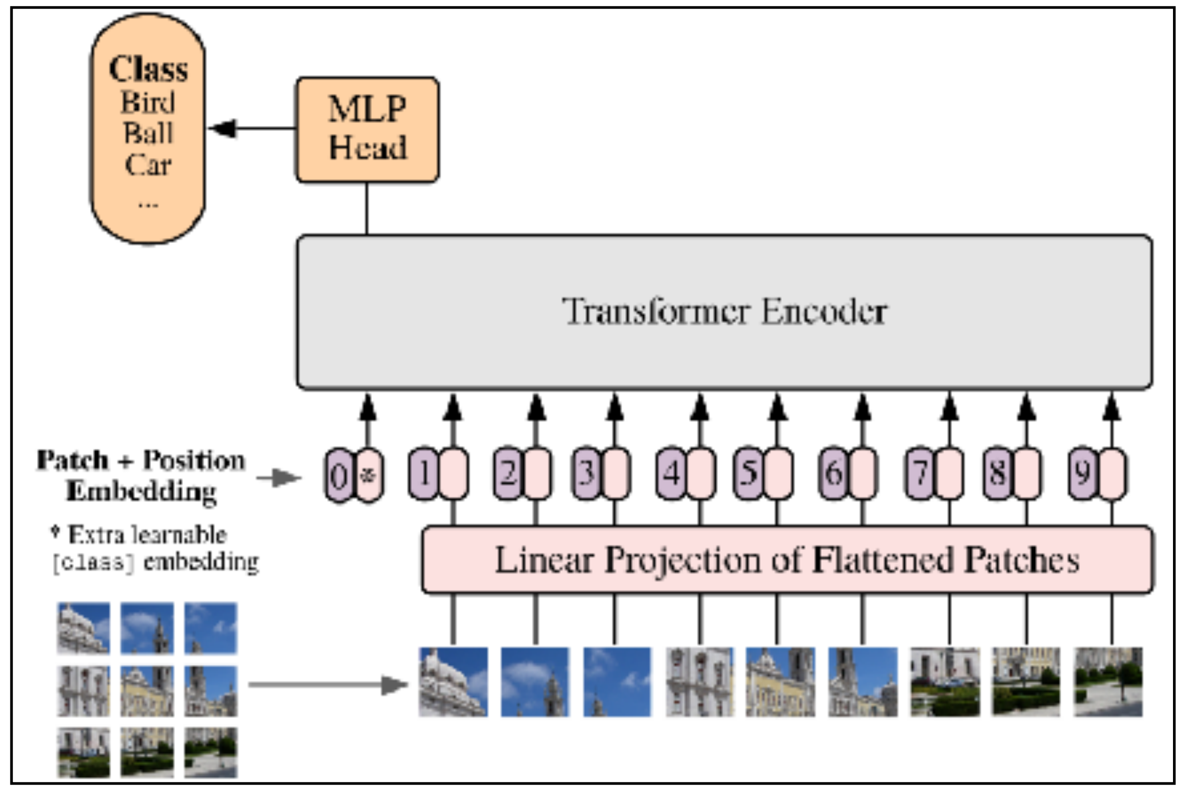


Generative Models

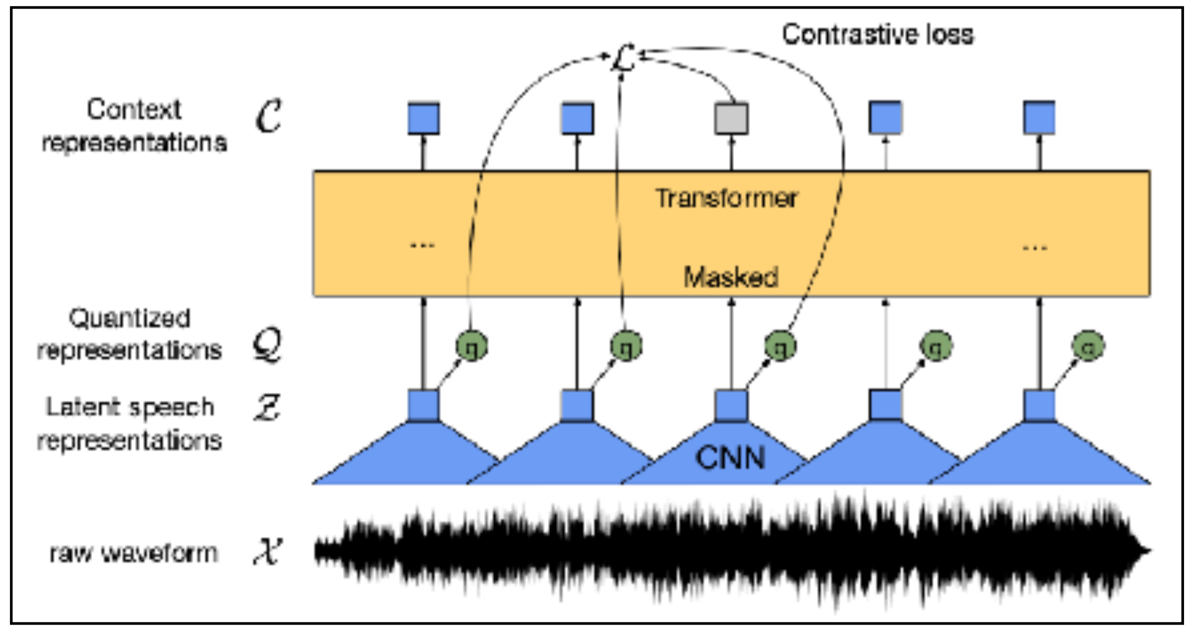
Philipp Krähenbühl, UT Austin

Discriminative models

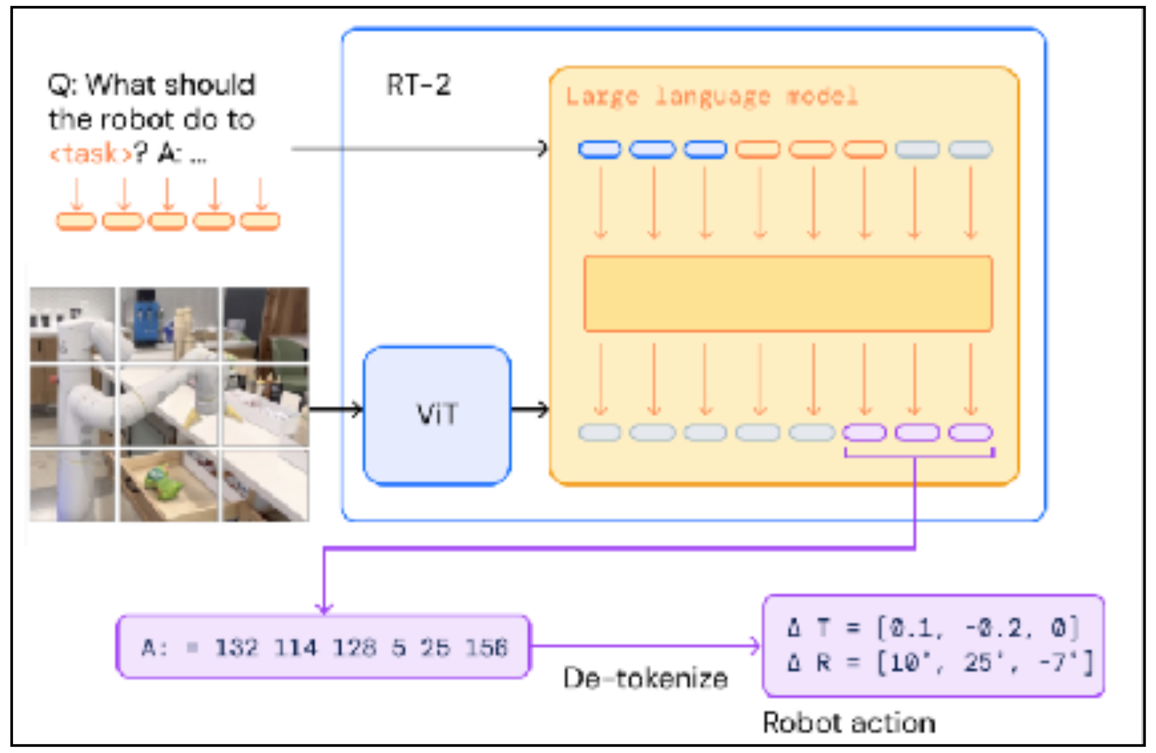
- Discriminative model: $P(Y|X)$
- Examples:
  - Image/video recognition
  - Speech recognition
  - Control policies
  - Weather prediction
  - ...



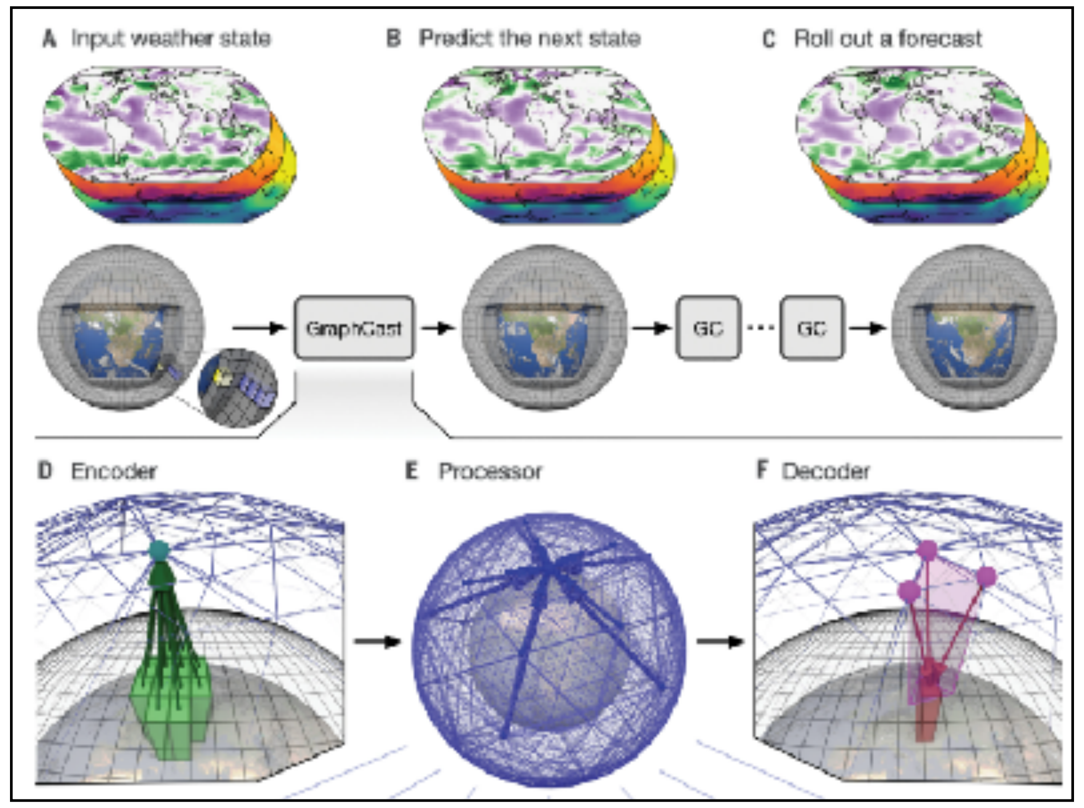
[1] Vision Transformer



[2] Wave2vec 2.0



[3] RT-2

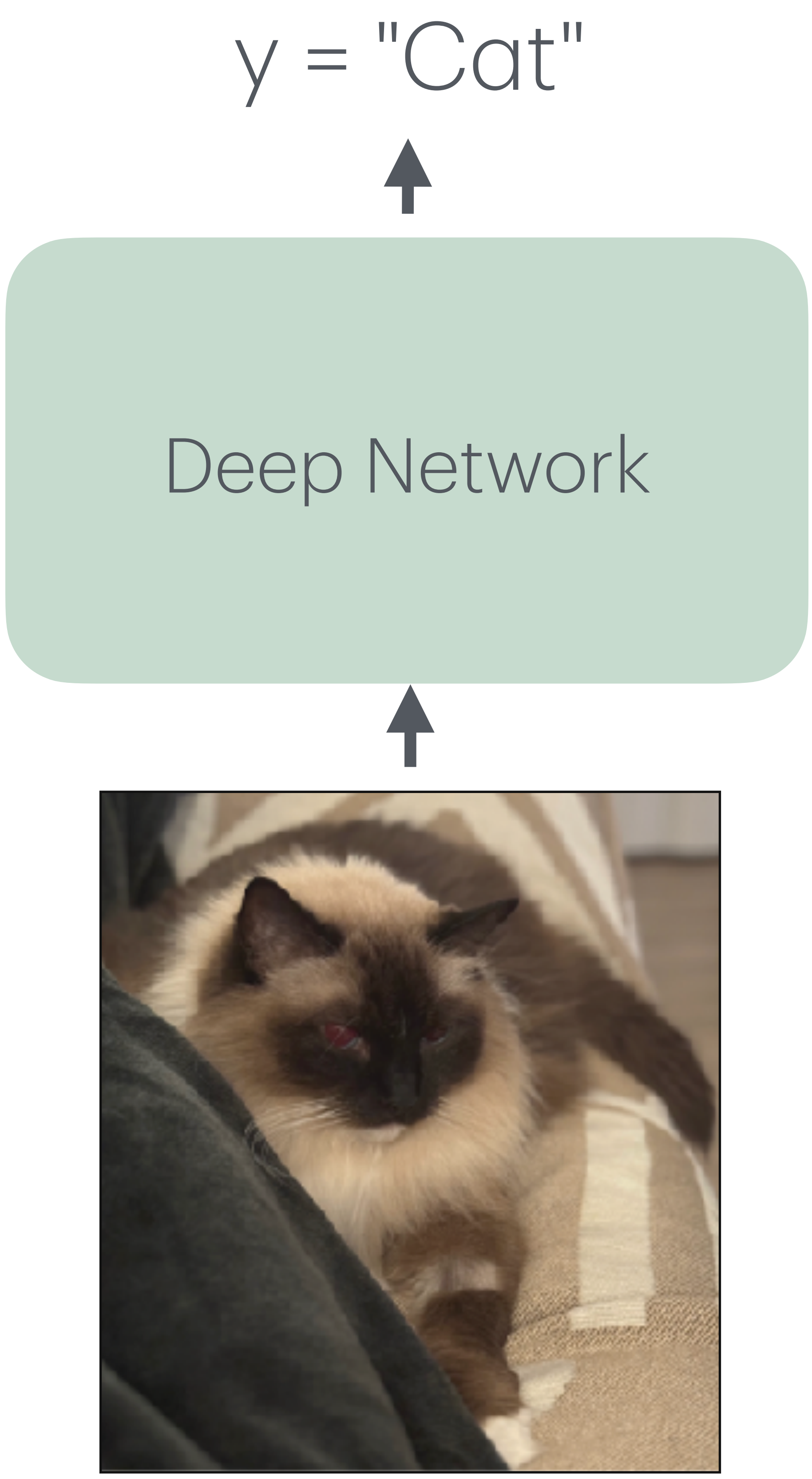


[4] GraphCast

[1] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.
[2] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.
[3] Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." arXiv preprint arXiv:2307.15818 (2023).
[4] Remi Lam et al. ,Learning skillful medium-range global weather forecasting.Science382,1416-1421(2023).

Discriminative models in deep learning

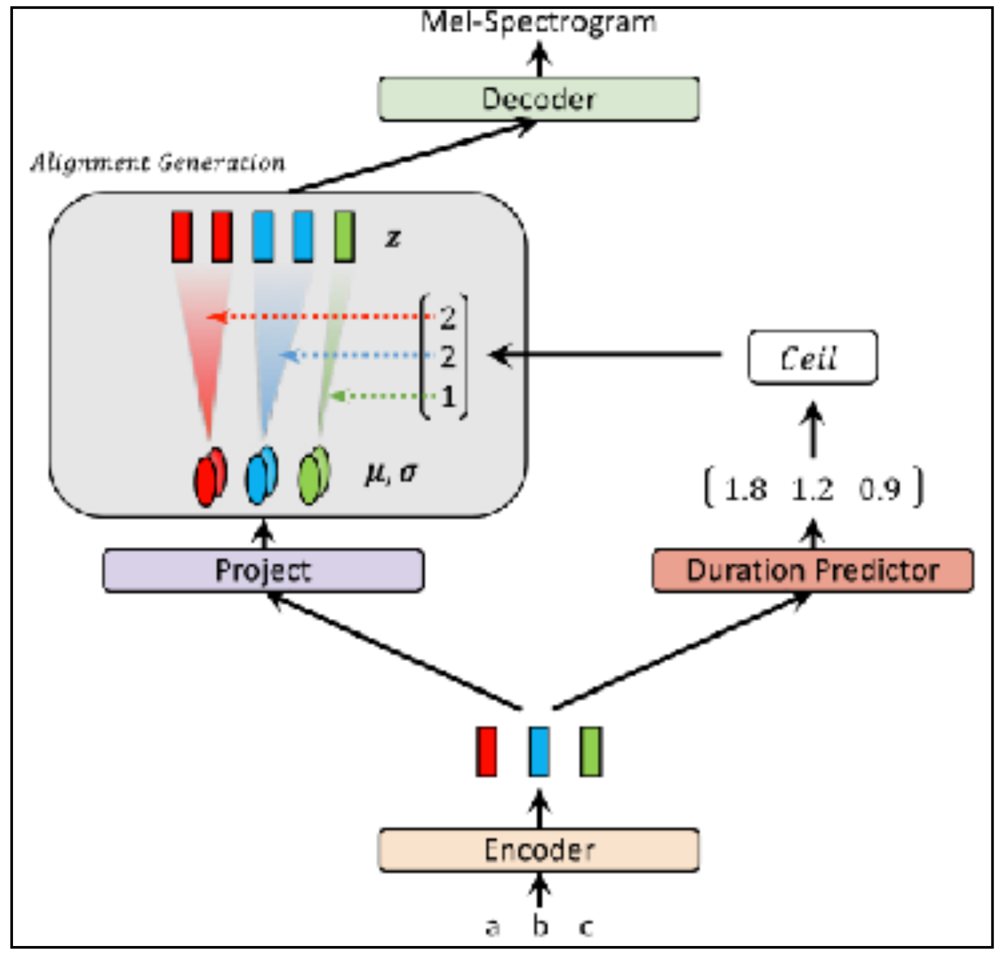
- Discriminative model: $P(Y|X)$
- Examples:
  - Image/video recognition
  - Speech recognition
  - Control policies
  - Weather prediction
  - ...



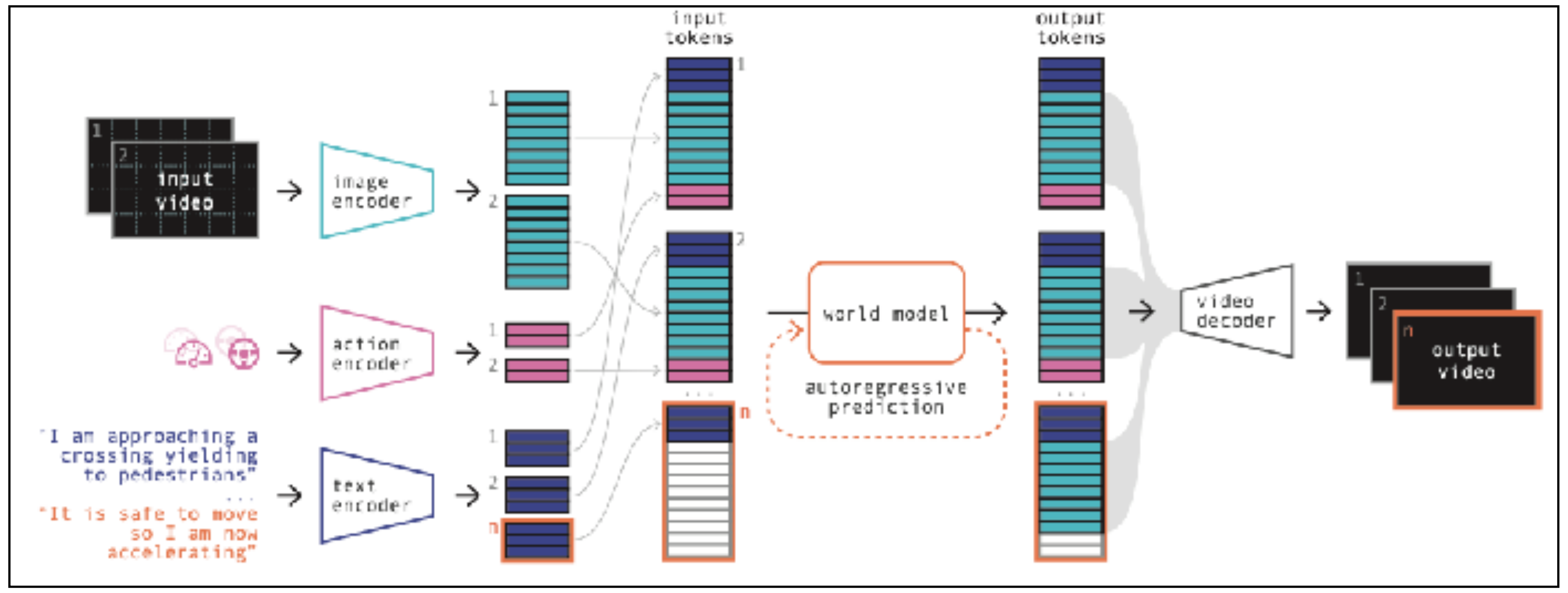
Generative models

- Generative model: $P(X)$
- Examples:
  - Image/video generation
  - Speech synthesis
  - Physics simulation / world modeling
  - Weather simulation (gaming)
  - ...

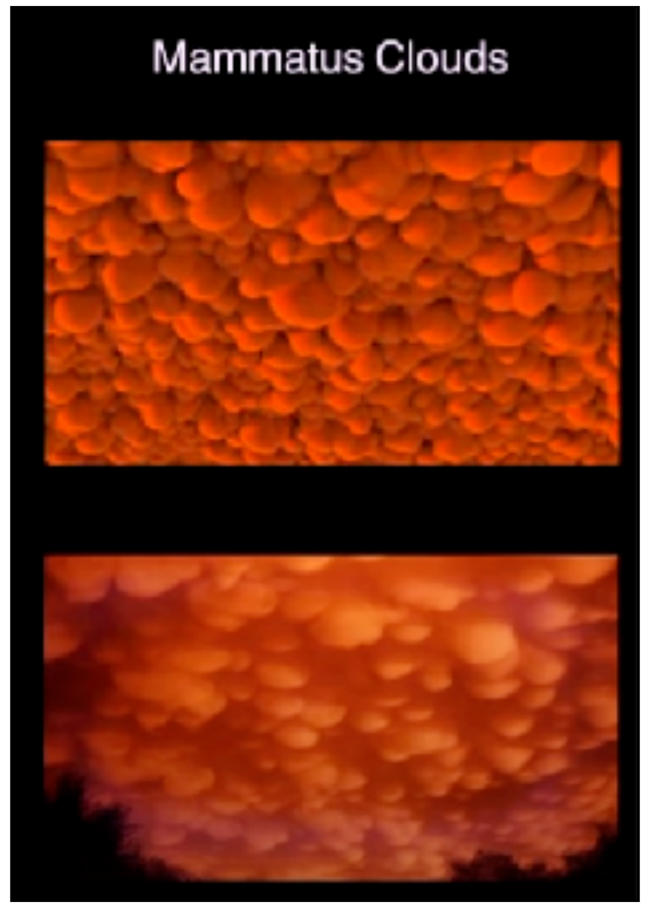
[1] Sora



[2] Glow-TTS



[3] GAIA-1



[4] Weatherscapes

[1] Brook, Tim, et al. "Video generation models as world simulators" OpenAI Blog (2024)
[2] Kim, Jaehyeon, et al. "Glow-tts: A generative flow for text-to-speech via monotonic alignment search." Advances in Neural Information Processing Systems 33 (2020): 8067-8077..
[3] Hu, Anthony, et al. "Gaia-1: A generative world model for autonomous driving." arXiv preprint arXiv:2309.17080 (2023).
[4]J. A. Amador Herrera, et al. "Weatherscapes: Nowcasting Heat Transfer and Water Continuity." ACM Transactions on Graphics (SIGGRAPH Asia 2021), Vol. 40, No. 6, Article 204..

Generative modeling in deep learning

- Generative model: $P(X)$
- Examples:
    - Image/video generation
    - Speech synthesis
    - Physics simulation / world modeling
    - Weather simulation (gaming)
    - ...

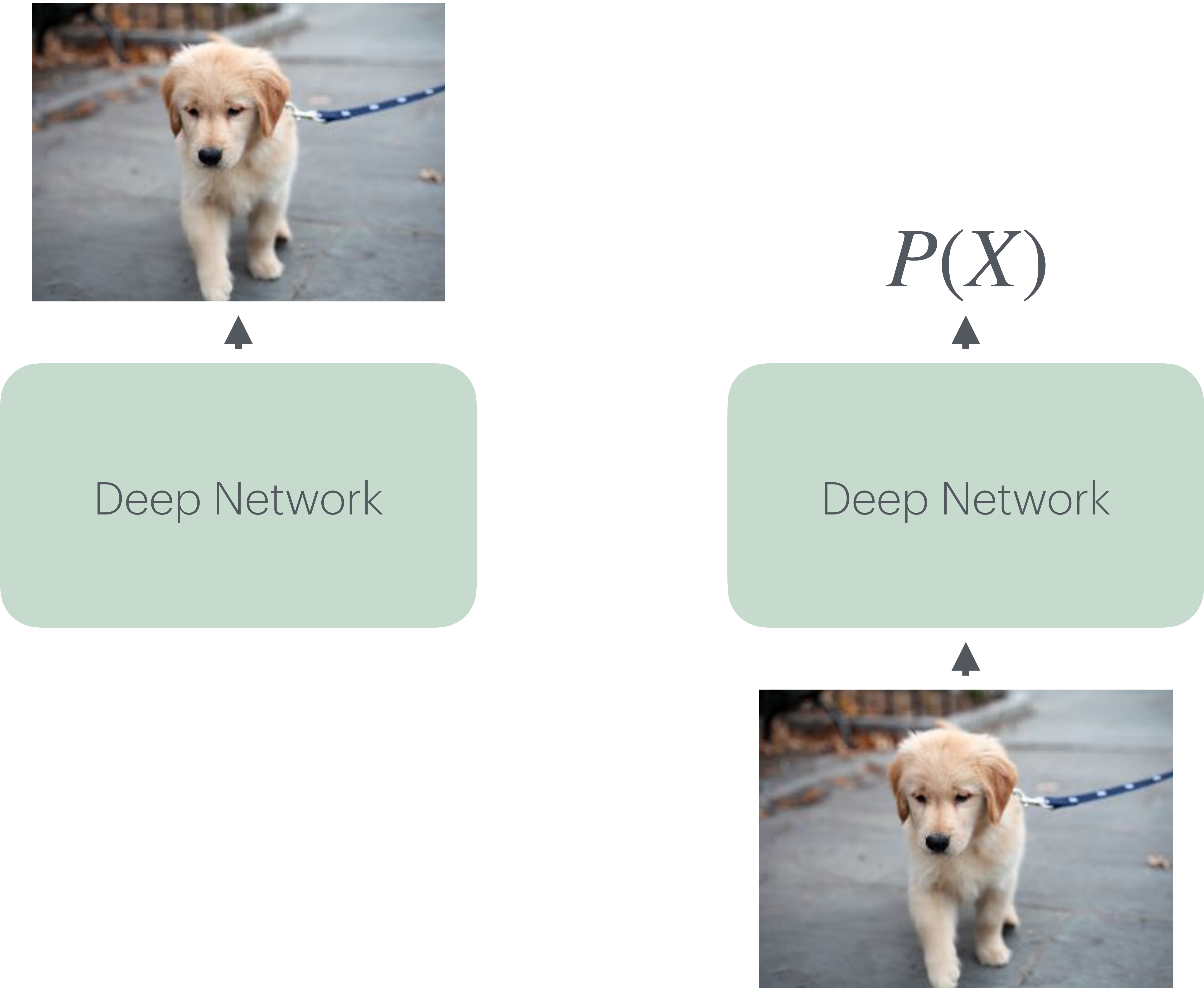
Deep Network

?

Training

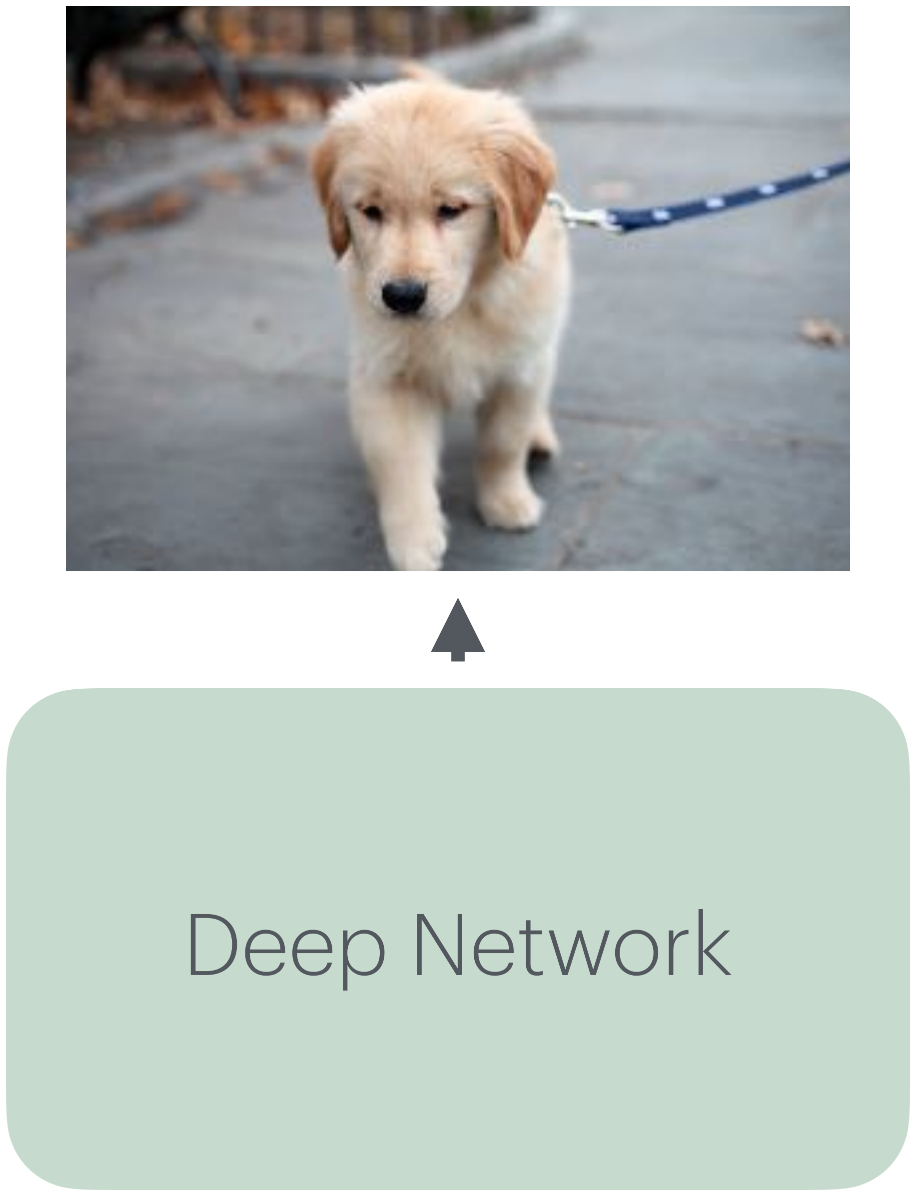
Generative models

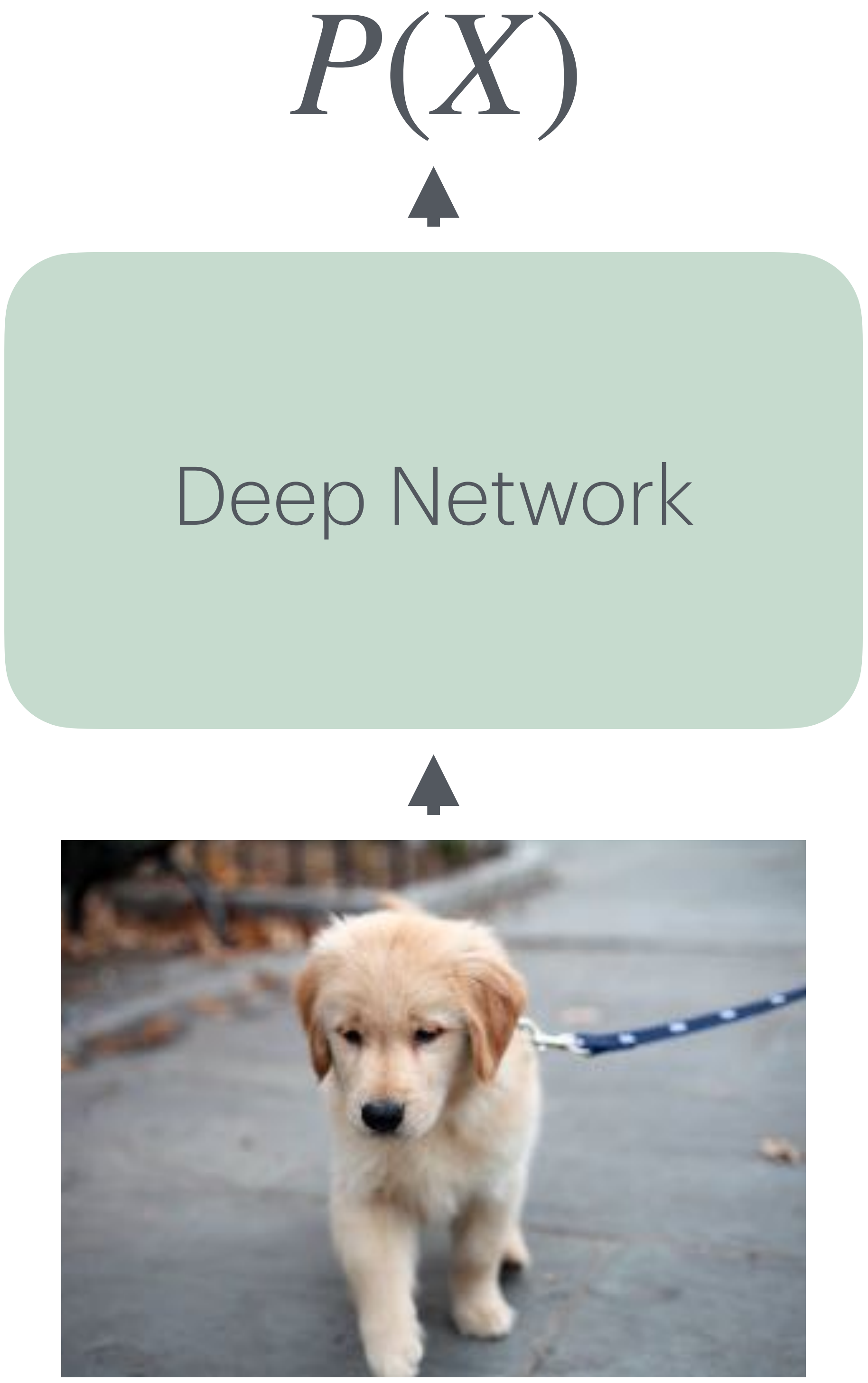
- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$



Generative modeling is hard

- Density estimation $P(X = x)$
  - How to ensure $\sum_x P(x) = 1$ for all x
  - Impossible to compute (in general)
- Sampling $x \sim P(X)$
  - What is the input to the network?





Generative vs Discriminative models

Generative

- Density estimation $P(X = x)$
  - How to ensure $\sum_x P(x) = 1$ for all x
  - Impossible to compute
- Sampling $x \sim P(X)$
  - What is the input to the network?

Discriminative

- Prediction $P(Y | X)$
  - Simple, explicit distribution
    - Discrete $P(y | x) = c_y^\top f(x)$
    - Continuous $P(y | x) = \mathcal{N}(y; \mu(x), \sigma(x))$
  - Well defined input y

Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x \mid z)$ or $f : z \rightarrow x$

z → Deep Network →

Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

→ Deep Network → $P(X)$

References

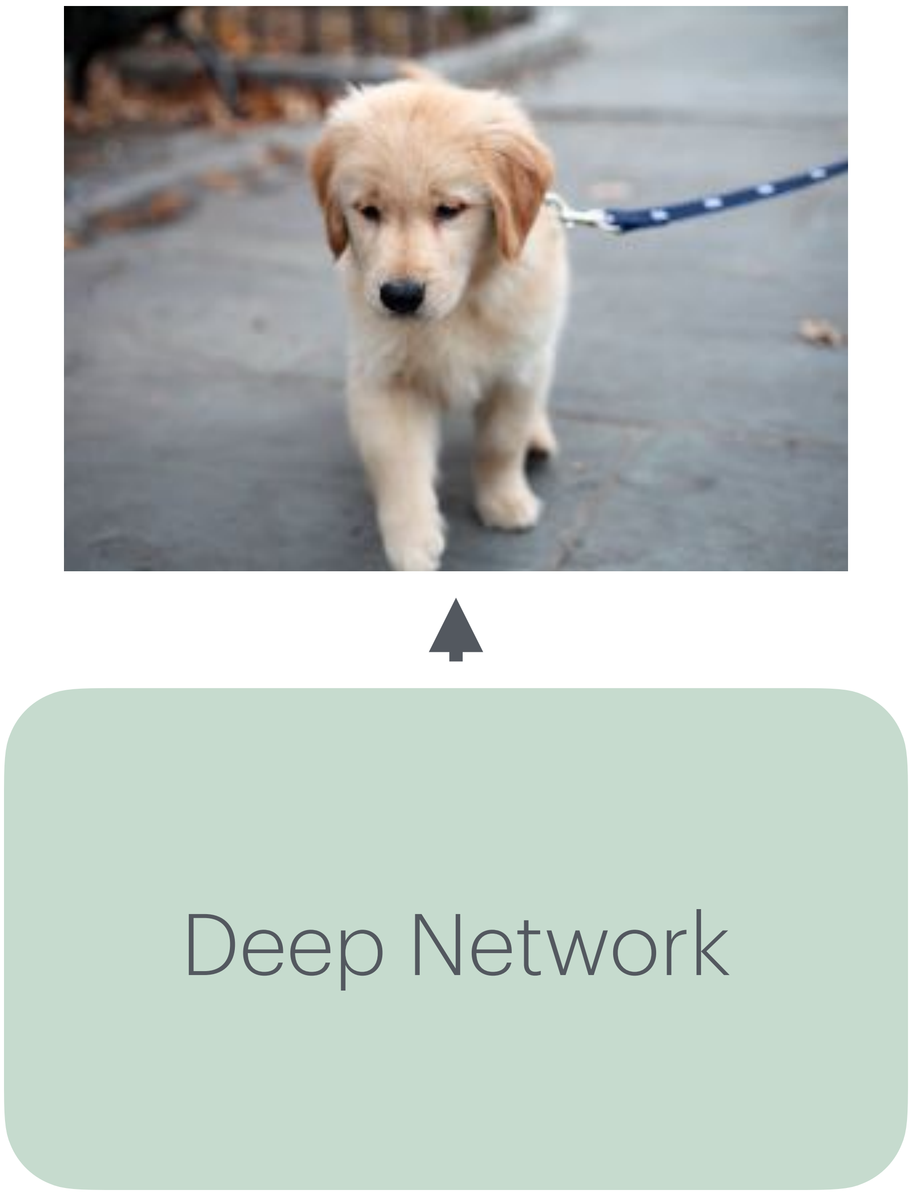
- [1] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.
- [2] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.
- [3] Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." arXiv preprint arXiv:2307.15818 (2023)
- [4] Remi Lam et al. "Learning skillful medium-range global weather forecasting." Science 382, 1416-1421(2023).
- [5] Brook, Tim, et al. "Video generation models as world simulators" OpenAI Blog (2024)
- [6] Kim, Jaehyeon, et al. "Glow-tts: A generative flow for text-to-speech via monotonic alignment search." Advances in Neural Information Processing Systems 33 (2020): 8067-8077.
- [7] Hu, Anthony, et al. "Gaia-1: A generative world model for autonomous driving." arXiv preprint arXiv:2309.17080 (2023).
- [8] J. A. Amador Herrera, et al. "Weatherscapes: Nowcasting Heat Transfer and Water Continuity." ACM Transactions on Graphics (SIGGRAPH Asia 2021), Vol. 40, No. 6, Article 204..

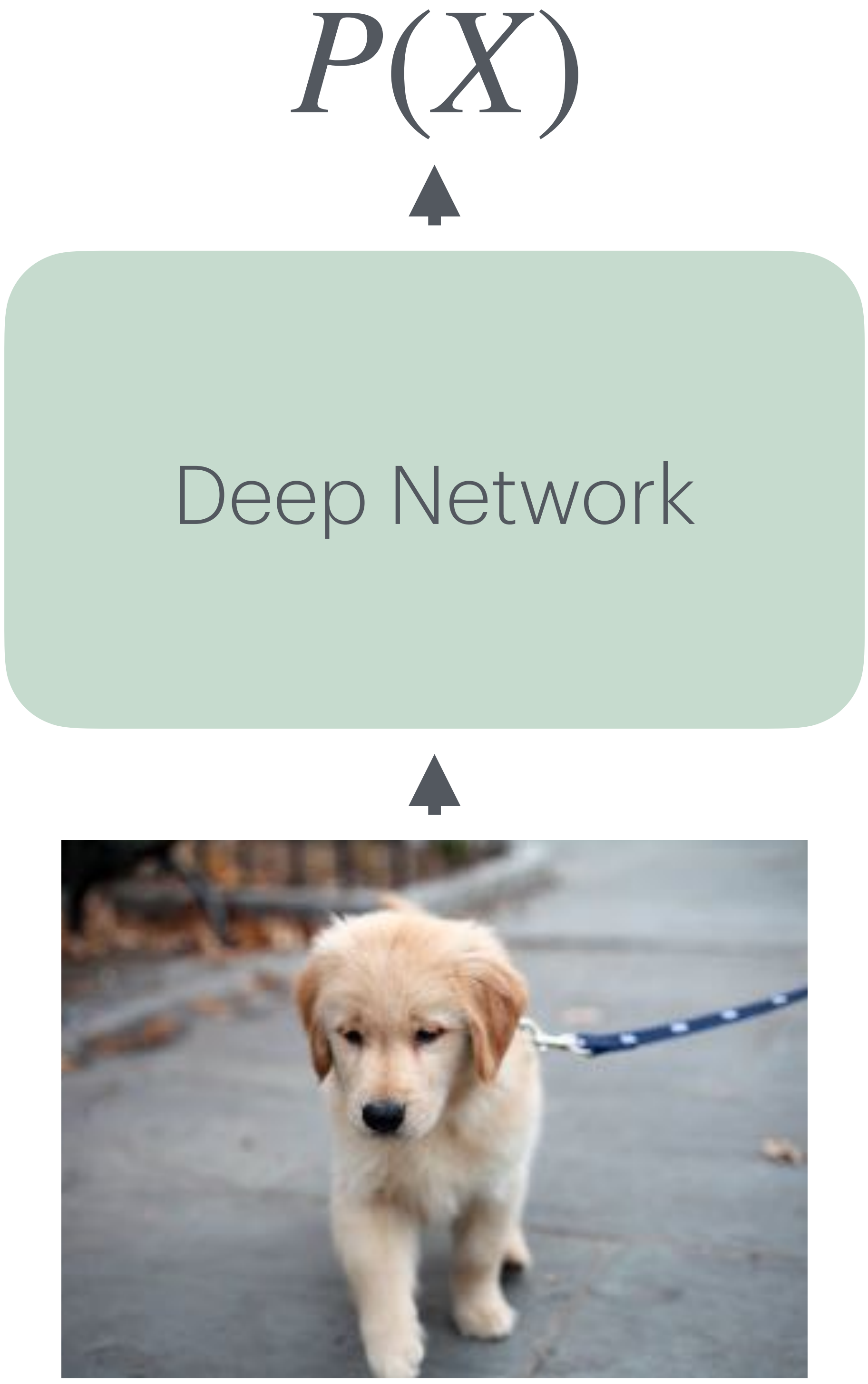
Variational Auto Encoders

Philipp Krähenbühl, UT Austin

Generative models

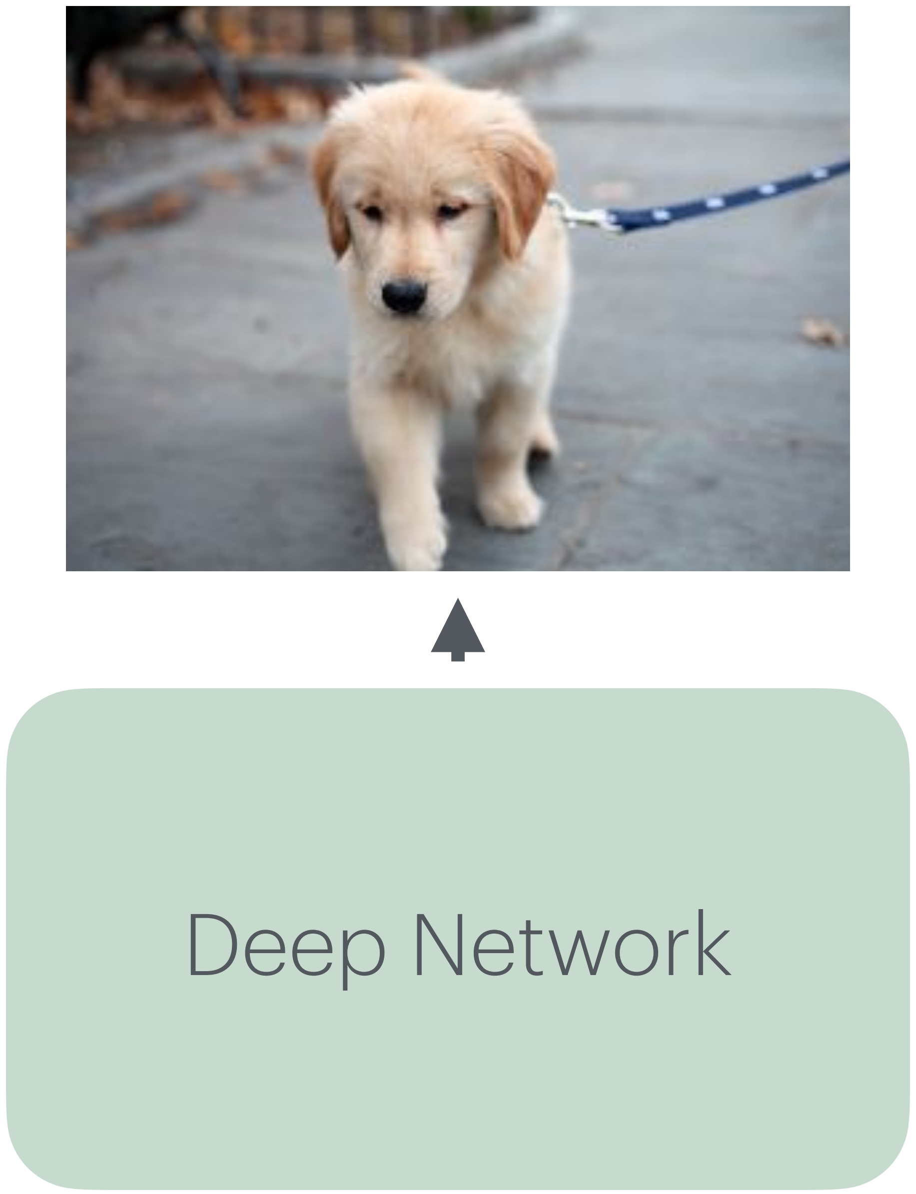
- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$

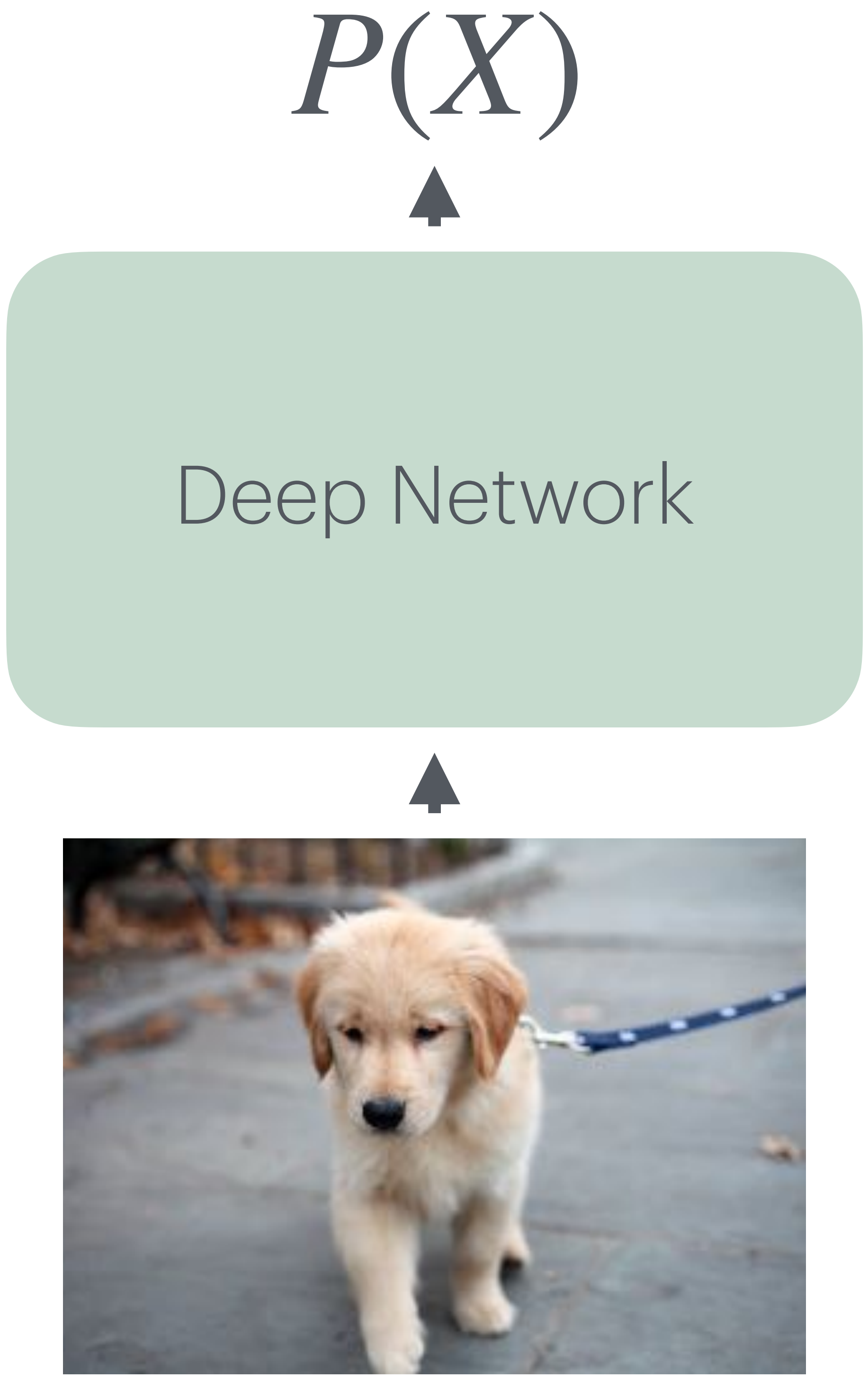




Generative modeling is hard

- Density estimation $P(X = x)$
  - How to ensure $\sum_x P(x) = 1$ for all x
  - Impossible to compute (in general)
- Sampling $x \sim P(X)$
  - What is the input to the network?





Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x|z)$ or $f : z \to x$

z

Deep Network

Density estimation based $P(X)$

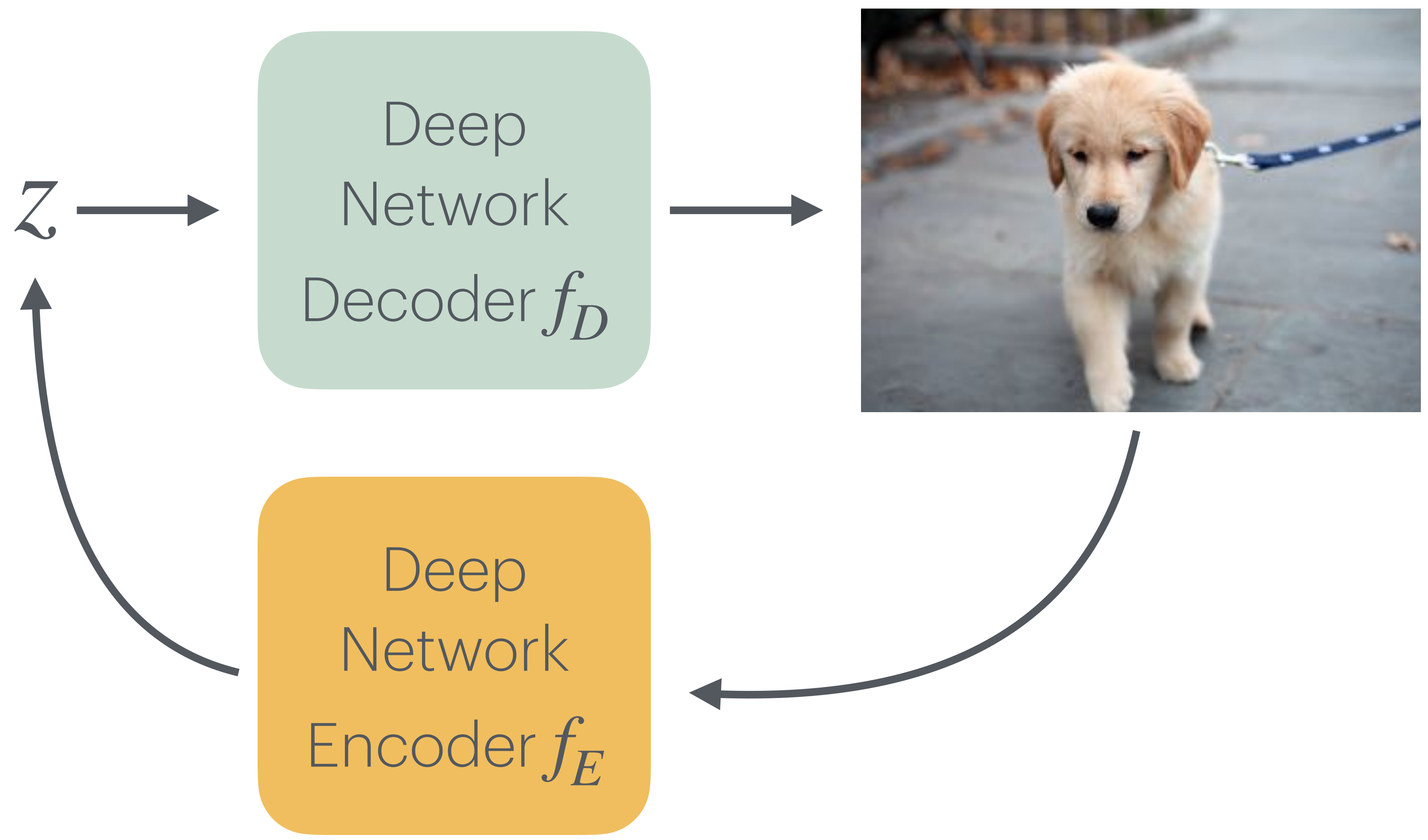
- Learn special form of $P(X)$
- Model specific sampling / generation

Deep Network

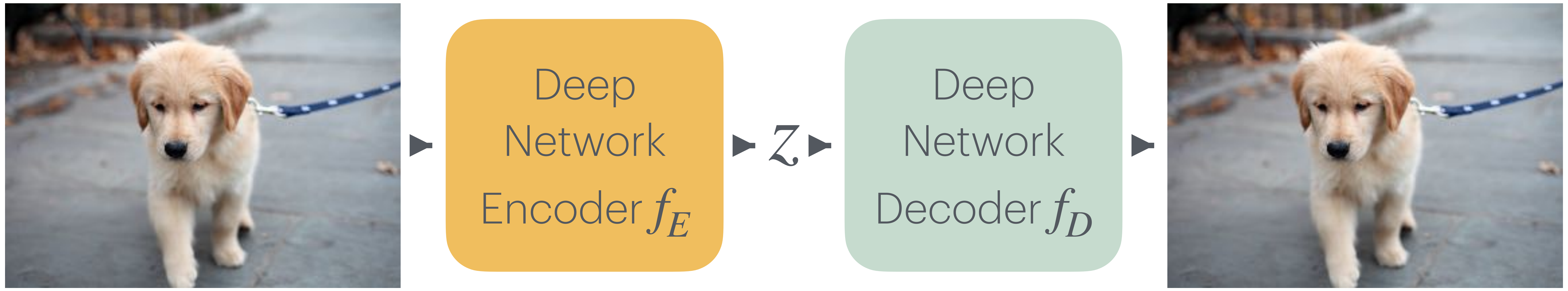
$P(X)$

Generative models

- Goal: Learn decoder $f_D : z \to x$
- What should z be?
  - Let a deep network decide
    - Encoder $f_E : x \to z$



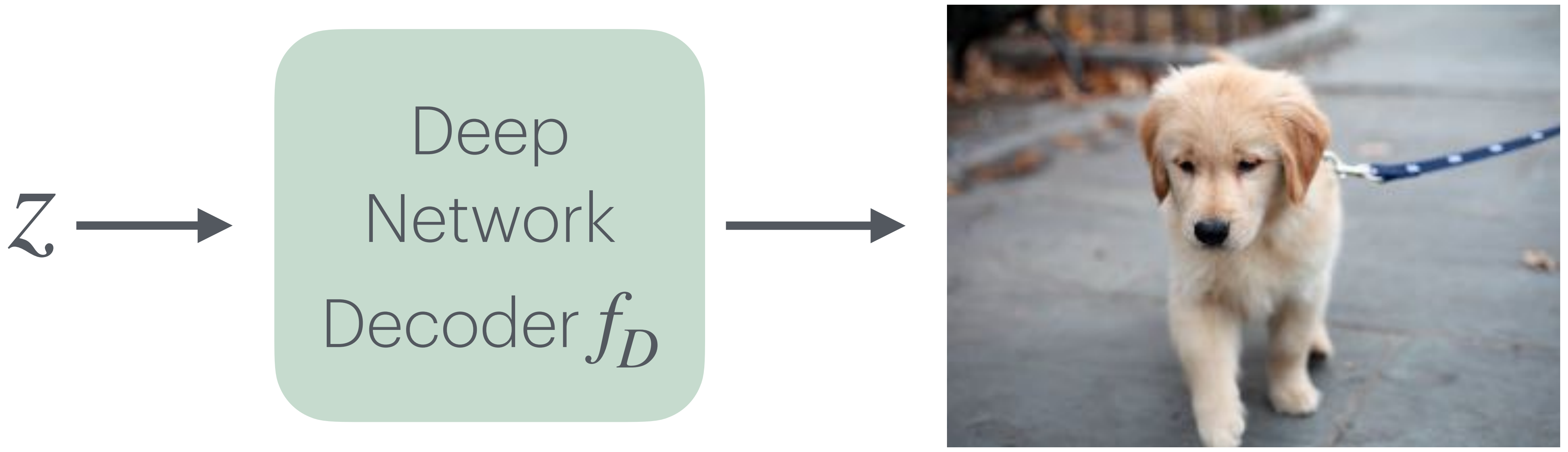
Auto-encoder



- encoder $z = f_E(x)$
- decoder $\hat{x} = f_D(z)$
- Training
  - Supervised learning on large dataset
  - $\ell = E_x \left[| f_D(f_E(x)) - x | \right]$

Auto-encoder

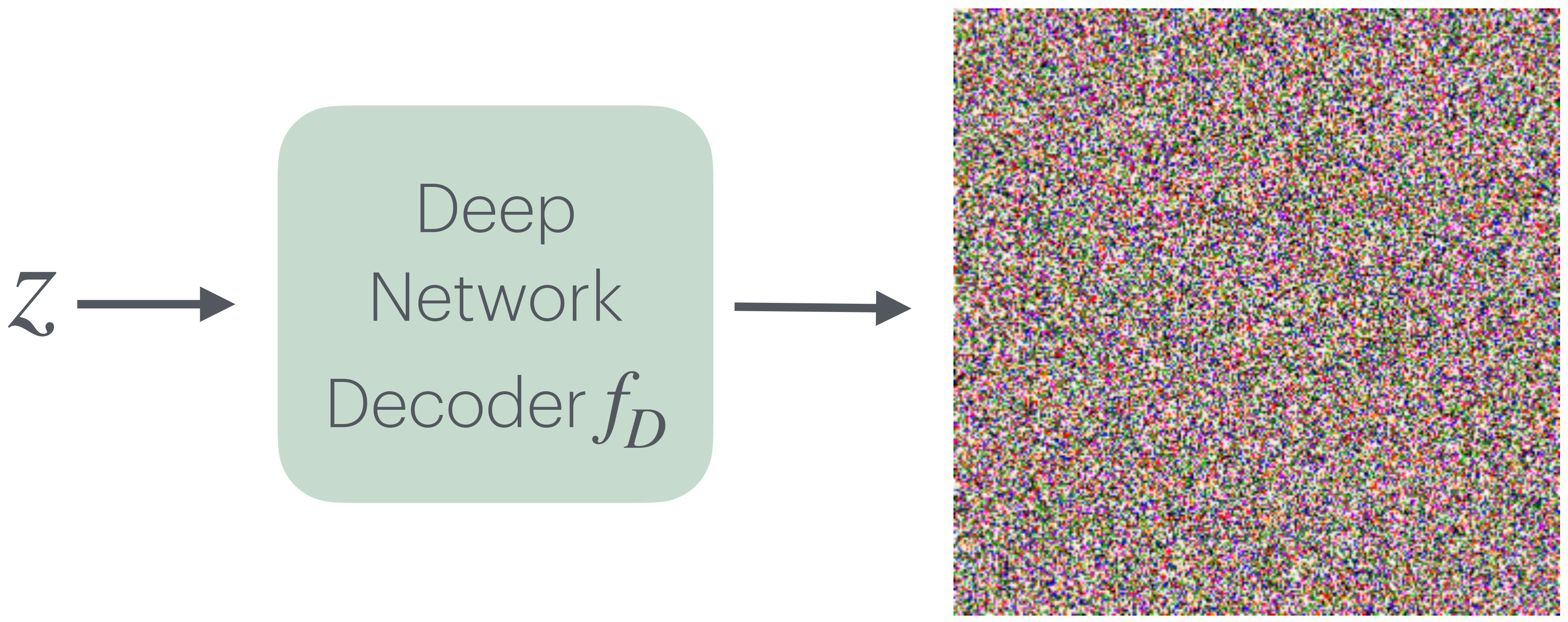
as a Generative model



- Decoder $f_D : z \rightarrow x$
- Inference / Sampling
  - What is z at test time?

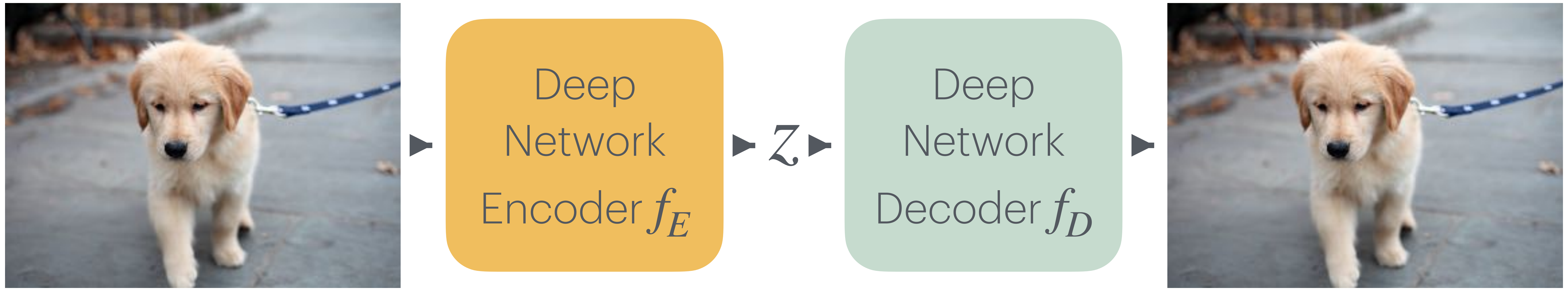
Auto-encoder

Generation



- Decoder $f_D : z \to x$
- Inference / Sampling
  - What is z at test time?
    - Network output -> no new image
    - Random input -> Garbage
    - Interpolation -> Garbage

What does an auto-encoder learn?

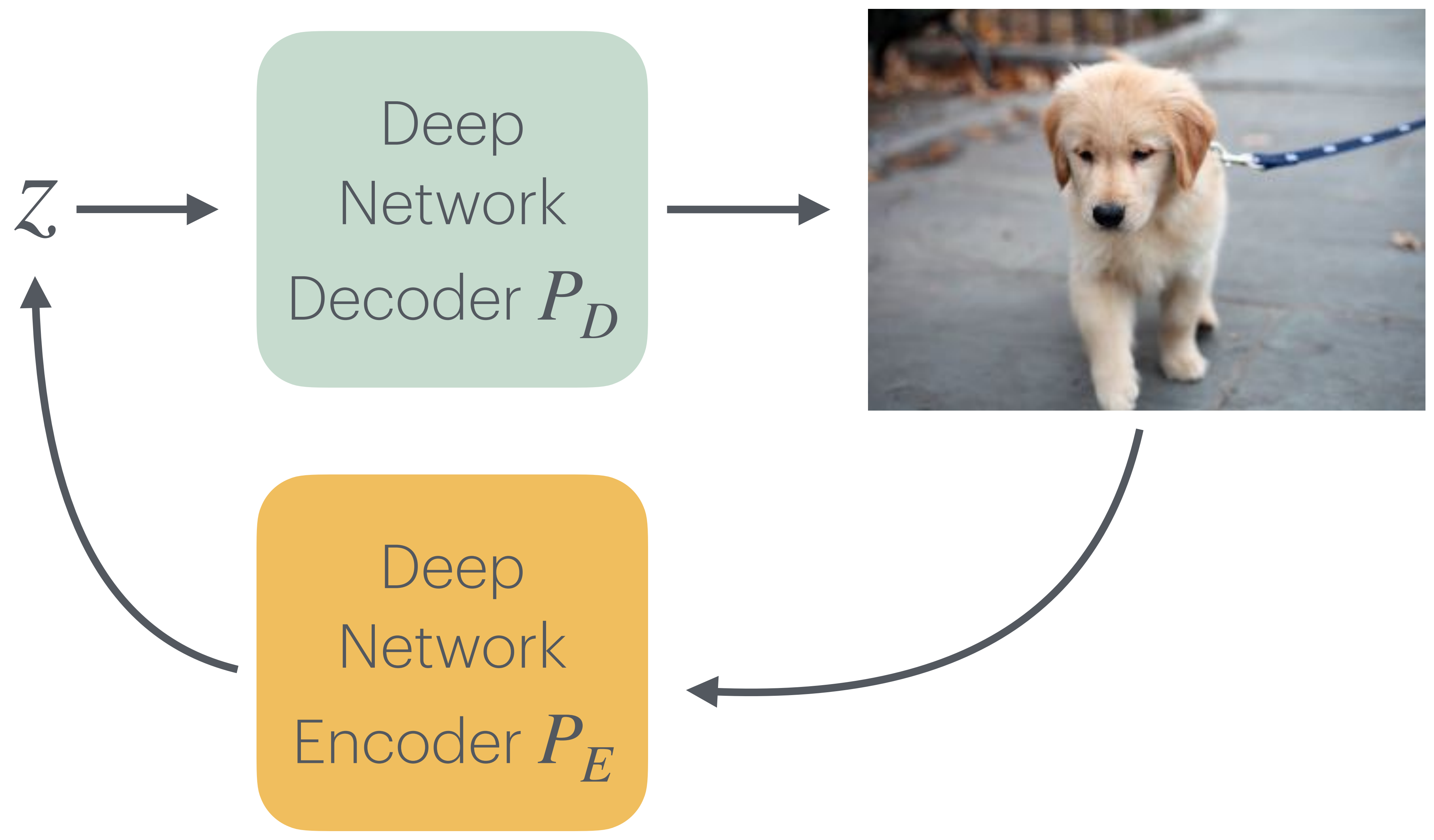


- Compression
- “Invertible” mapping
- Does it learn the structure of images?
    - Only in the limit
    - Perfect compression = understanding
- Poor generation

Variational auto-encoder

A "probabilistic" auto-encoder

- Goal: Learn decoder $P_D(x \mid z)$
- What should z be?
  - Let a deep network decide
    - Encoder $P_E(z \mid x)$

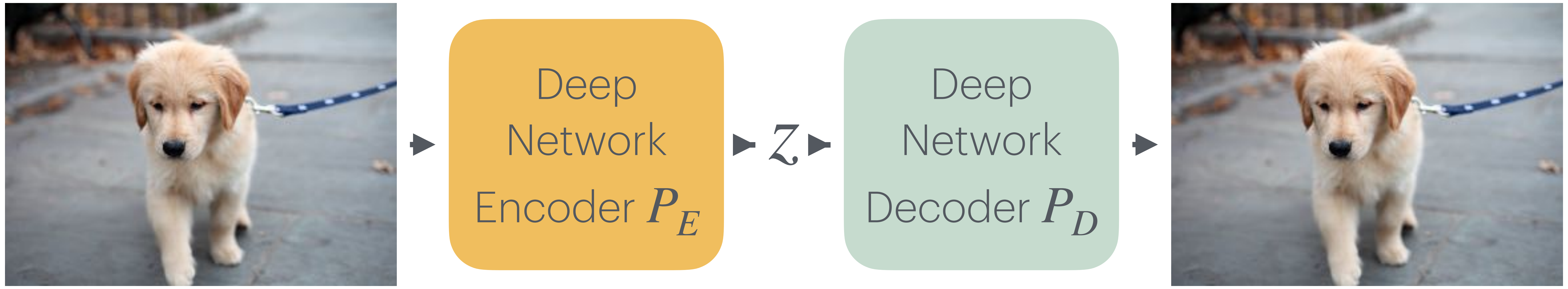


[1] Auto-Encoding Variational Bayes. Kingma et al. 2014.

Variational auto-encoder

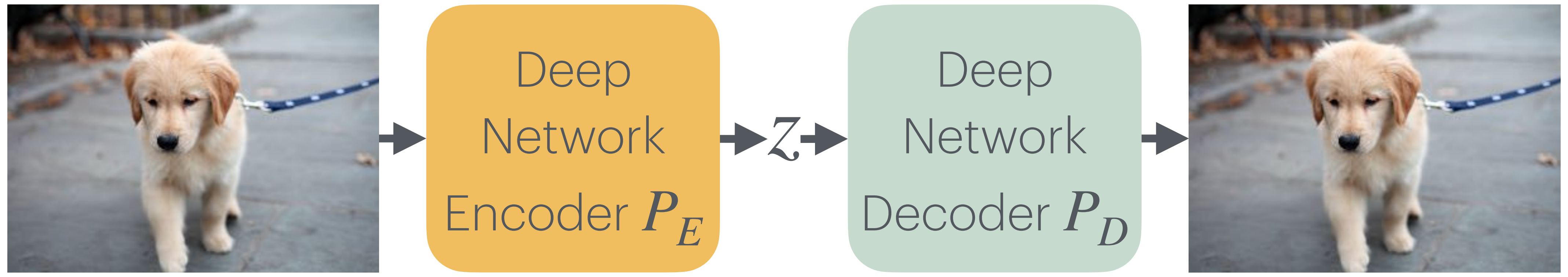
A "probabilistic" auto-encoder

- Decoder $P_D(x \mid z)$ (similar to discriminative model)
- Encoder $P_E(z \mid x)$ (similar to discriminative model)
- Assume $P(Z) = \mathcal{N}(0,1)$
- $P(x) = \sum_z P_D(x \mid z) P(z)$
- $z \sim P(X)$ is equivalent to $z \sim P(Z)$ and $x \sim P(x \mid z)$



Variational auto-encoder

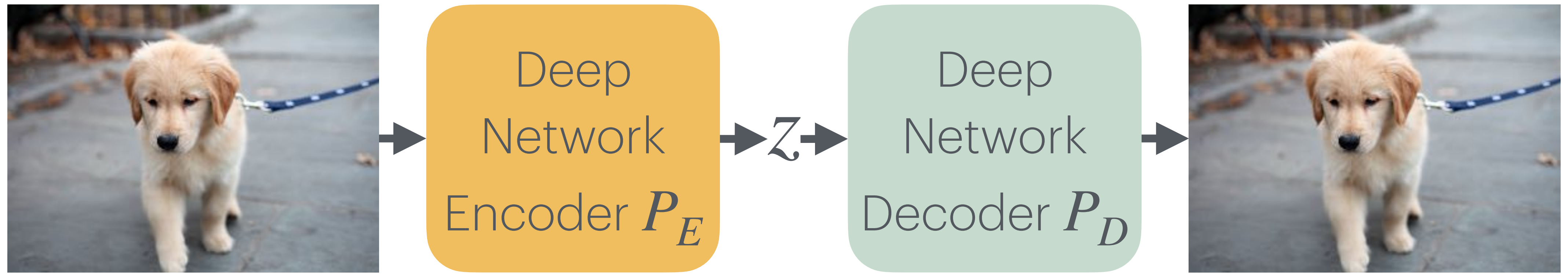
A "probabilistic" auto-encoder



- Decoder $P_D(x \mid z)$ (similar to discriminative model)
- Encoder $P_E(z \mid x)$ (similar to discriminative model)
- Assume $P(Z) = \mathcal{N}(0,1)$
- Bayes rule $P_E(z \mid x) = \dfrac{P_D(x \mid z)P(z)}{P(x)}$ ←intractable

Variational auto-encoder

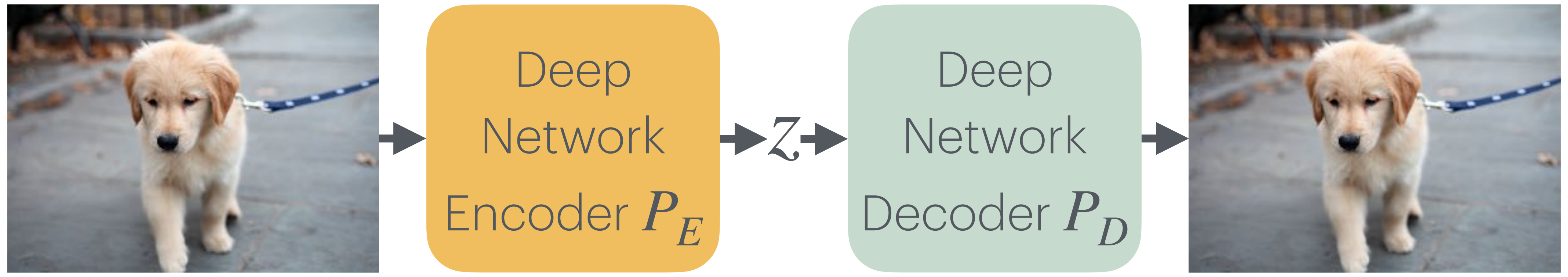
A "probabilistic" auto-encoder



- Decoder $P_D(x \mid z)$ (similar to discriminative model)
- Encoder $Q(z \mid x)$ (similar to discriminative model)
- Assume $P(Z) = \mathcal{N}(0,1)$
  - Bayes rule $P_E(z \mid x) = \dfrac{P_D(x \mid z)P(z)}{P(x)}$ ←intractable
  - Learn $Q \approx P_E$ that minimizes $D_{KL}(Q \mid P_E)$

Variational auto-encoder

A "probabilistic" auto-encoder



- Learn $Q \approx P_E$ that minimizes

$$D_{KL}(Q(z\,|\,x) \| P_E(z\,|\,x)) = \log P(x) + E_{z\sim Q}\left[\log \frac{P(z)P_D(x\,|\,z)}{Q(z\,|\,x)}\right]$$

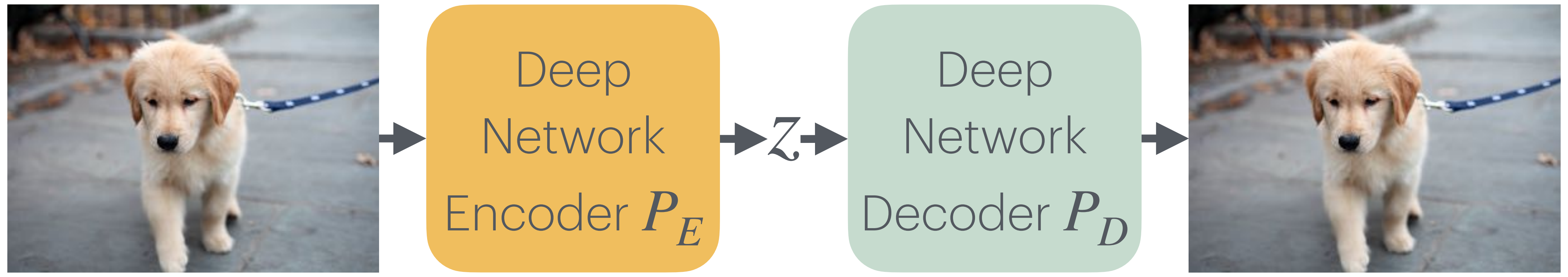
- Maximize $\log P(x)$ of real data, minimize D_{KL}

$$\log P(x) - D_{KL}(Q(z\,|\,x) \| P_E(z\,|\,x)) = E_{z\sim Q}\left[\log \frac{Q(z\,|\,x)}{P(z)P_D(x\,|\,z)}\right]$$

  - Known as ELBO (Evidence Lower BOund)

Variational auto-encoder

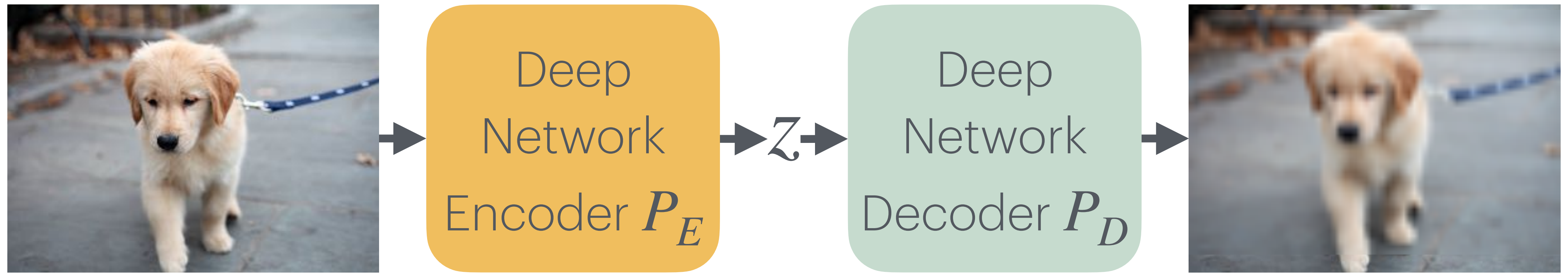
A "probabilistic" auto-encoder



- ELBO $E_{z\sim Q}\left[\log \frac{Q(z\,|\,x)}{P(z)P_D(x\,|\,z)}\right]$ for Gaussians
  - $-\frac{1}{2}\mathbb{E}_{z\sim Q}\left[\|x-\mu_D(z)\|_2^2\right]-\frac{1}{2}\left(N\sigma_Q(x)^2+\|\mu_Q(x)\|_2^2-2N\log\sigma_Q(x)\right)+Const$
- Reparametrization trick
  - For $Q(z\,|\,x)=\mathcal{N}(z;\mu_Q(x),\sigma_Q^2(x))$
  - $\mathbb{E}_{z\sim Q}\left[\|x-\mu_D(z)\|_2^2\right]=\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\left[\|x-\mu_D(\mu_Q(x)+\varepsilon\sigma_Q(x))\|_2^2\right]$

Variational auto-encoder

A "probabilistic" auto-encoder

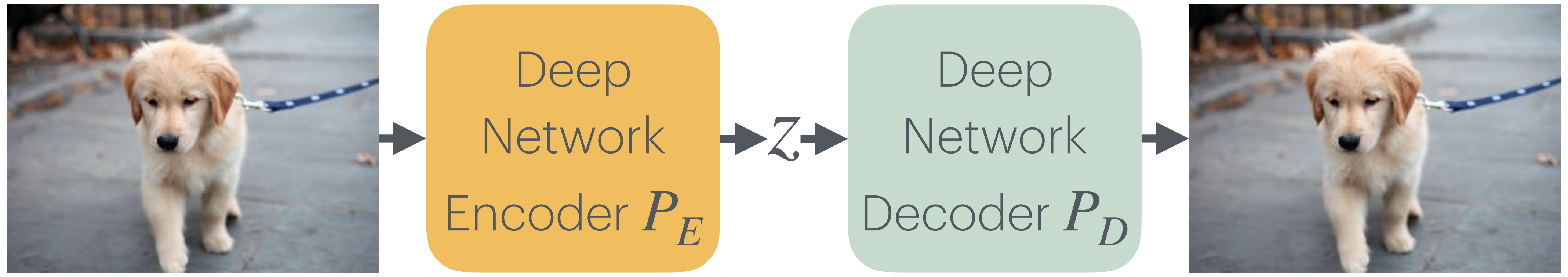


- Can learn $P(X)$ and sampling function $x \sim P$
- Issues
  - Reconstruction loss: Pixel-level l2 loss
    - Blurry outputs
  - Approximation Q: Gaussian assumption
    - Sphere packing in higher dimensions
    - Lots of empty space

[1] Auto-Encoding Variational Bayes. Kingma et al. 2014.

Variational auto-encoder

A "probabilistic" auto-encoder



- Learn a model of $P(x) = P_D(x\,|\,z)P(z)$ with $P(z) = \mathcal{N}(z; 0,1)$
  - Training: Maximize $P(x)$ of data
  - Approximate $Q \approx P_E$

[1] Auto-Encoding Variational Bayes. Kingma et al. 2014.

References

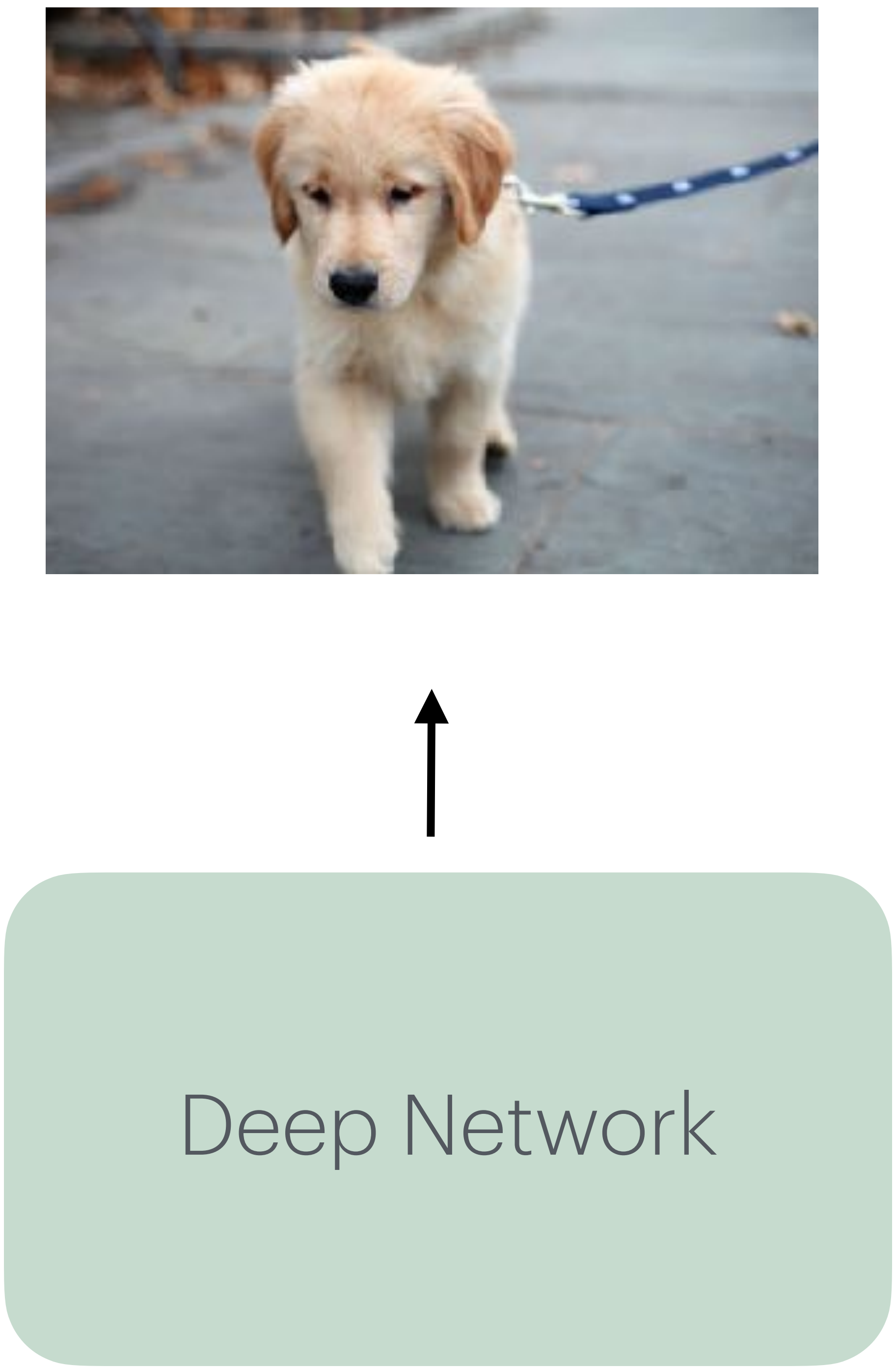
- [1] Auto-Encoding Variational Bayes. Kingma et al. 2014.

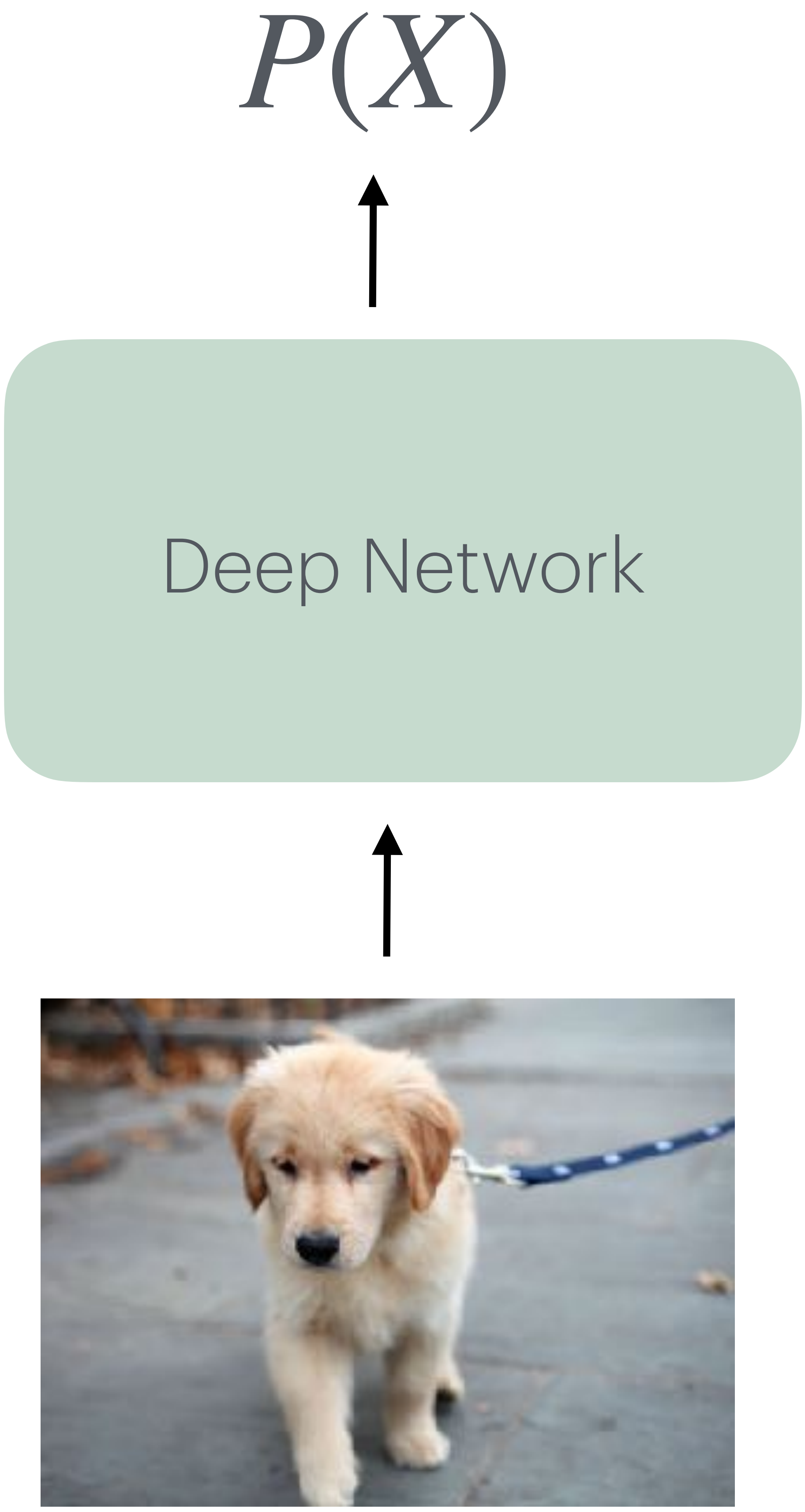
Auto-regressive generation

Philipp Krähenbühl, UT Austin

Generative models

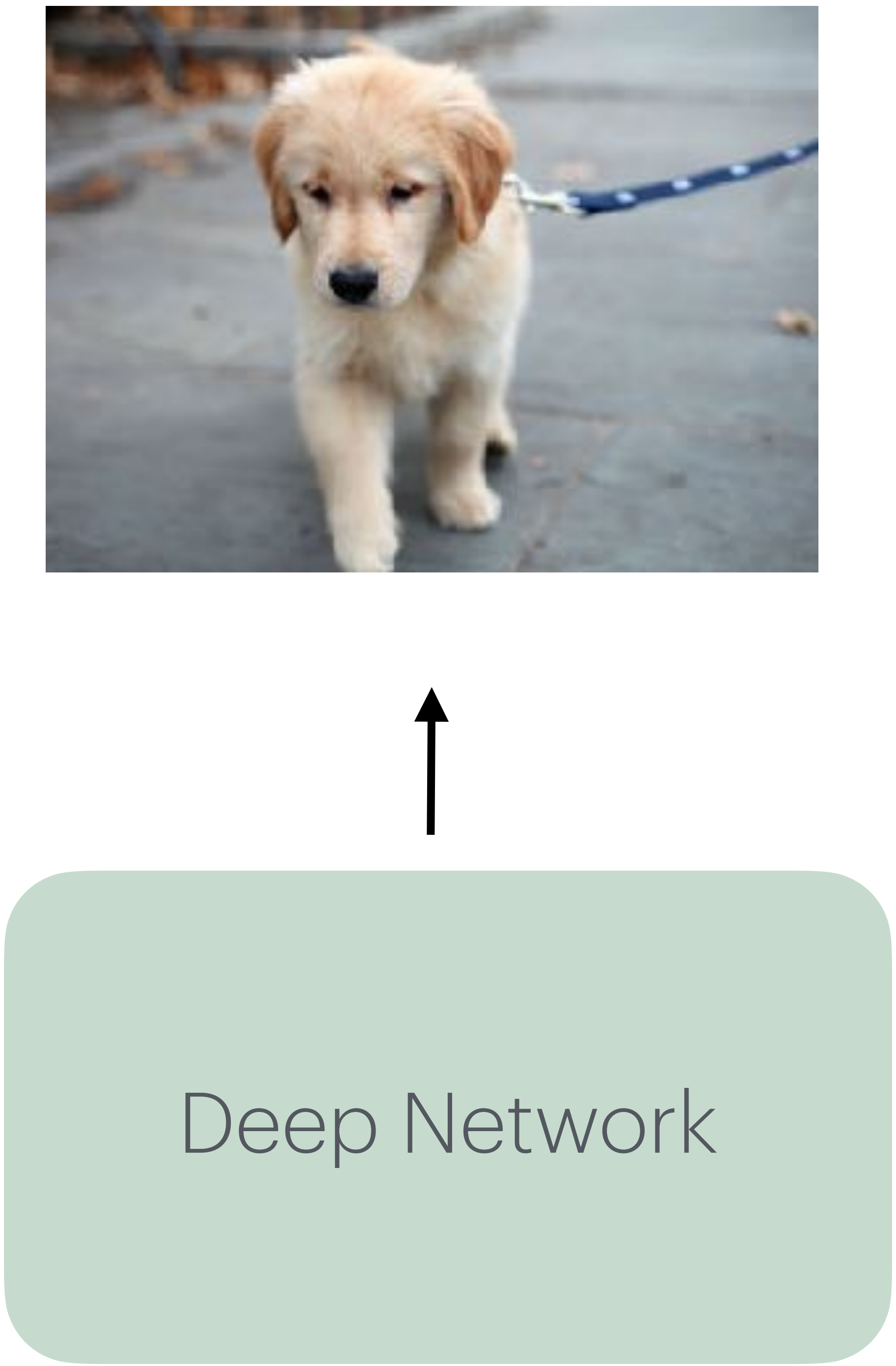
- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$

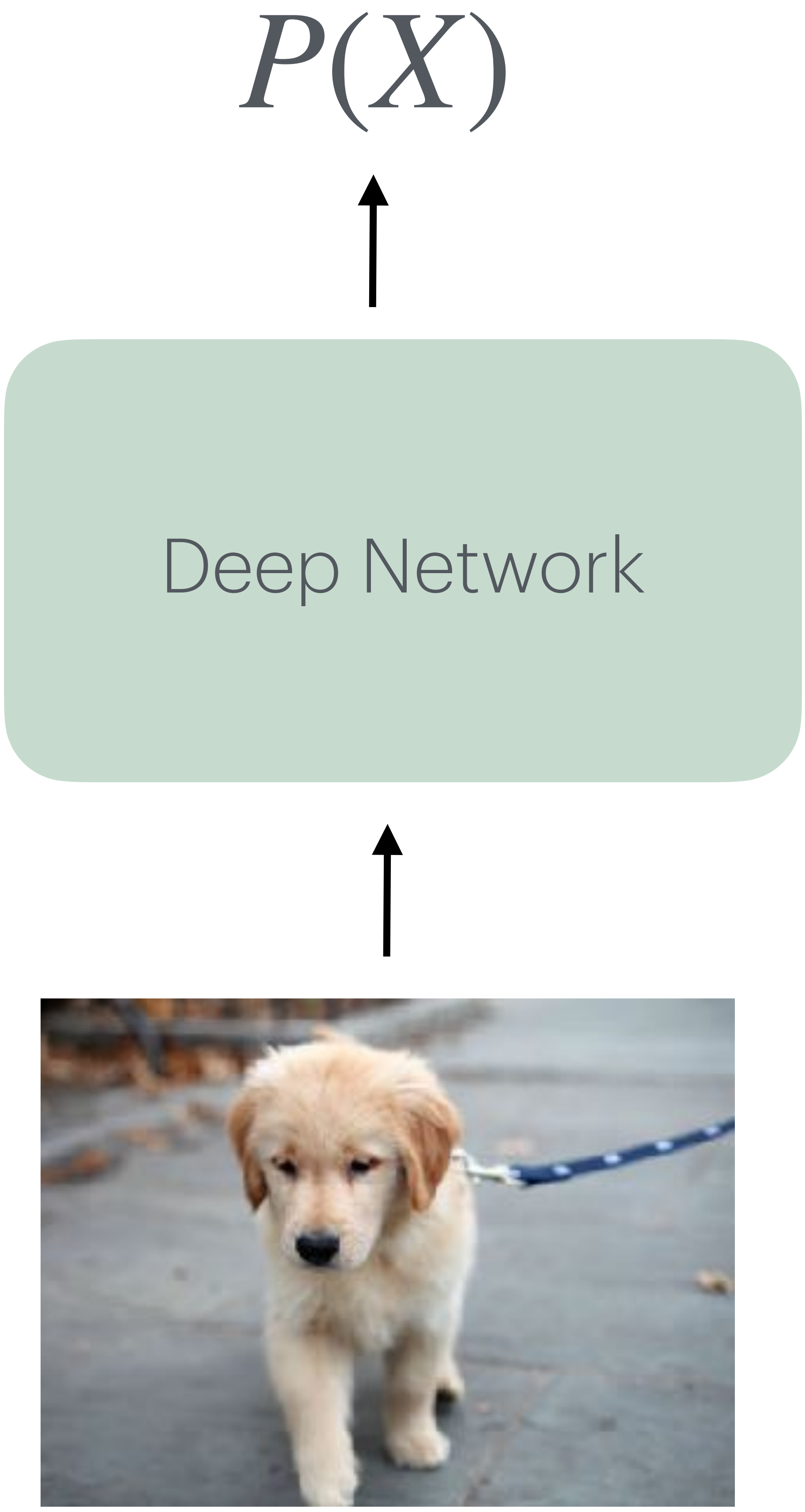




Generative modeling is hard

- Density estimation $P(X = x)$
  - How to ensure $\sum_x P(x) = 1$ for all x
  - Impossible to compute (in general)
- Sampling $x \sim P(X)$
  - What is the input to the network?



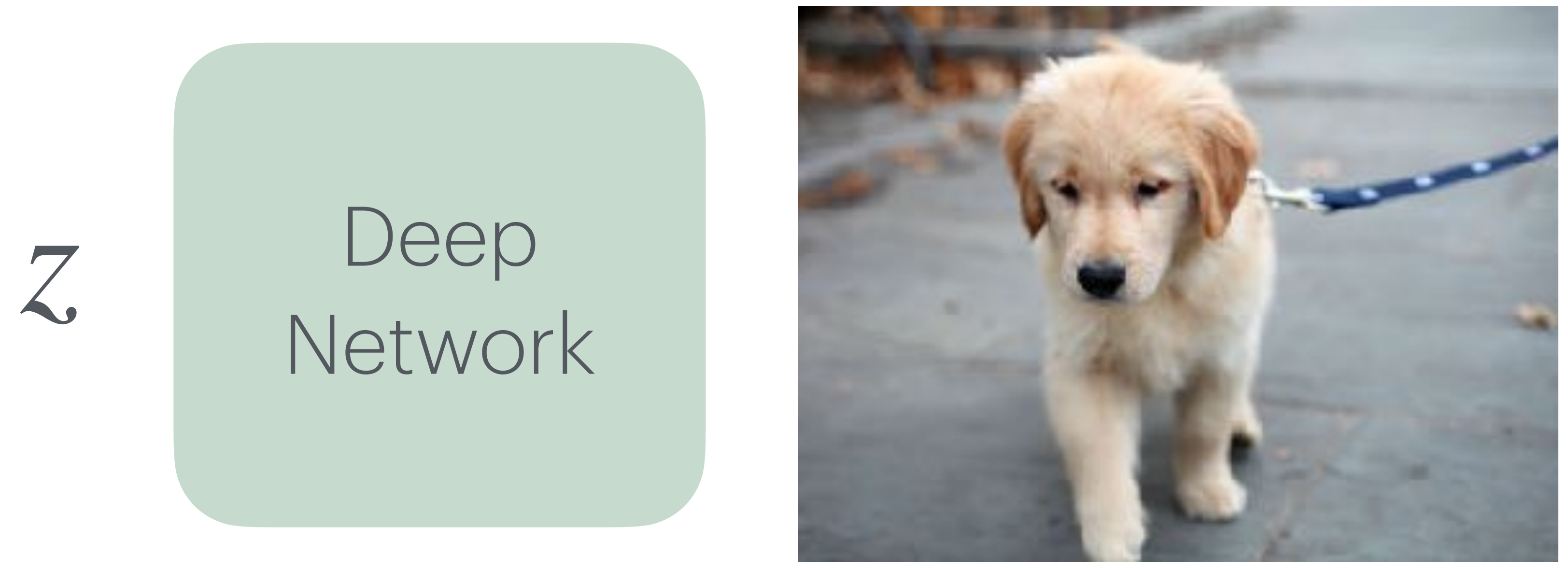


Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x \mid z)$ or $f : z \rightarrow x$

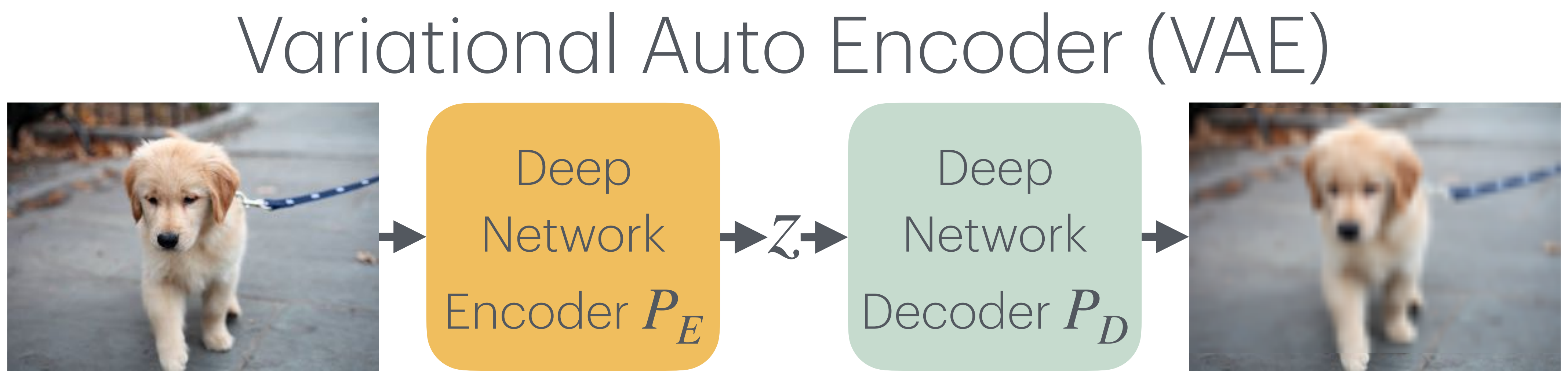


Density estimation based $P(X)$

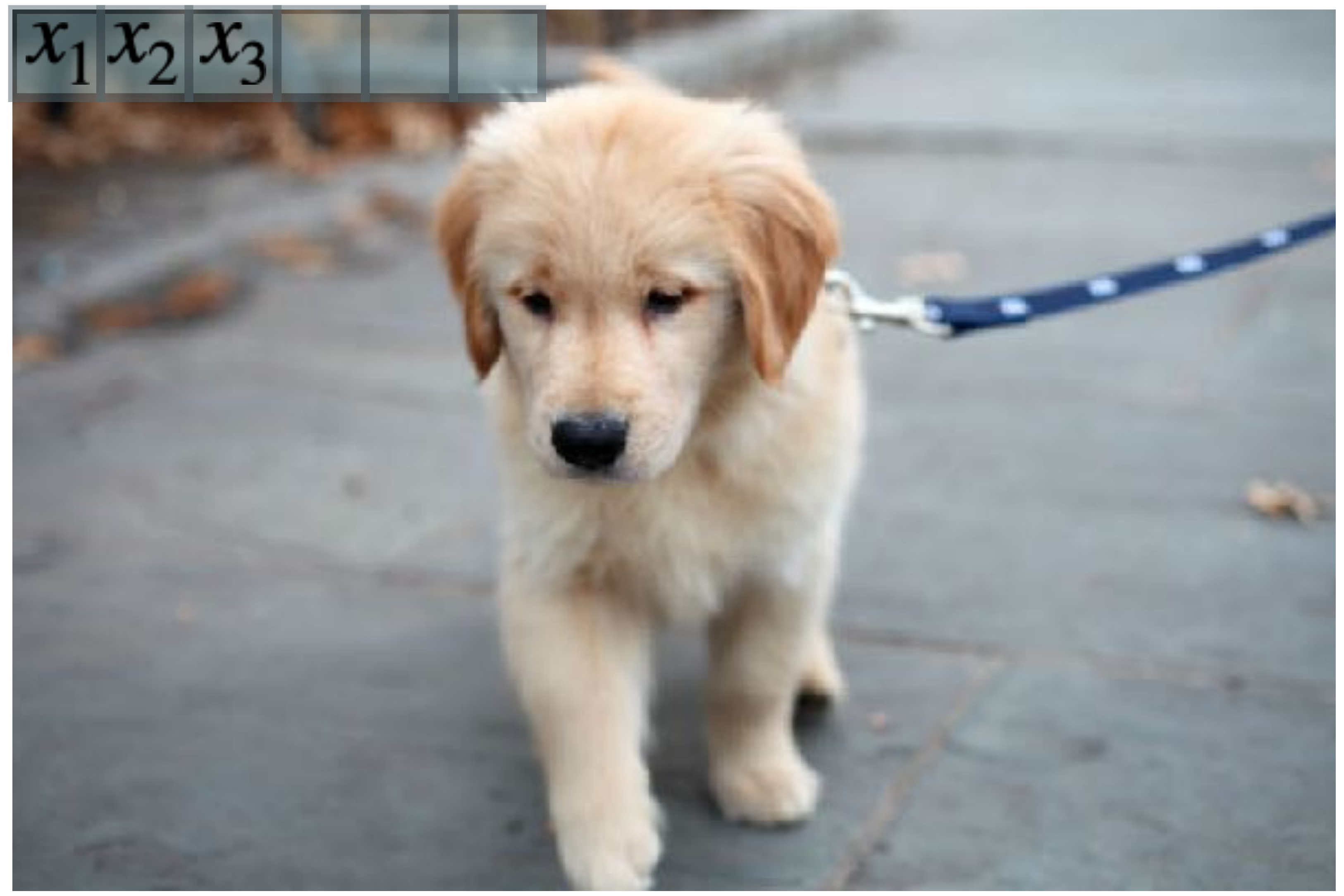
- Learn special form of $P(X)$
- Model specific sampling / generation

Recap

- VAE
  - Image -> latent space -> Image
  - Loss encourages Gaussian latent



Auto-regressive models



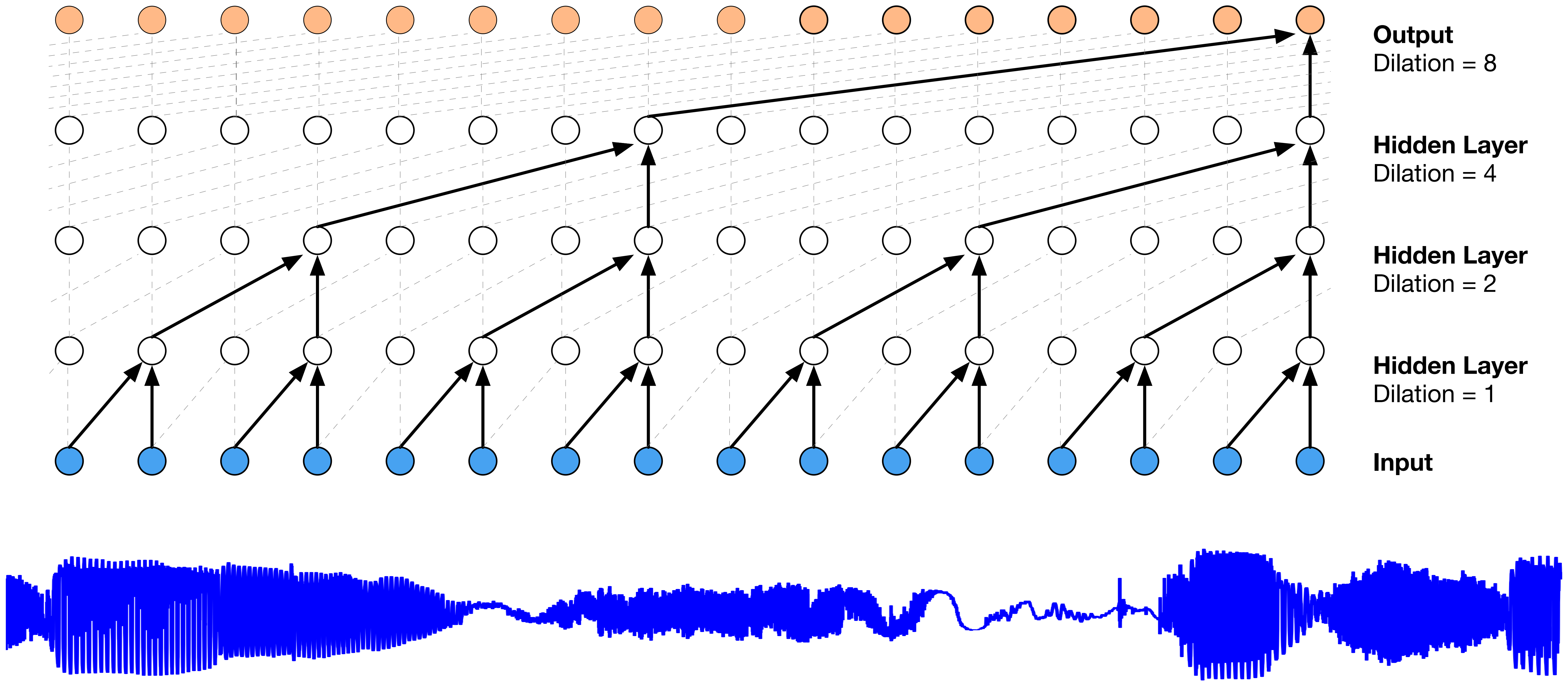
$$P(x) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)P(x_4|x_1\ldots x_3)\ldots$$

- $P(x_i|x_1\ldots x_{i-1}) = \mathrm{softmax}(f(x_1\ldots x_{i-1}))$
- Basis of most LLM models
- Easy estimation of $P(x)$
- Easy sampling
  $x_1 \sim P(X_1); x_2 \sim P(X_2|x_1)$
  - Slow sampling

[1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016
[2] Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. Songwei Ge, et al. 2022

Example: WaveNet

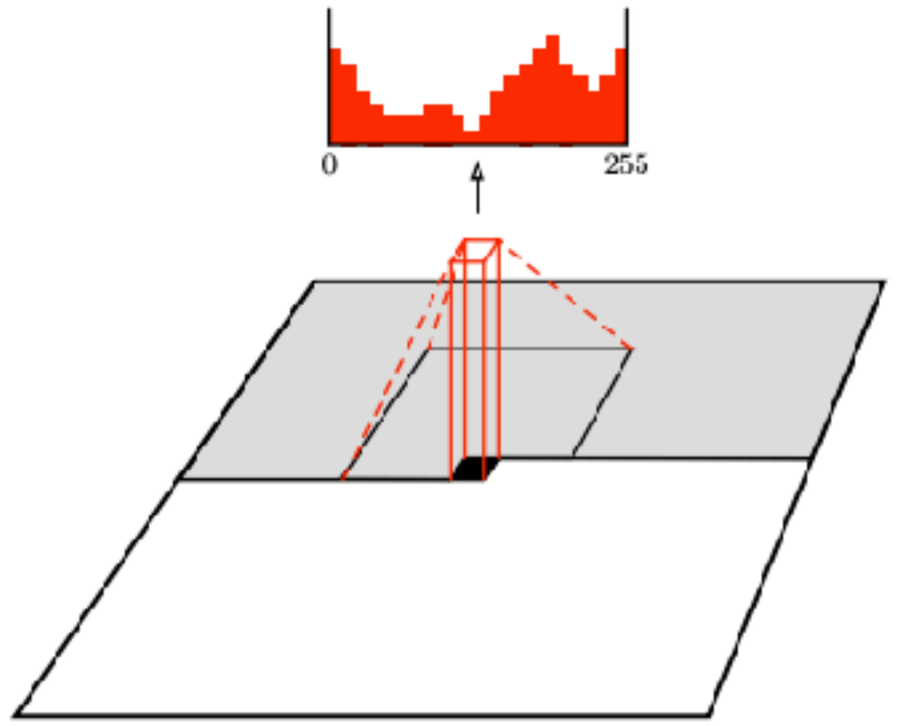
- Input: Raw waveform $\mathbf{x}_{1\ldots t-1}$
- Output: Quantized next value
  $\mathbf{x}_t \in \{1\ldots 256\}$
- Model: $P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \,|\, \mathbf{x}_{1\ldots t-1})$
- Conditioned model:
  $P(\mathbf{x} \,|\, \mathbf{h}) = \prod_{t=1}^{T} P(x_t \,|\, \mathbf{x}_{1\ldots t-1} \,|\, \mathbf{h})$

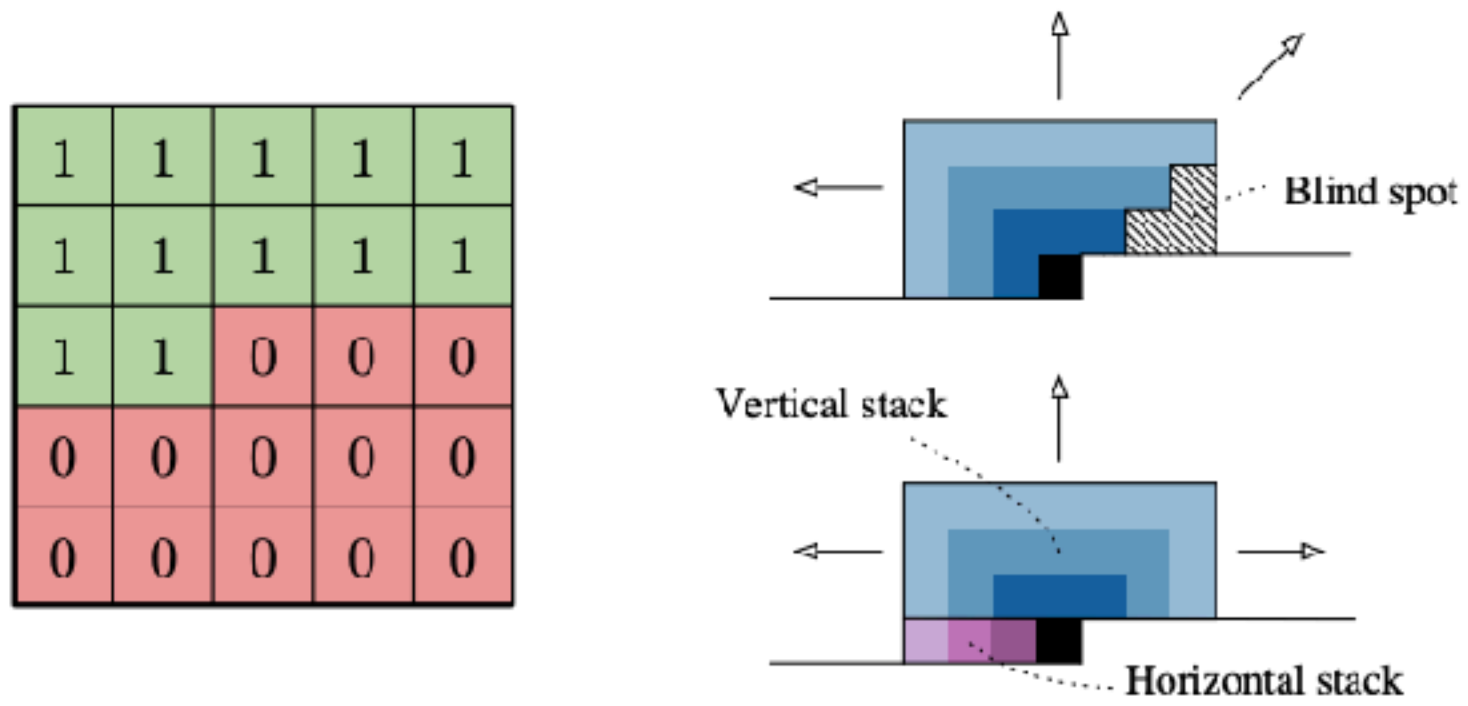


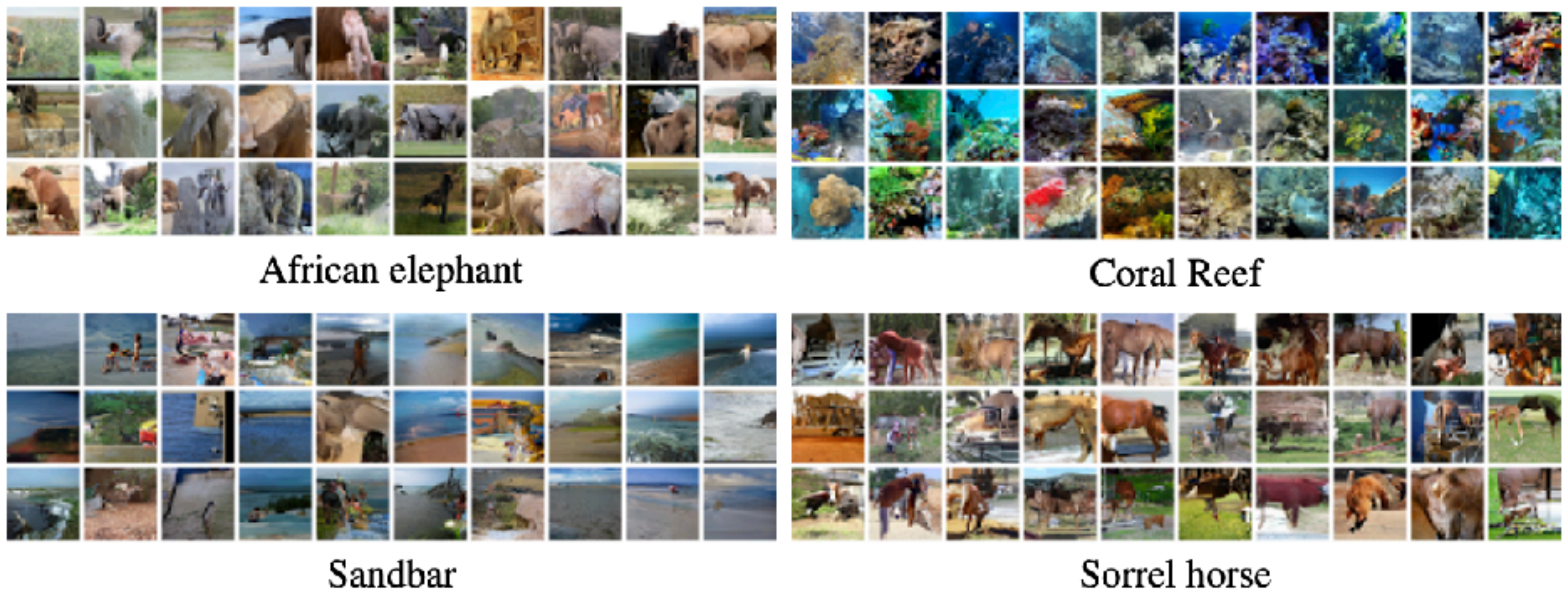
[1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016

Example: PixelCNN

- Input: Raw pixels $\mathbf{x}_{1\ldots t-1}$
- Output: Quantized next color value $\mathbf{x}_t \in \{1 \ldots 256\}$
- Model: $P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \mid \mathbf{x}_{1\ldots t-1})$
- Conditioned model: $P(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} P(x_t \mid \mathbf{x}_{1\ldots t-1} \mid \mathbf{h})$







Auto-regressive models

Issues

$$P(x) = P(x_1)P(x_2 | x_1)P(x_3 | x_1, x_2)P(x_4 | x_1 \ldots x_3)\ldots$$

- Difficult learning problem for long sequences (requires good model)
- Solution: Tokenization/Vector-Quantization (next class)
    - More complex x_i
    - Shorter sequence

[1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016
[2] Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. Songwei Ge, et al. 2022

Generation vs Compression

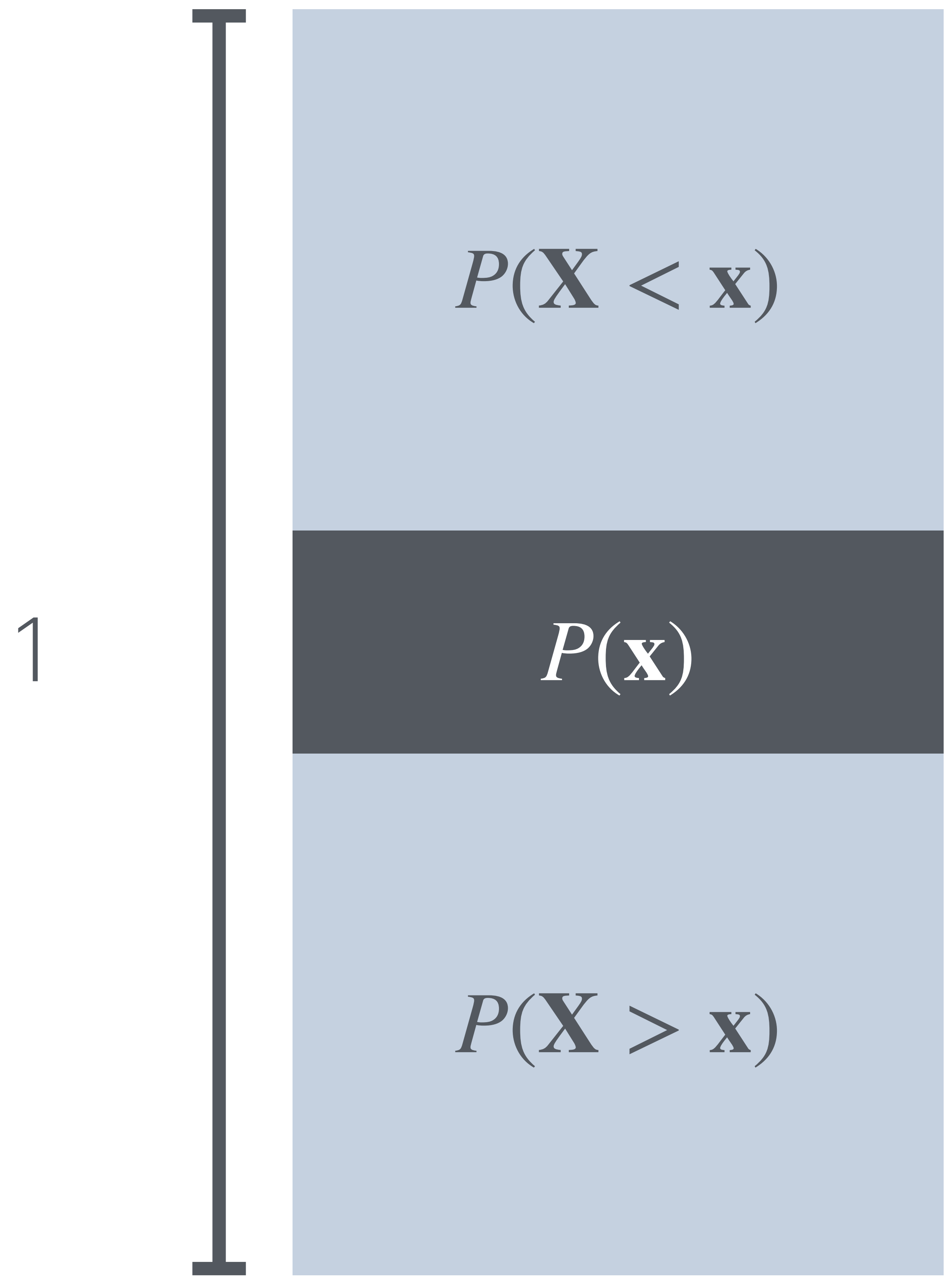
- Knowing $P(\mathbf{x})$ leads to best lossless compression within one bit
    - #bits $= \lfloor -\log_2 P(\mathbf{x}) \rfloor + 1$
- Why?

[1] Lossless Image Compression through Super-Resolution. Sheng Cao, et al. 2020
[2] Practical Full Resolution Learned Lossless Image Compression. Fabian Mentzer, et al. 2019

Arithmetic coding

$\lfloor -\log_2 P(\mathbf{x}) \rfloor + 1$ bit lossless compression

- Sort $\mathbf{x}$ lexicographically
  - Compute CDF $P(\mathbf{X} < \mathbf{x})$
  - Split interval between 0...1 into $2^{\lfloor -\log_2 P(\mathbf{x}) \rfloor + 1}$ numbers
  - Since $2^{\lfloor -\log_2 P(\mathbf{x}) \rfloor + 1} > \dfrac{1}{P(\mathbf{x})}$, at least one number n will end in range $P(\mathbf{X} < \mathbf{x}) \ldots P(\mathbf{X} \leq \mathbf{x})$
  - n is our $\lfloor -\log_2 P(\mathbf{x}) \rfloor + 1$ code



[1] Lossless Image Compression through Super-Resolution. Sheng Cao, et al. 2020
[2] Practical Full Resolution Learned Lossless Image Compression. Fabian Mentzer, et al. 2019

Arithmetic coding in practice

- CDF $P(\mathbf{X} < \mathbf{x})$ generally hard to compute
  - Easy for $P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \,|\, \mathbf{x}_{1\ldots t-1})$
  - $P(\mathbf{X} \leq \mathbf{x}) = \prod_{t=1}^{T} P(X_t \leq x_t \,|\, \mathbf{x}_{1\ldots t-1})$
  - Leads to adaptive arithmetic coding

[1] Lossless Image Compression through Super-Resolution. Sheng Cao, et al. 2020
[2] Practical Full Resolution Learned Lossless Image Compression. Fabian Mentzer, et al. 2019

Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x\,|\,z)$ or $f: z \to x$

z Deep Network 

Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

 Deep Network $P(X)$

References

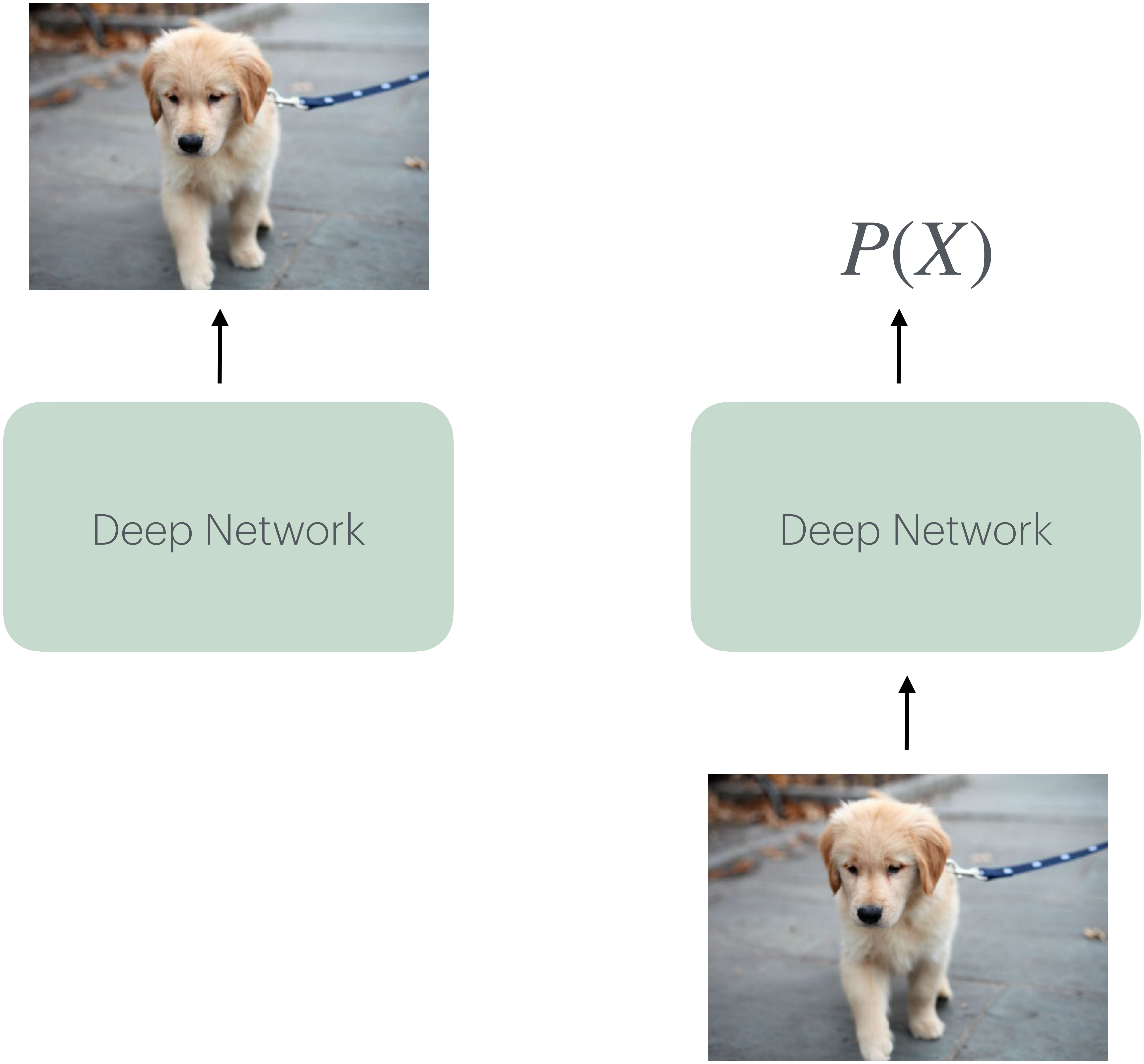
- [1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016
- [2] Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. Songwei Ge, et al. 2022
- [3] Lossless Image Compression through Super-Resolution. Sheng Cao, et al. 2020
- [4] Practical Full Resolution Learned Lossless Image Compression. Fabian Mentzer, et al. 2019

Vector Quantization

Philipp Krähenbühl, UT Austin

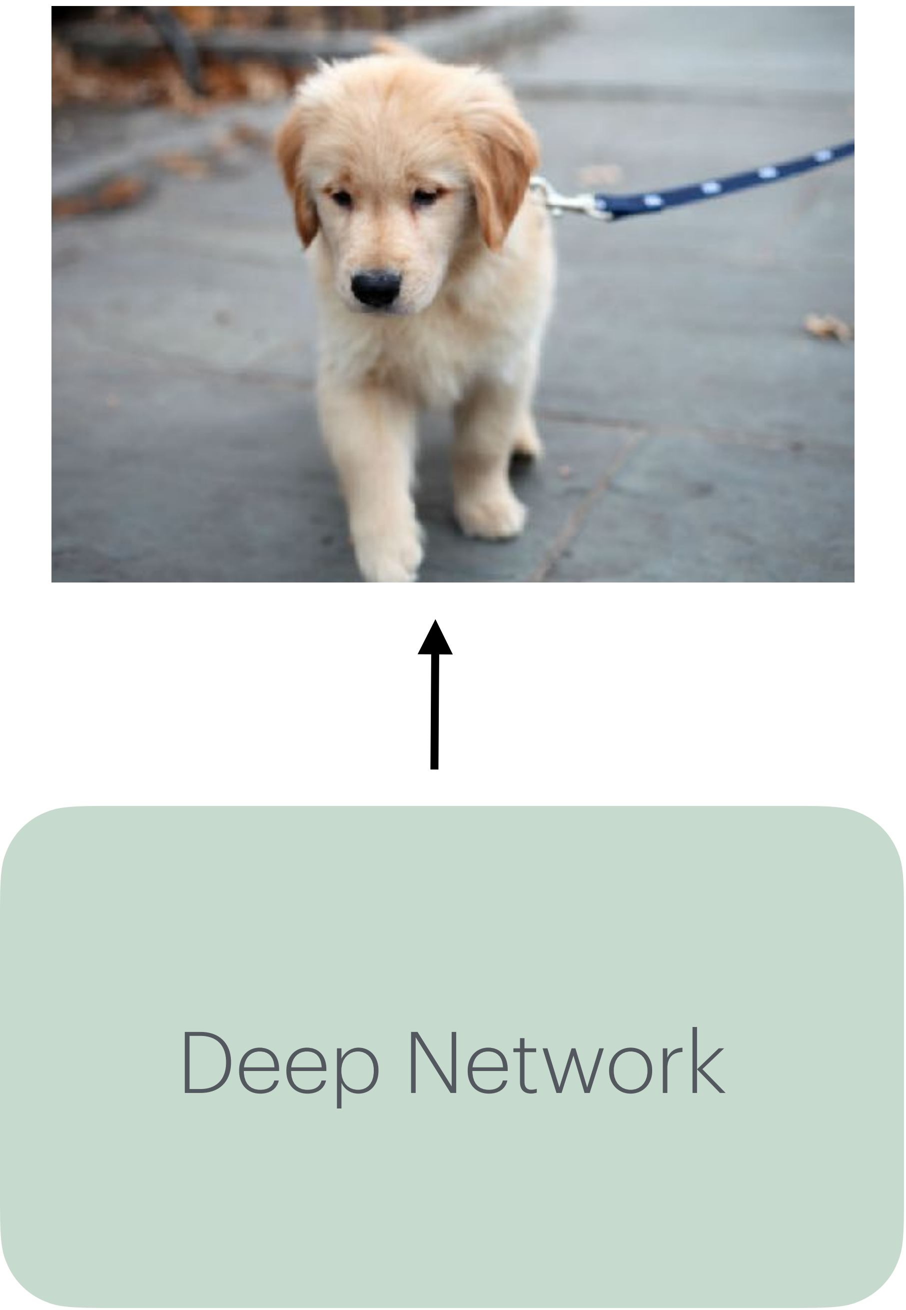
Generative models

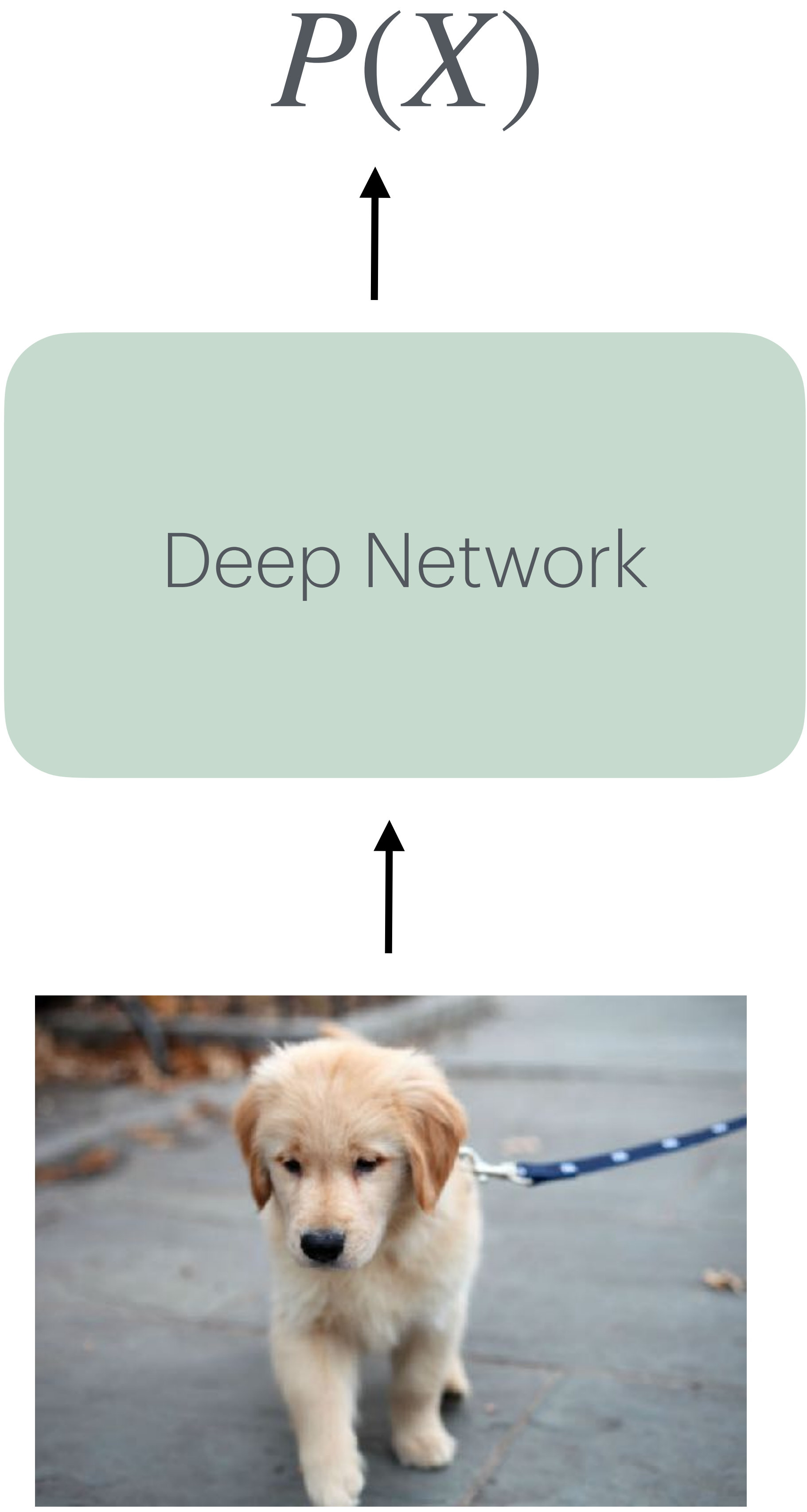
- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$



Generative modeling is hard

- Density estimation $P(X = x)$
  - How to ensure $\sum_x P(x) = 1$ for all x
  - Impossible to compute (in general)
- Sampling $x \sim P(X)$
  - What is the input to the network?





Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x|z)$ or $f : z \to x$

z Deep Network

Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

Deep Network $P(X)$

Auto-regressive models

Issues

$$P(x) = P(x_1)P(x_2 | x_1)P(x_3 | x_1, x_2)P(x_4 | x_1 \dots x_3)\dots$$

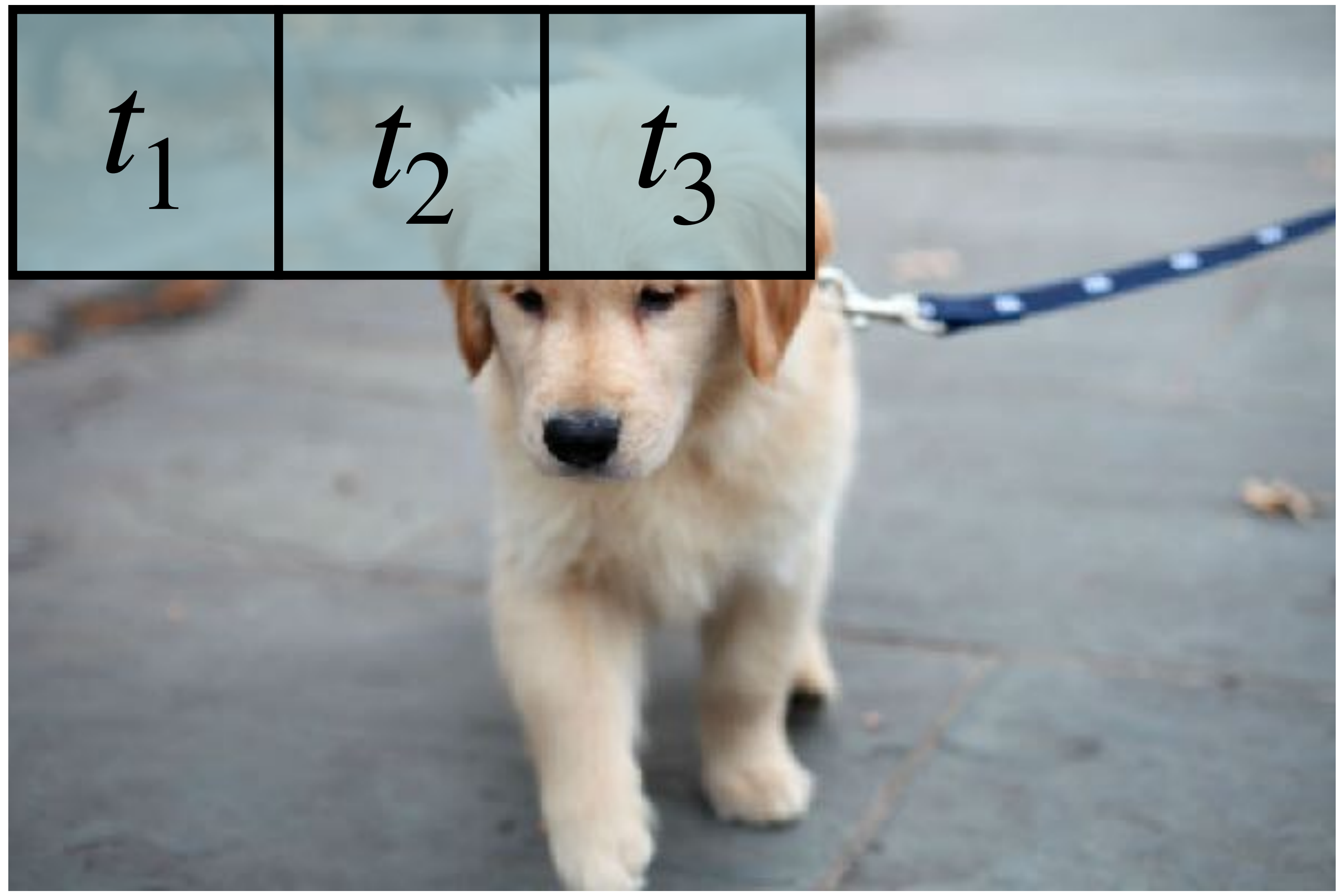
- Difficult learning problem for long sequences (requires good model)

[1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016
[2] Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. Songwei Ge, et al. 2022

Tokenization

- Image [1]
  - Convert patch p_i of pixels into token $t_i \in \{1, \ldots, K\}$
- Text [2]
  - Convert set of characters into token
- Protein-sequence [3]
  - Convert local protein structure to token

Vanilla auto-regressive model



Tokenized auto-regressive model

[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017
[2] Language models are unsupervised multitask learners. Alec Radford, et al. 2019
[3] Simulating 500 million years of evolution with a language model. Thomas Hayes, et al. 2024

Auto-regressive models on tokens

$$P(\mathbf{t}) = P(t_1)P(t_2 \mid t_1)P(t_3 \mid t_1, t_2)P(t_4 \mid t_1 \ldots t_3)\ldots$$

- Shorter sequence = easier to learn structure

[1] MAGVIT: Masked Generative Video Transformer. Lijun Yu, et al. 2023

Learning Tokenization

Vector Quantization

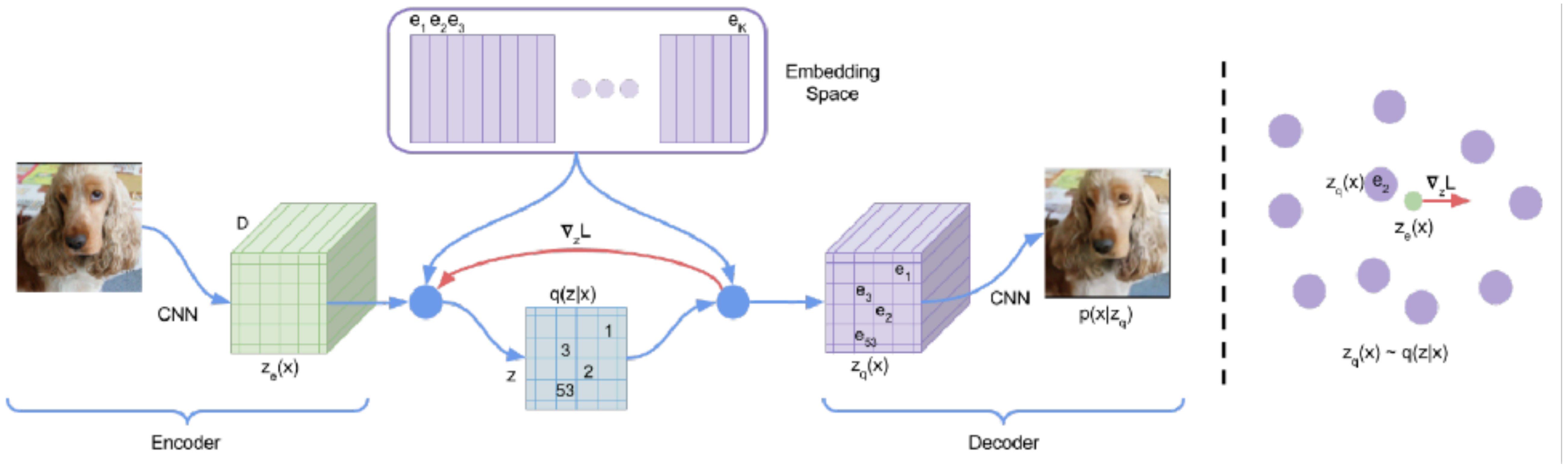
- Input: Image (or patch)
  $x \in \mathbb{R}^{H \times W \times 3}$

- Output: "Image" of tokens
  $z \in \{1 \ldots K\}^{h \times w}$

- Why is this hard to learn?

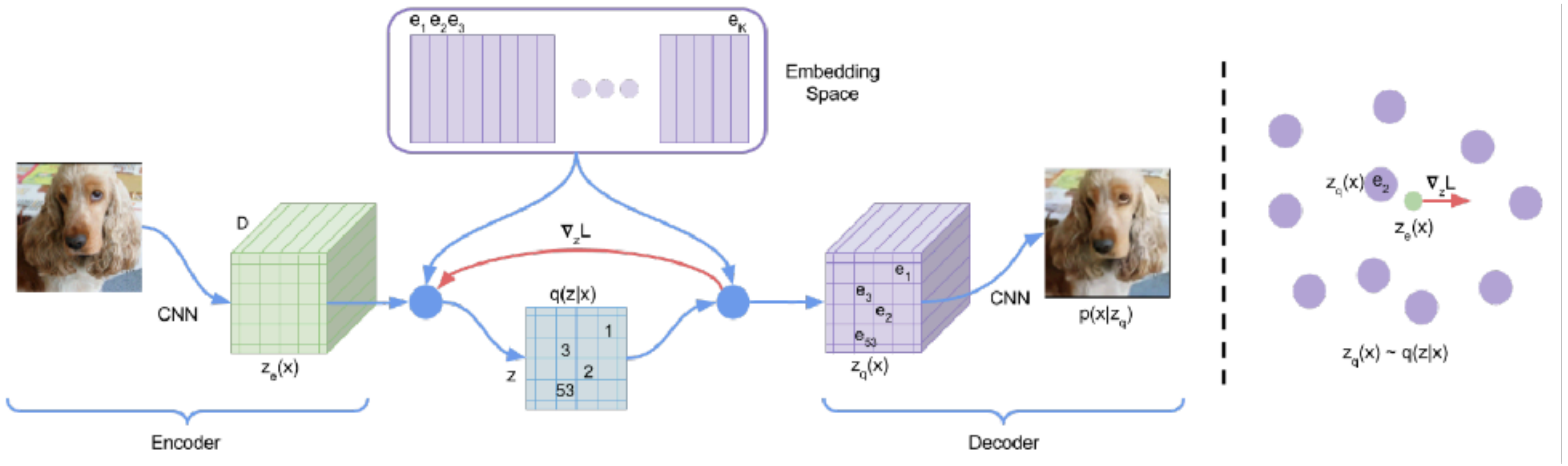
  - $z \rightarrow x$ (easy, reconstruction)

  - $x \rightarrow z \rightarrow x$ (hard, z is discrete and non-differentiable)



[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

VQ-VAE

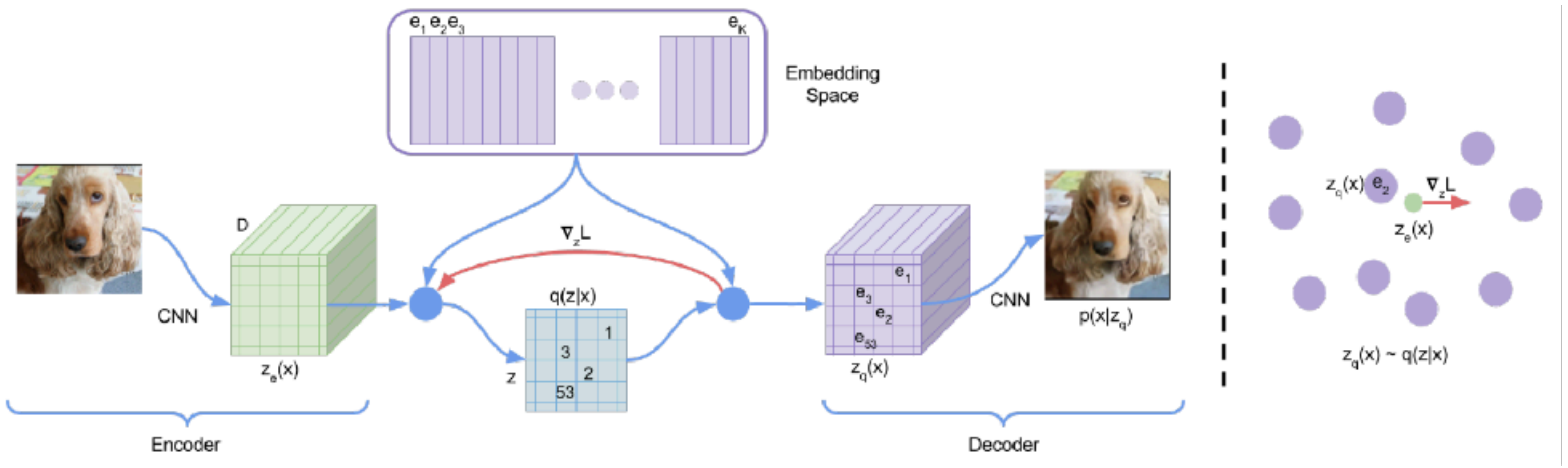


- Variational Auto-Encoder
  - Decoder $P_D(x \mid z)$ Encoder $Q(z \mid x)$
- Vector Quantizer
  - $q(z) = \arg\min_{e_k} ||z - e_k||$
  - Learn codebook $\{e_1 \ldots e_K\}$
  - What is $\nabla q(z)$?

[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

VQ-VAE

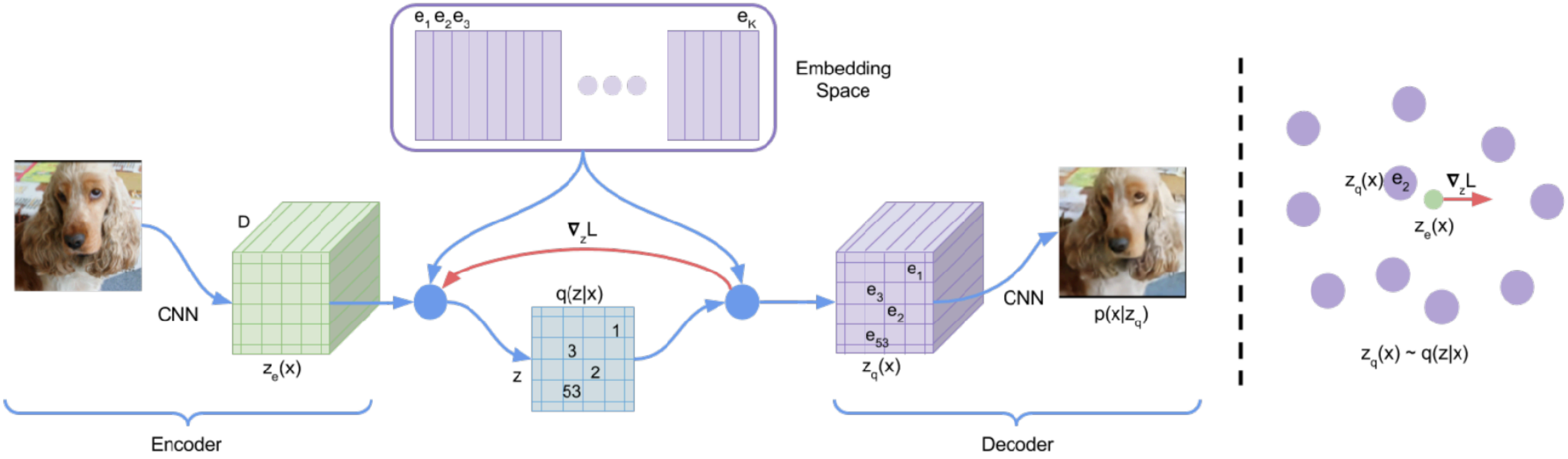
Gradient



- What is $\nabla q(z)$?
  - Let's assume $\nabla q(z) = \mathbf{I}$ (identity)
  - Straight-Through Estimator
    - Works in practice because errors average out over large enough batches
    - No reason it should work

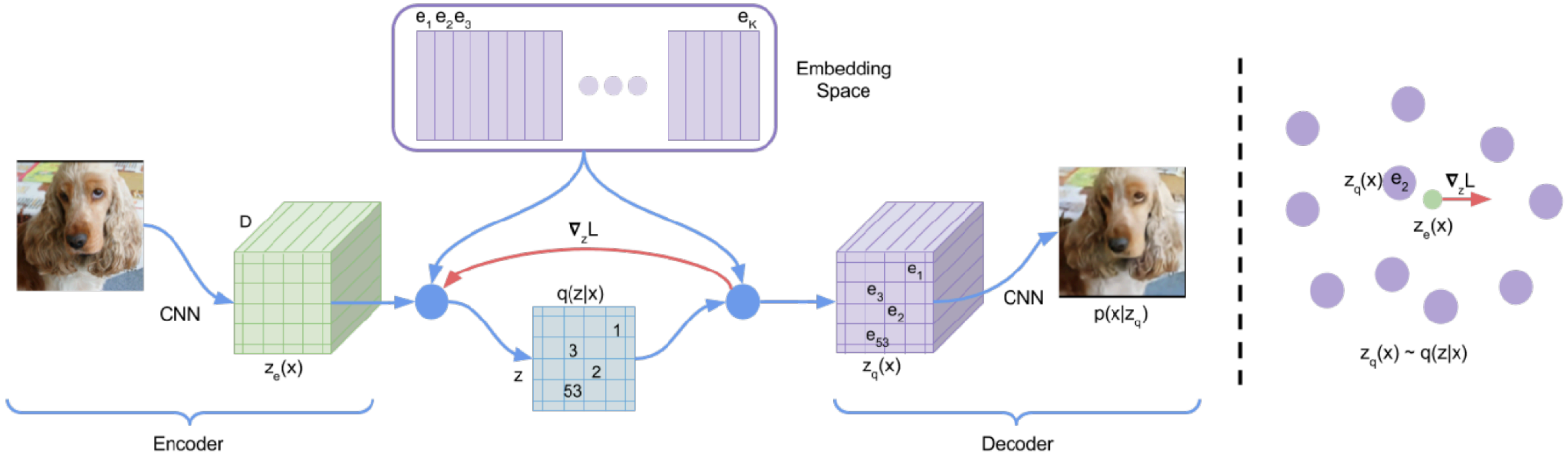
[1] Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. Yoshua Bengio, et al. 2013

VQ-VAE



[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

VQ-VAE

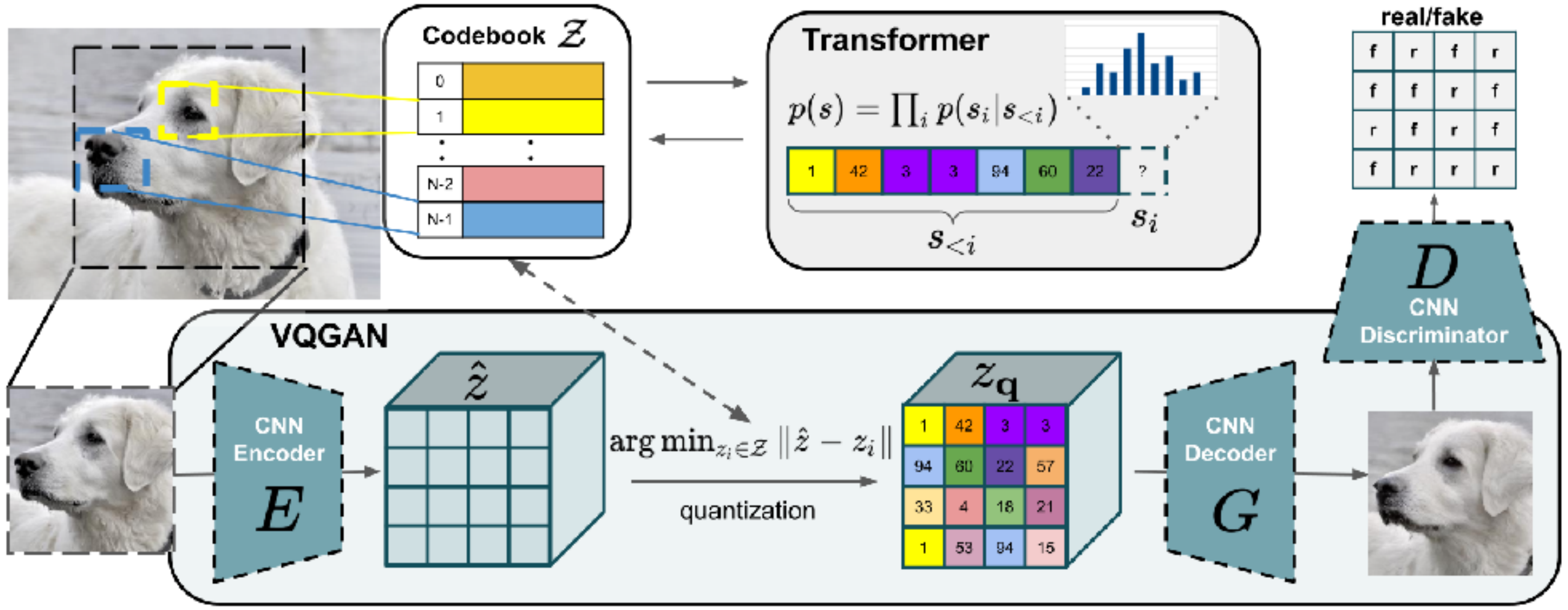


- Only as good as VAE
- Does not scale well with codebook size
  - Codebook grows exponentially in #bits
  - Many entries → sparse gradients
  - Slow

[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

VQ-GAN

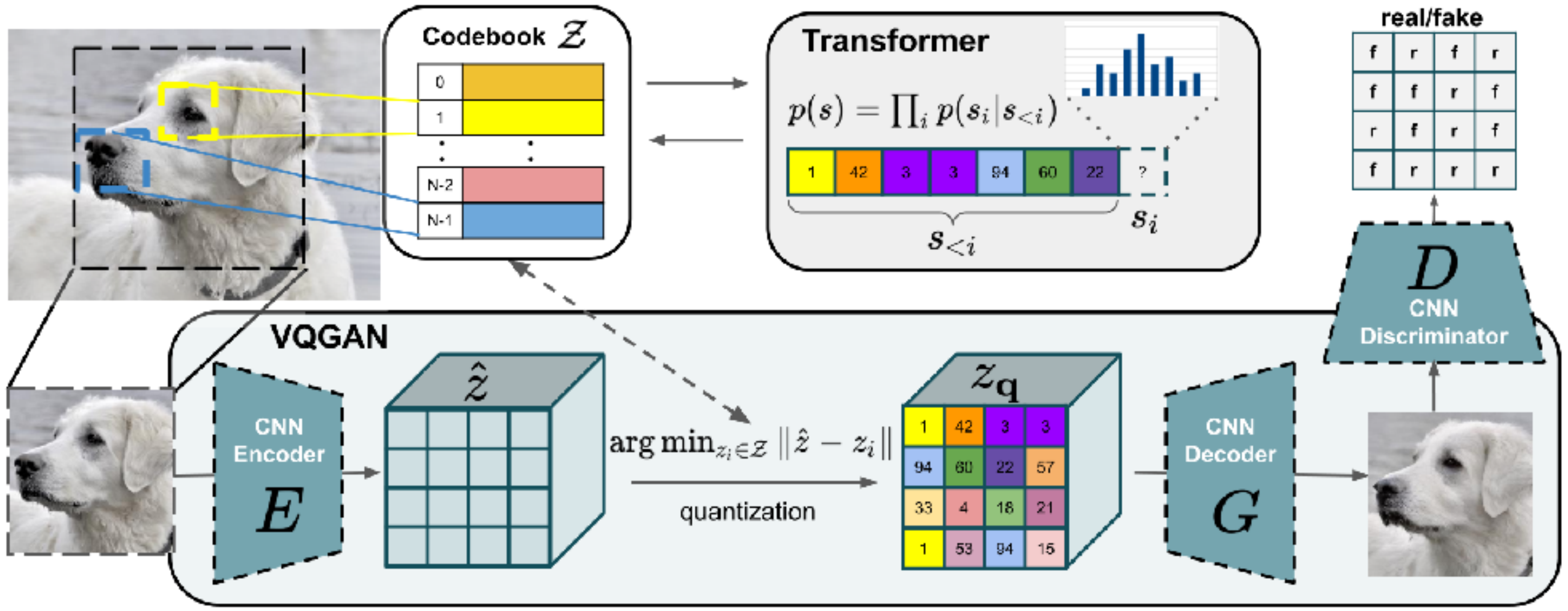
- Replace VAE with GAN
    - Auto-encoder with vector quantization
      $$q(z) = \arg\min_{e_k} ||z - e_k||$$
    - GAN + Reconstruction loss
- Learn a sequence model on top
- Default image tokenizer nowadays

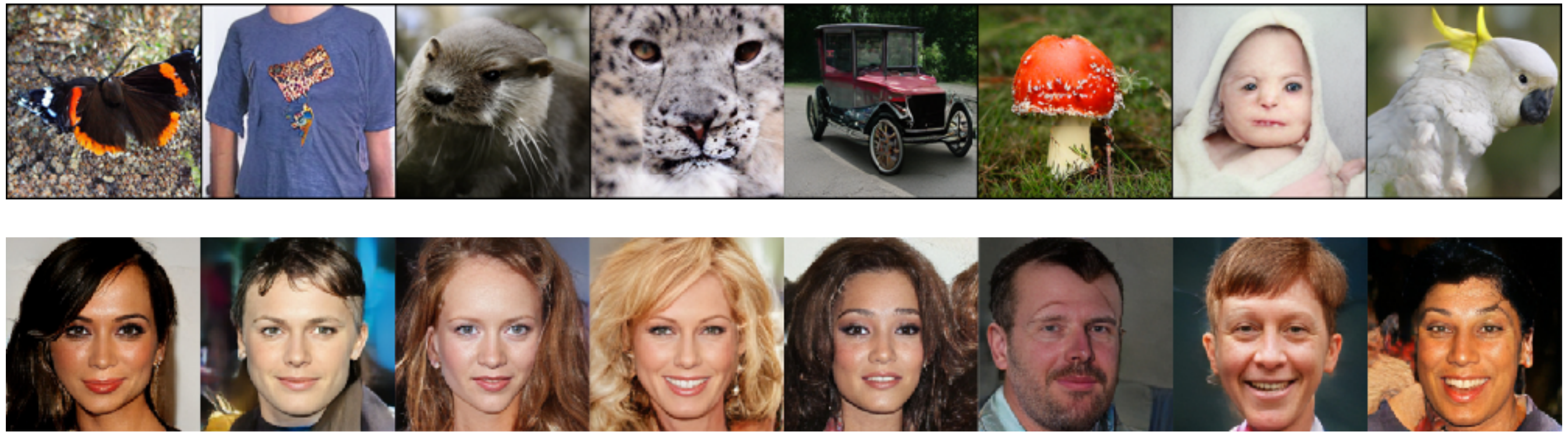


[1] Taming transformers for high-resolution image synthesis. Patrick Esser et al. 2021

VQ-GAN

- Great tokenizer, ok sequence model
- Does not scale well with codebook size
    - Codebook grows exponentially in #bits
    - Many entries → sparse gradients
    - Slow



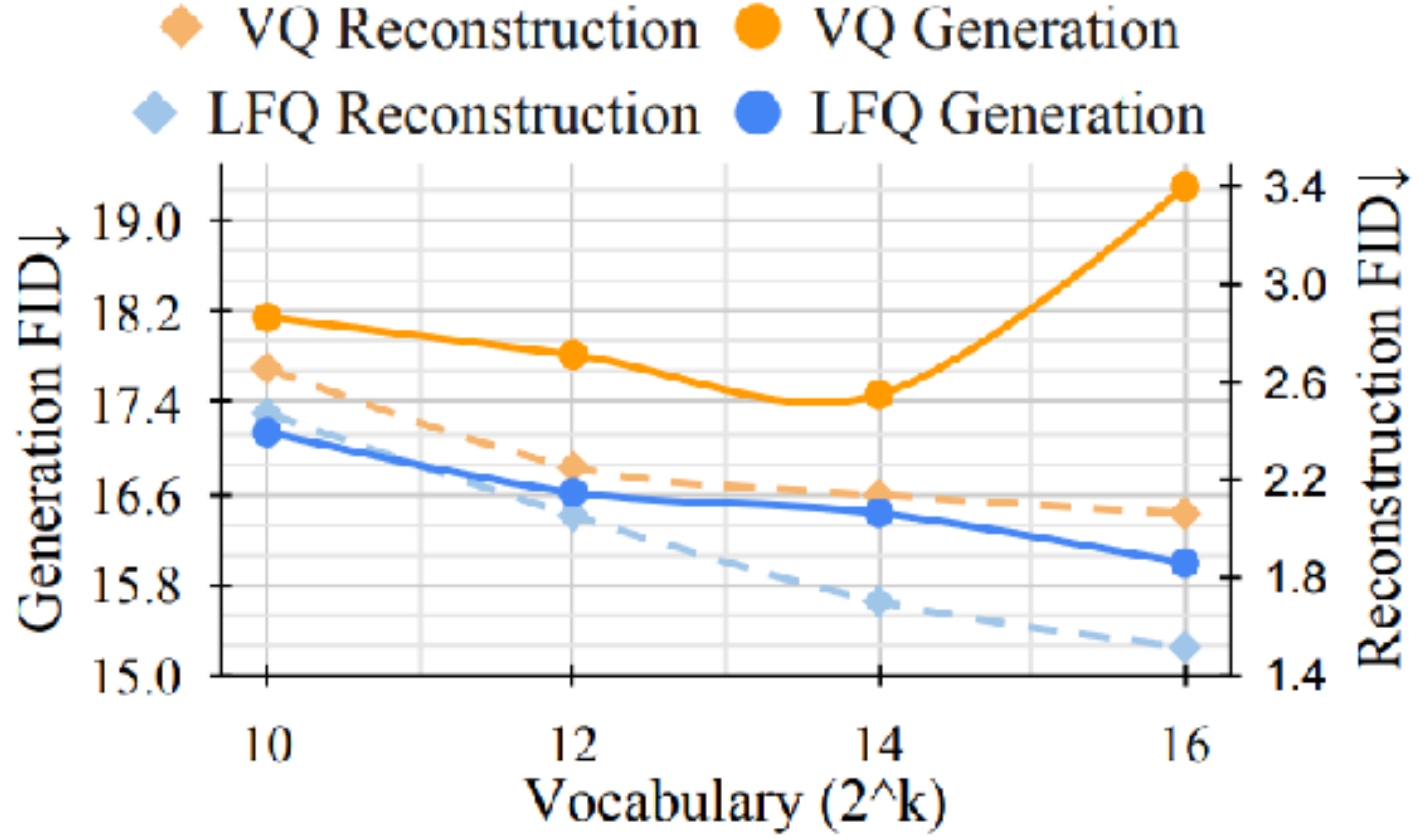


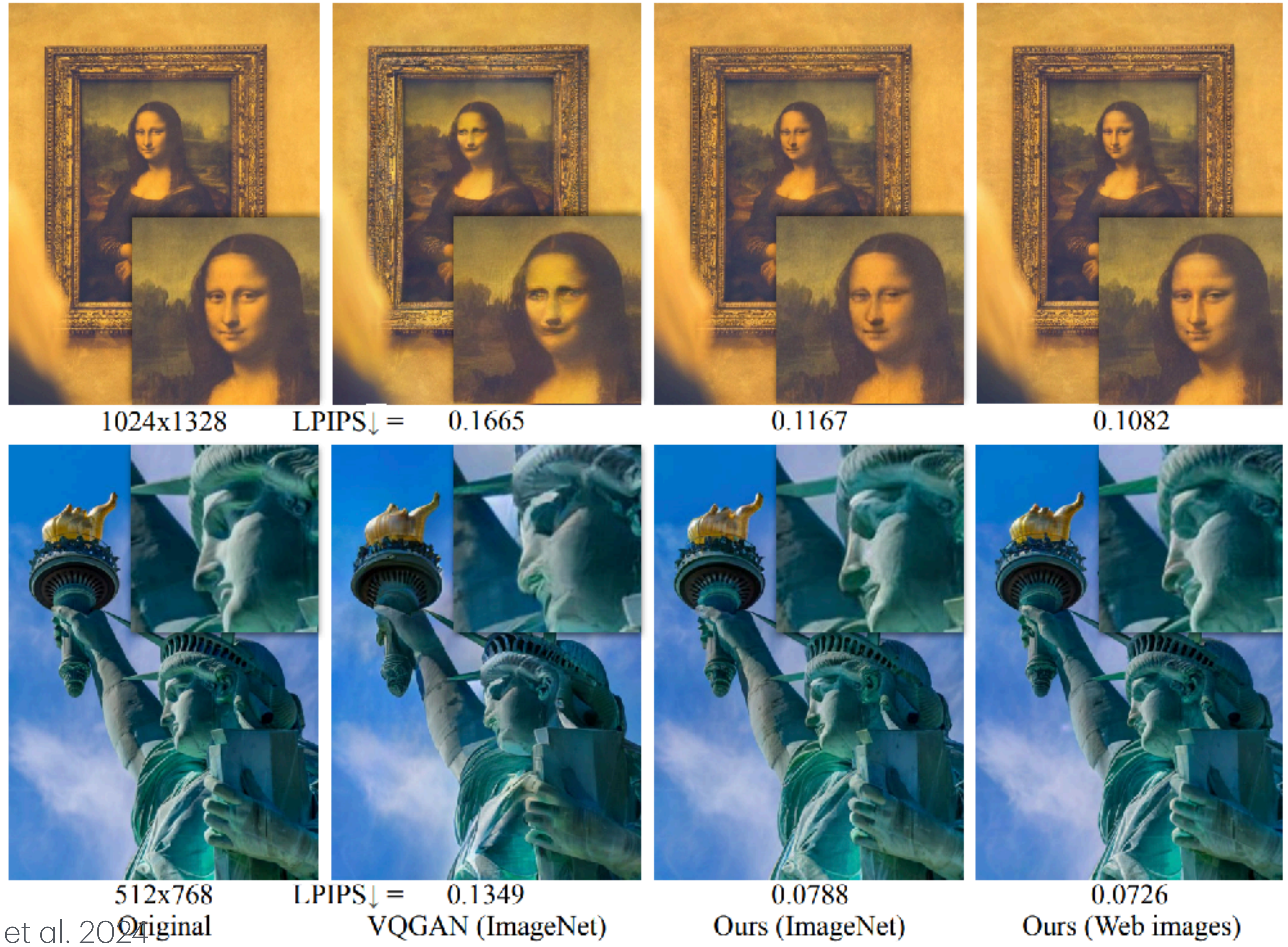
[1] Taming transformers for high-resolution image synthesis. Patrick Esser et al. 2021

LFQ

Lookup-Free Quantization

- Different quantizer
  - $q(z) = \text{sign}(z)$ where $\text{sign}(z_i) = 1_{[z_i \le 0]} - 1_{[z_i > 0]}$
- Scales linearly with #bits in bottleneck
- No learned parameters

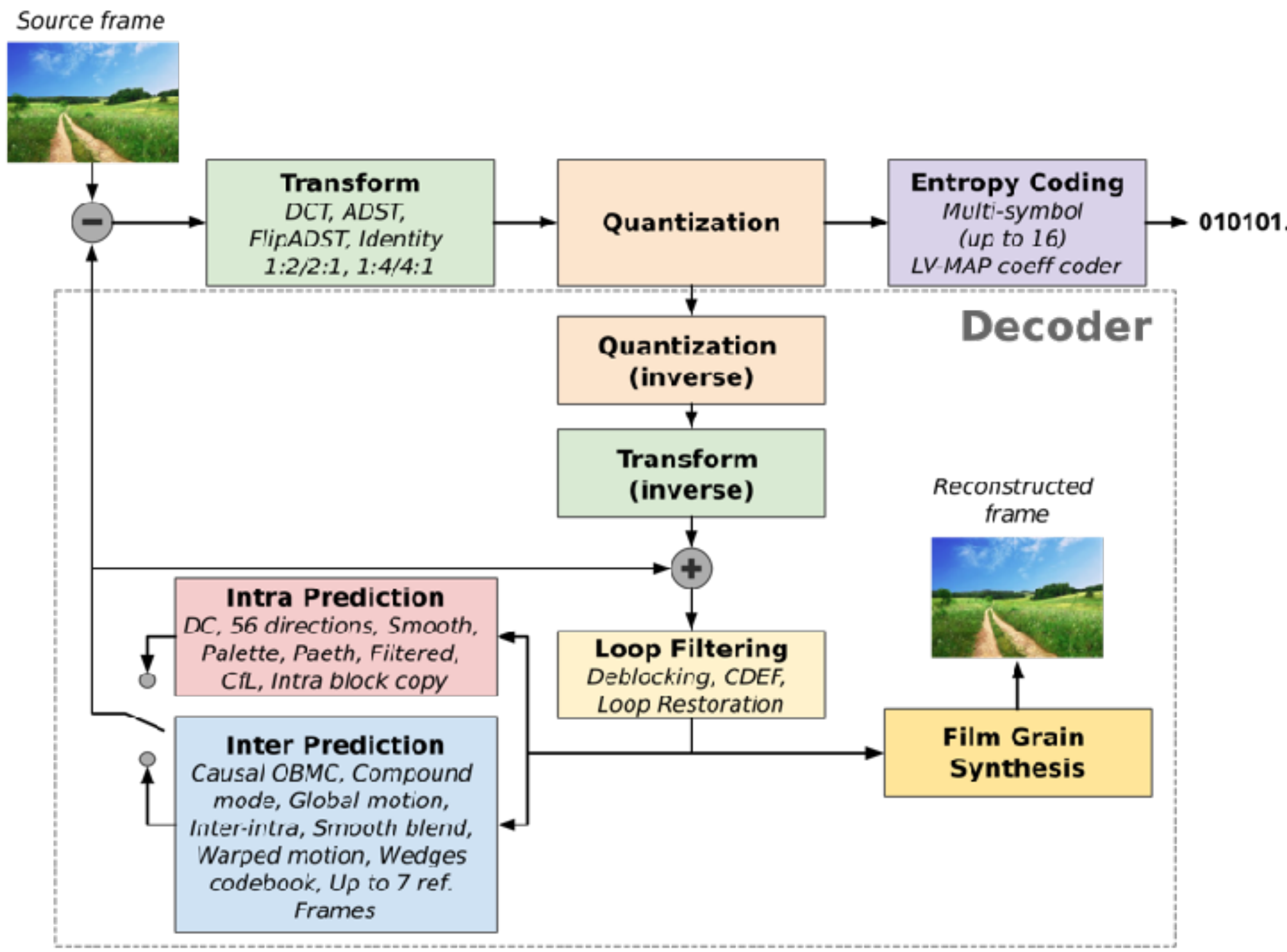




[1] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation. Lijun Yu, et al. 2024

Generation vs Compression

- Auto-regressive model
  - Lossless compression (fancy gzip)
- Tokenization (VQ)
  - Lossy compression
- Similar to how JPEG most video codecs work



Source: https://commons.wikimedia.org/wiki/File:The_Technology_Inside_Av1.svg

Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x \mid z)$ or $f: z \to x$

z Deep Network

Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

Deep Network $P(X)$

References

- [1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016
- [2] Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. Songwei Ge, et al. 2022
- [3] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017
- [4] Language models are unsupervised multitask learners. Alec Radford, et al. 2019
- [5] Simulating 500 million years of evolution with a language model. Thomas Hayes, et al. 2024
- [6] MAGVIT: Masked Generative Video Transformer. Lijun Yu, et al. 2023
- [7] Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. Yoshua Bengio, et al. 2013
- [8] Taming transformers for high-resolution image synthesis. Patrick Esser et al. 2021
- [9] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation. Lijun Yu, et al. 2024