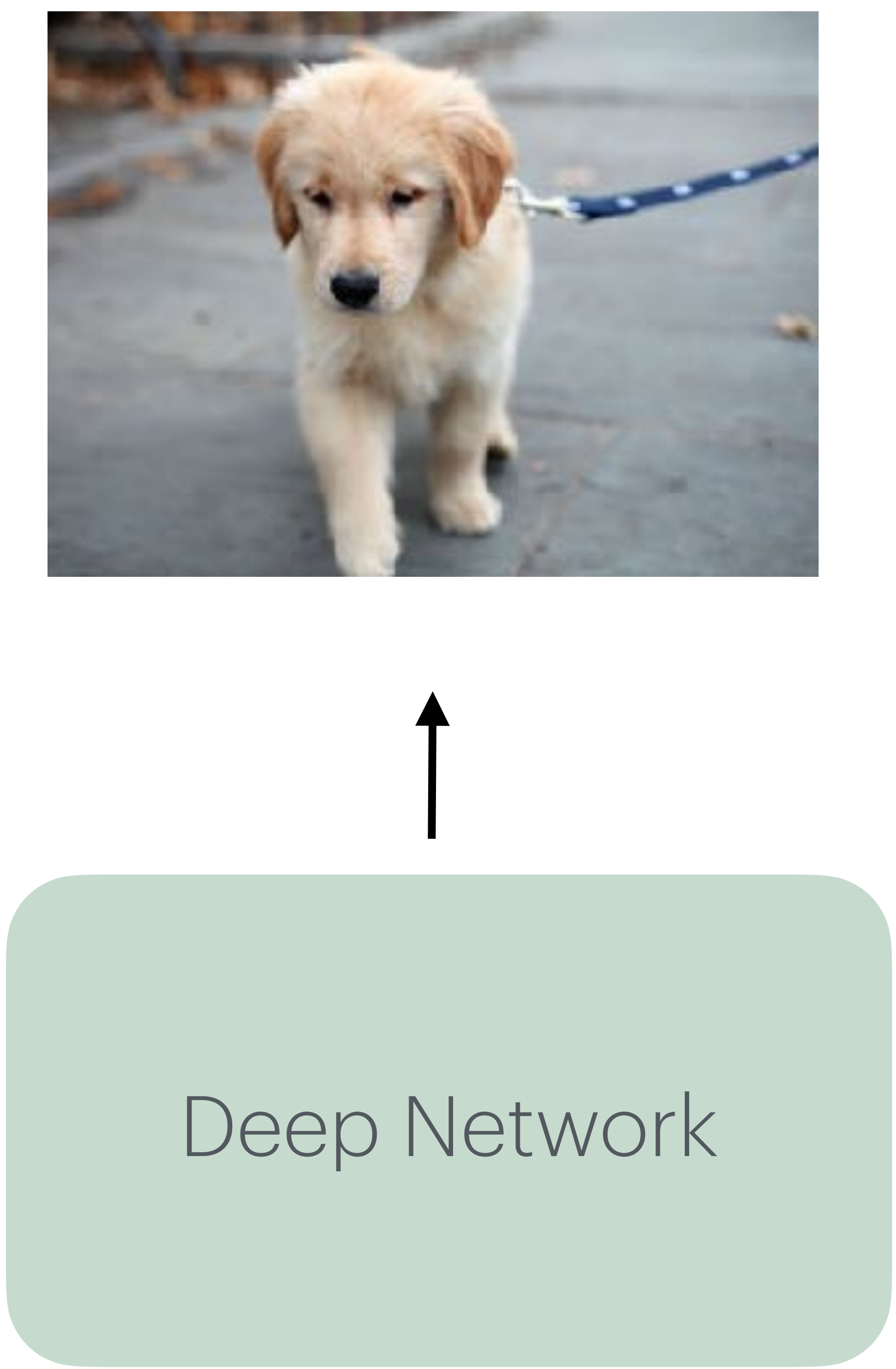
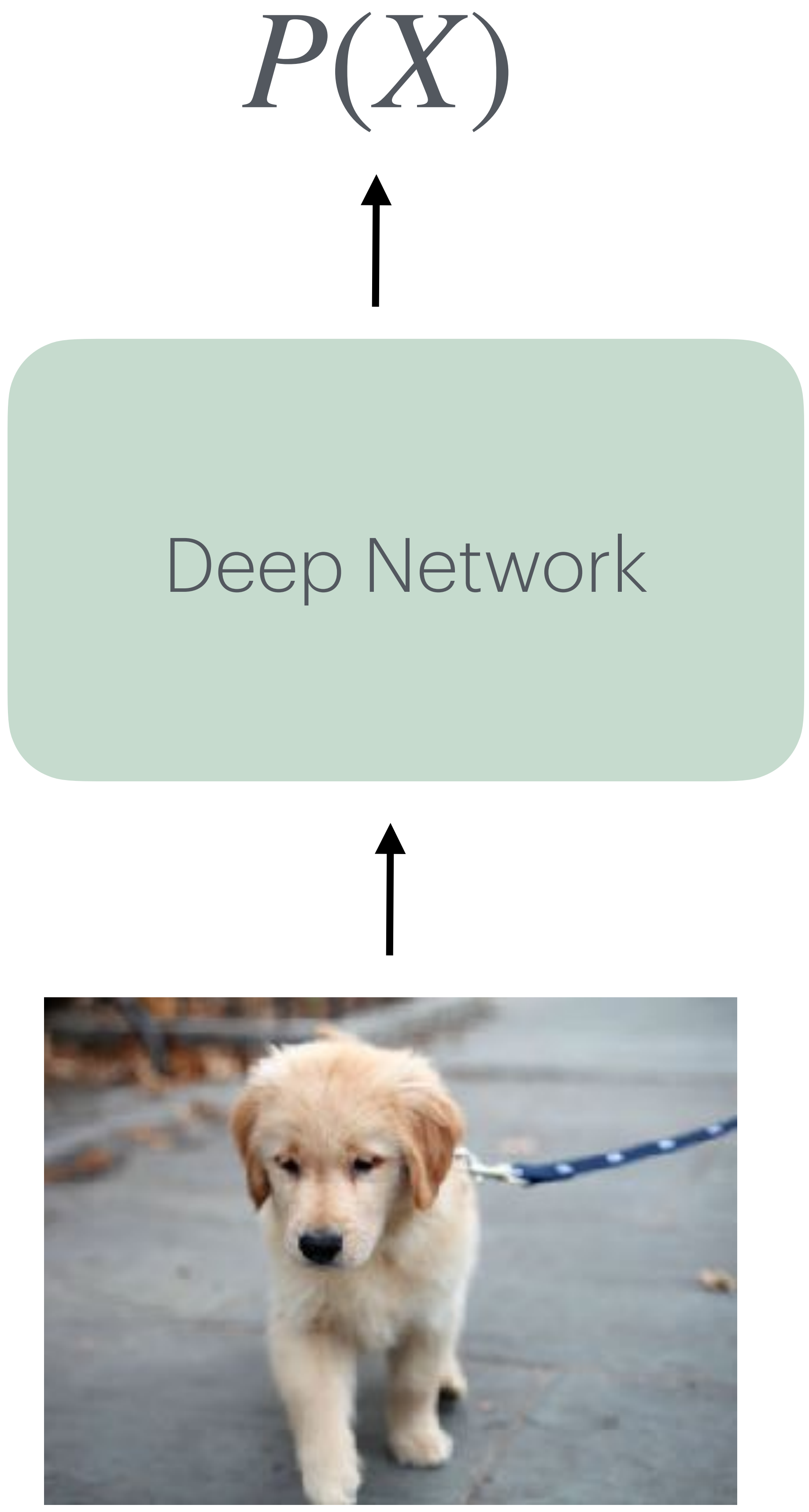
Auto-regressive generation

Philipp Krähenbühl, UT Austin

Generative models

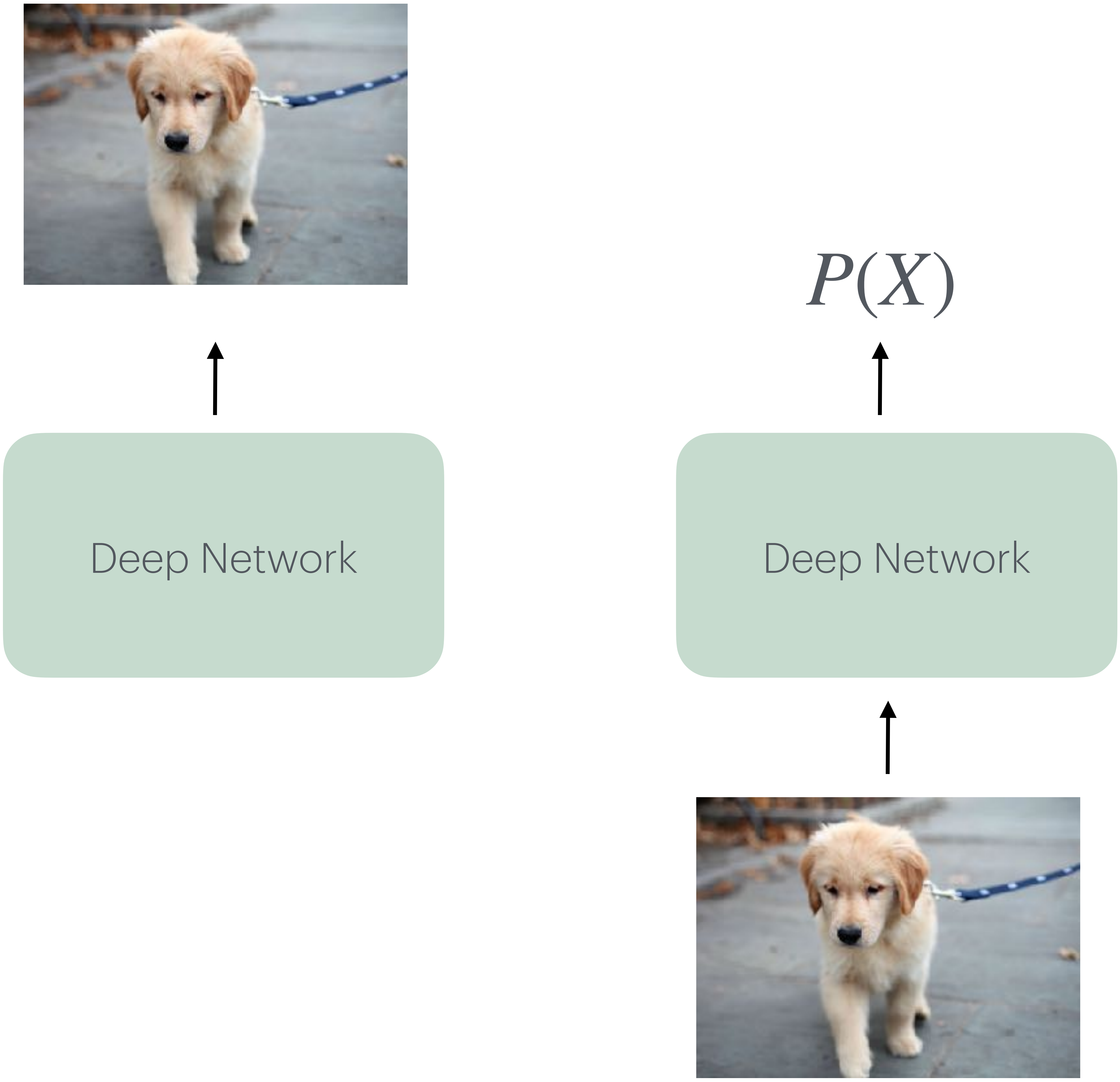
- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$





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?

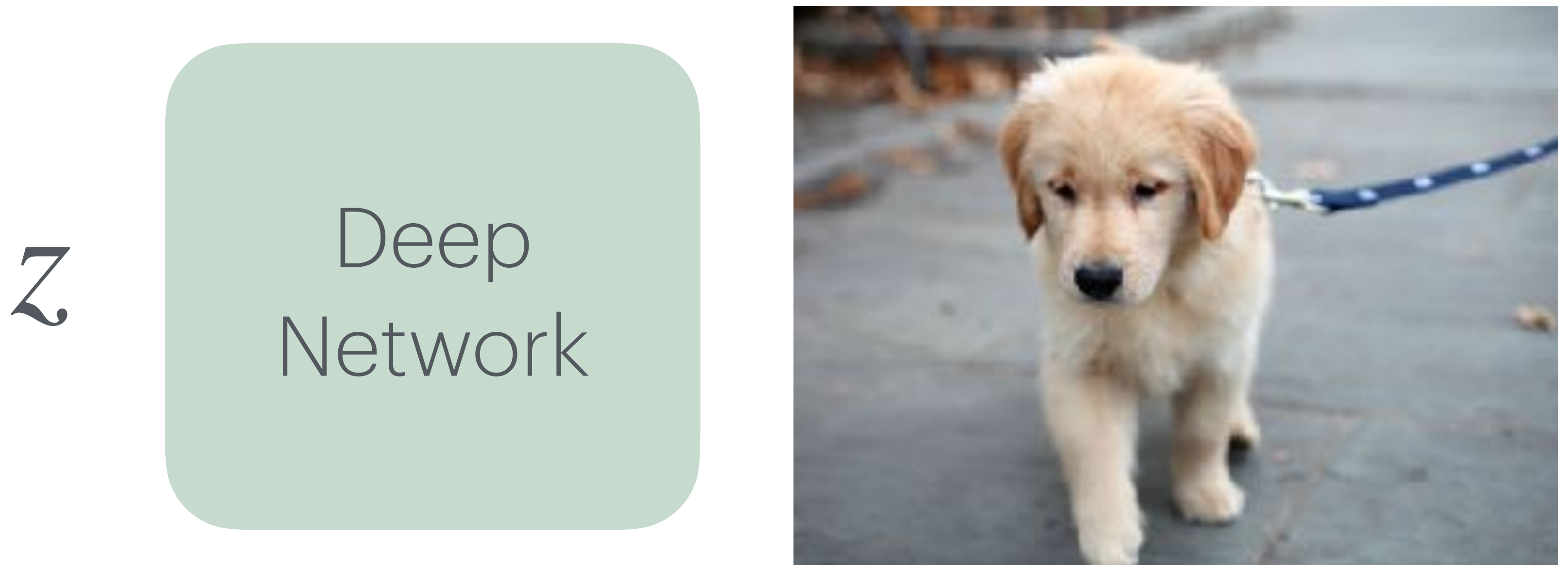


Generative models

Two kinds of models

Sampling based $x \sim P(X)$

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



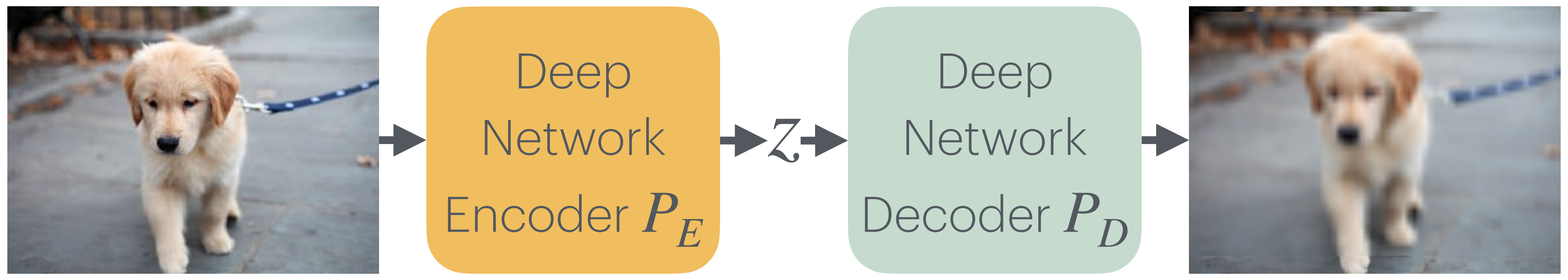
Density estimation based $P(X)$

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

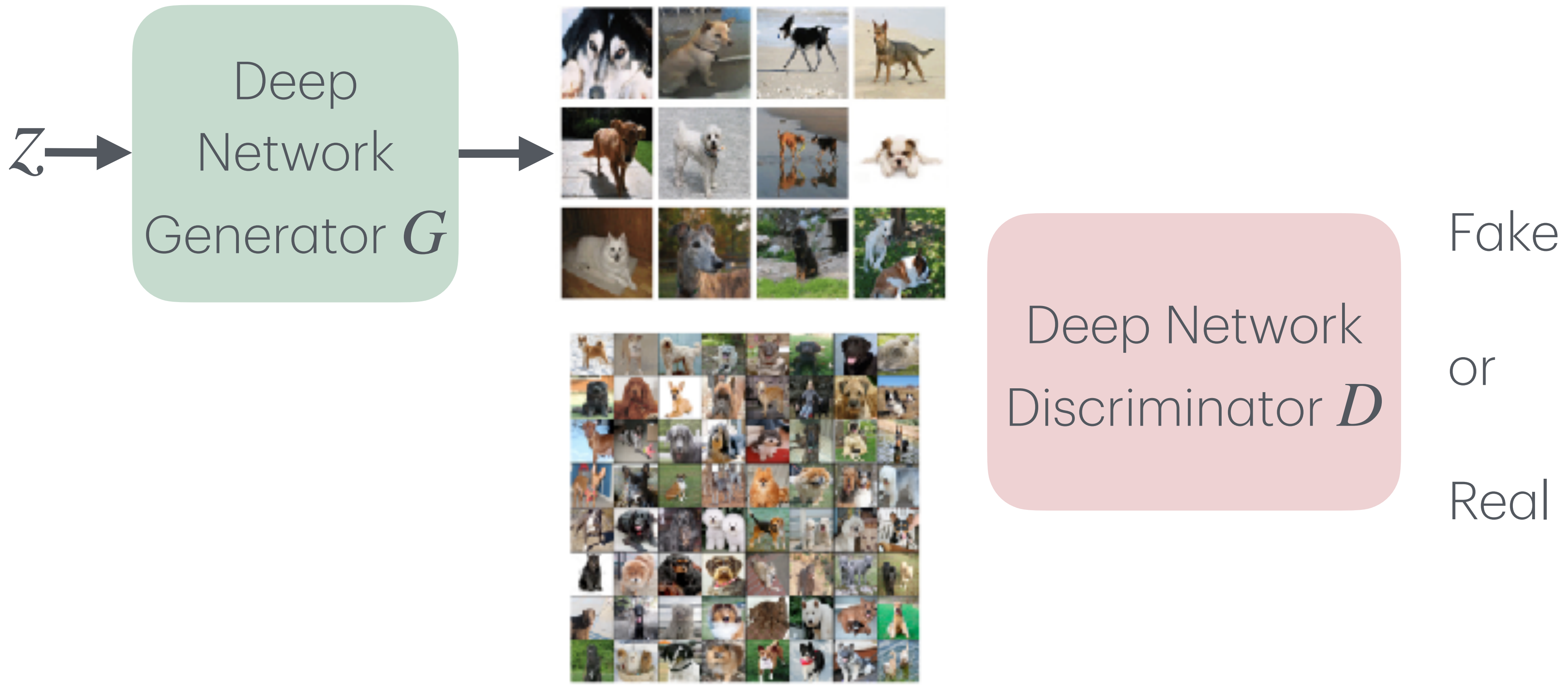
Recap

- VAE
  - Image -> latent space -> Image
  - Loss encourages Gaussian latent
- GAN
  - Gaussian -> Image
  - Loss compares distributions
- Flow-based
  - Gaussian ↔ Image
  - Requires Invertible architecture

Variational Auto Encoder (VAE)

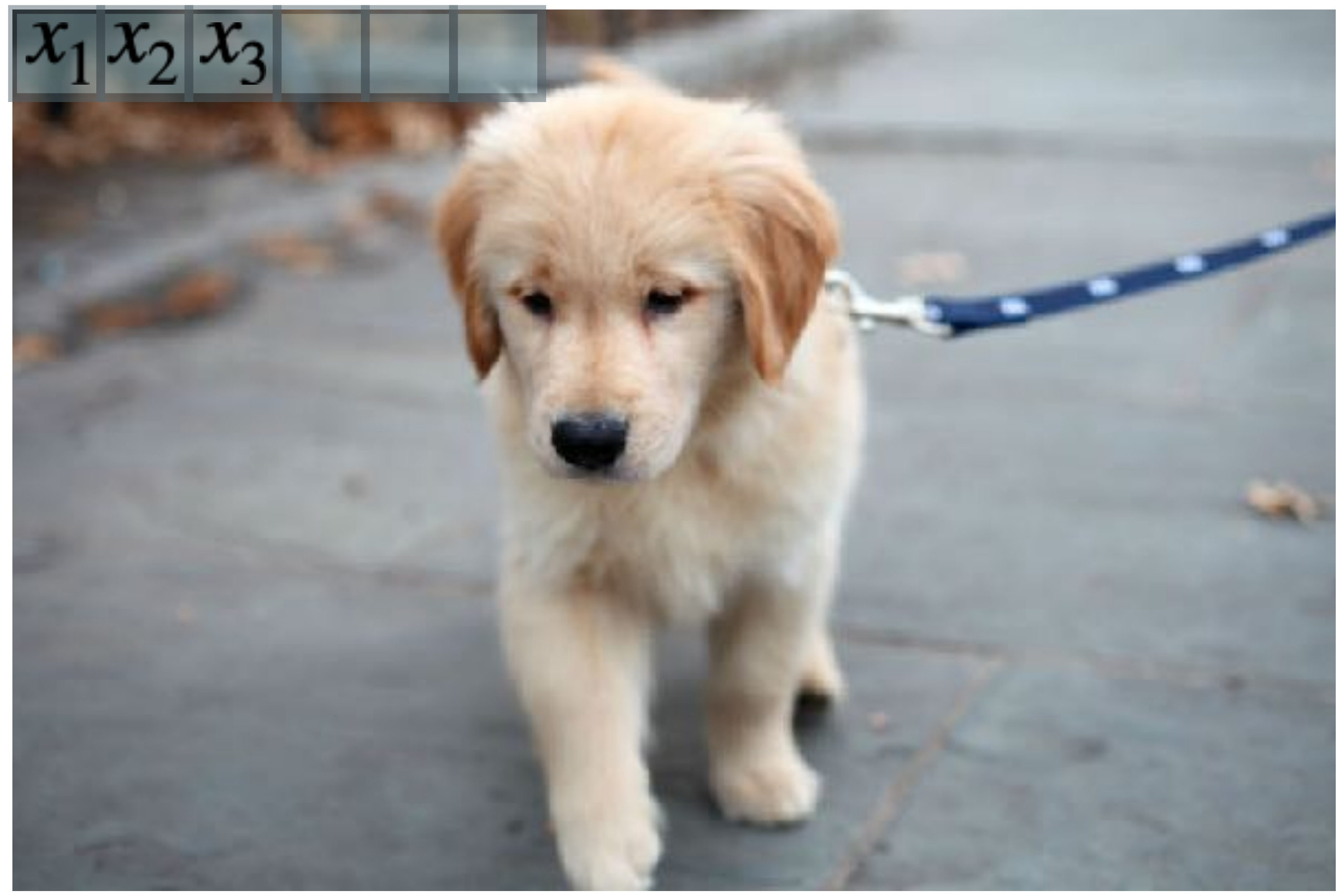


Generative Adversarial Network (GAN)



Flow-based models

Auto-regressive models

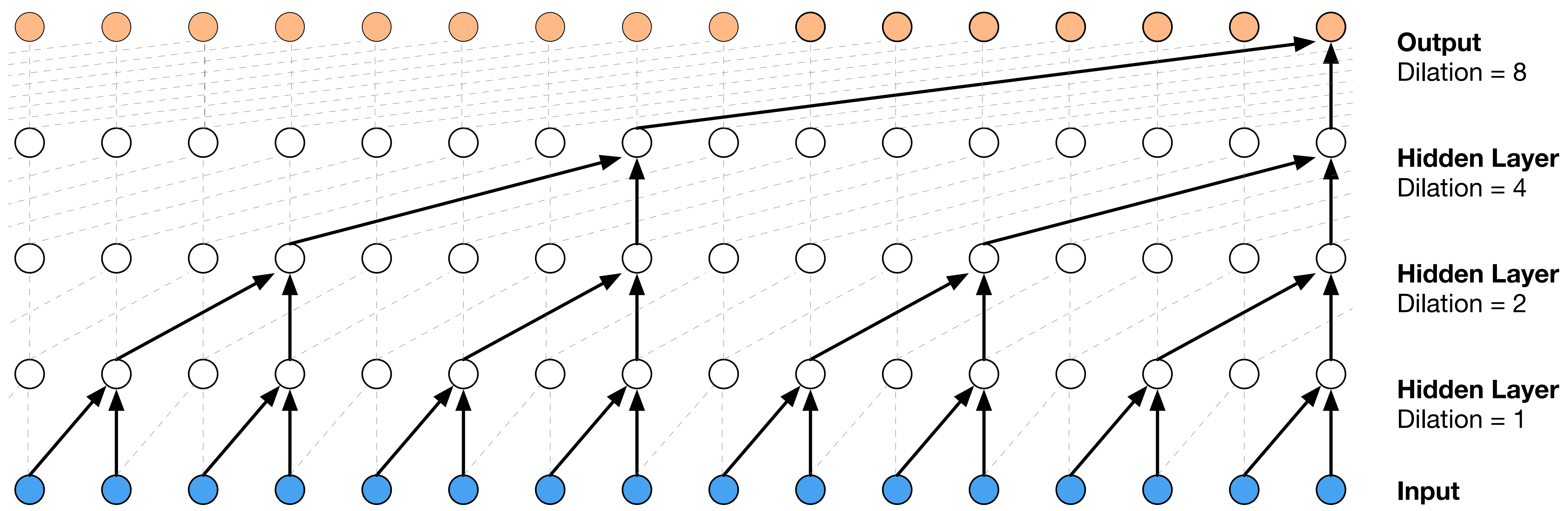


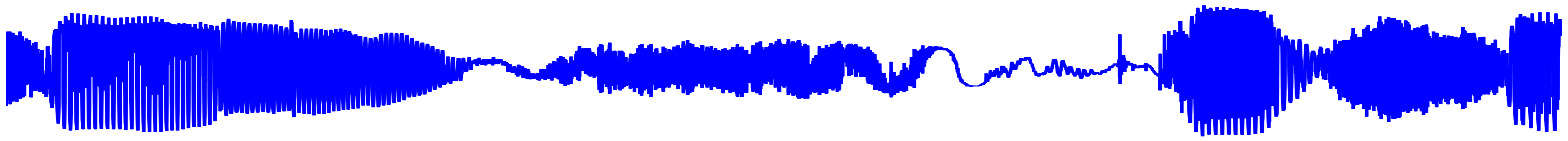
$P(x) = P(x_1)P(x_2 | x_1)P(x_3 | x_1, x_2)P(x_4 | x_1 \ldots x_3)\ldots$

- $P(x_i | x_1 \ldots x_{i-1}) = \text{softmax}(f(x_1 \ldots x_{i-1}))$
- 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



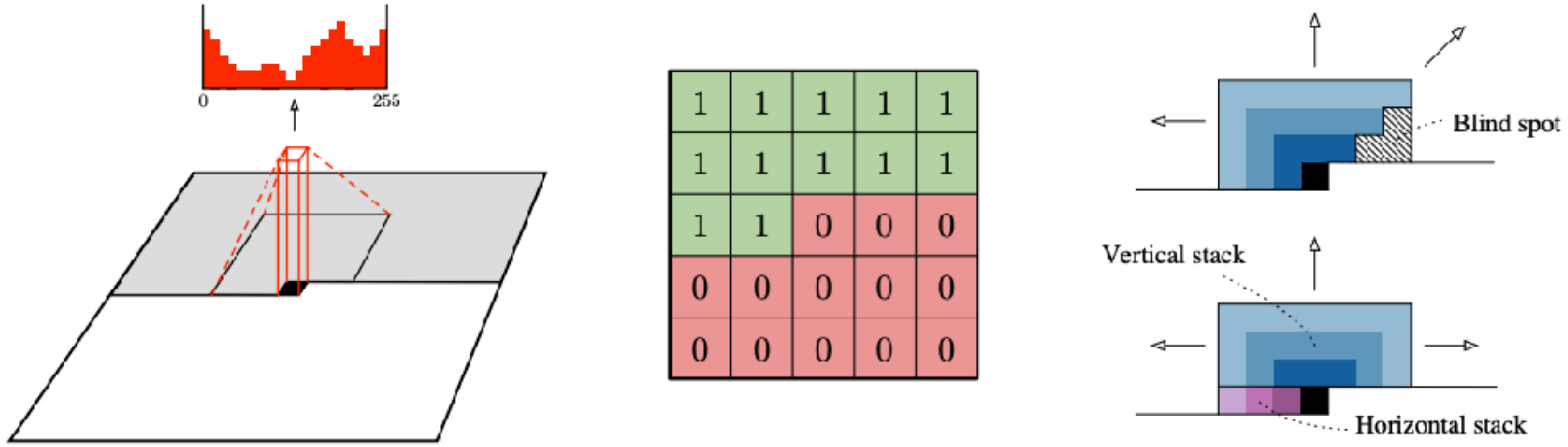


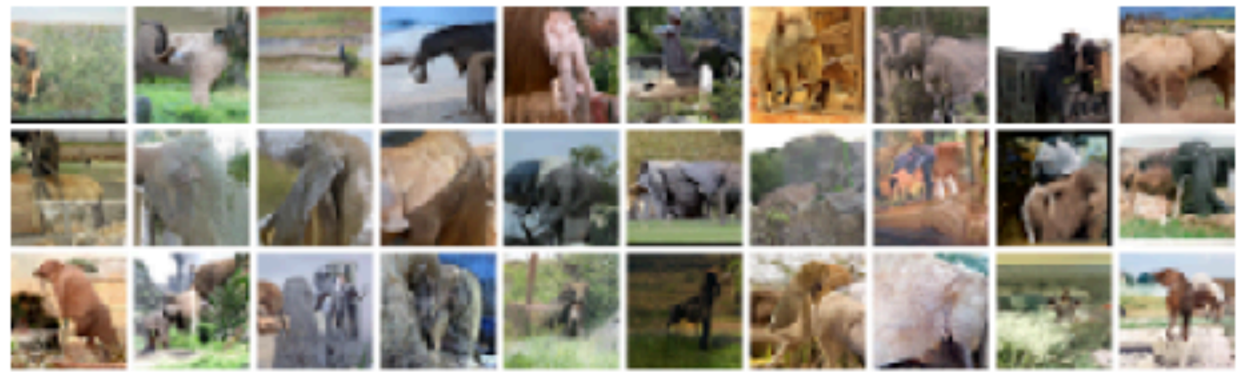
- Input: Raw waveform $\mathbf{x}_{1\ldots t-1}$
- Output: Quantized next value $\mathbf{x}_t \in \{1\ldots256\}$
- Model: $P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \mid \mathbf{x}_{1\ldots t-1})$
- Conditioned model: $P(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} P(x_t \mid \mathbf{x}_{1\ldots t-1} \mid \mathbf{h})$

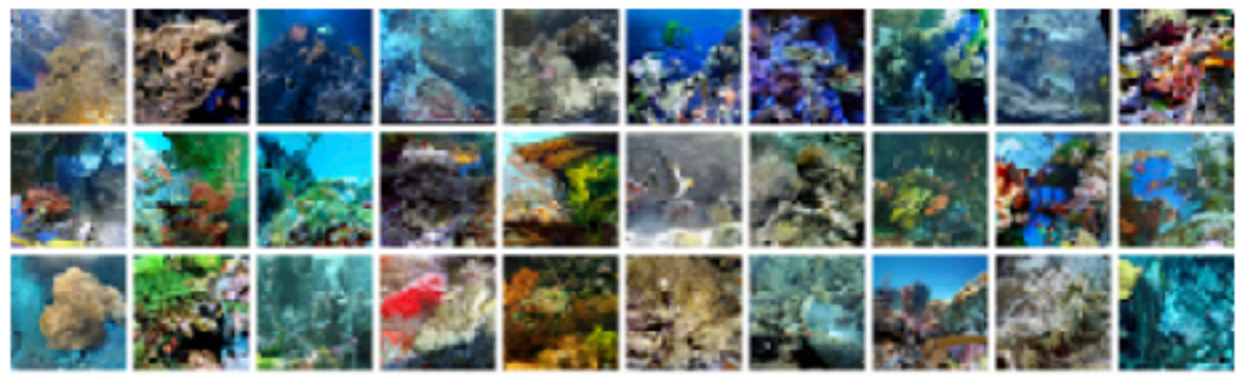
[1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016

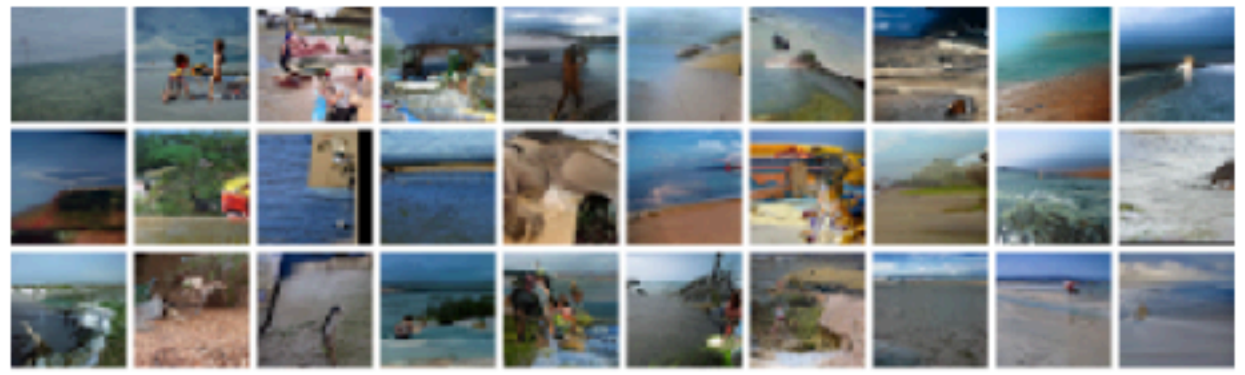
Example: PixelCNN

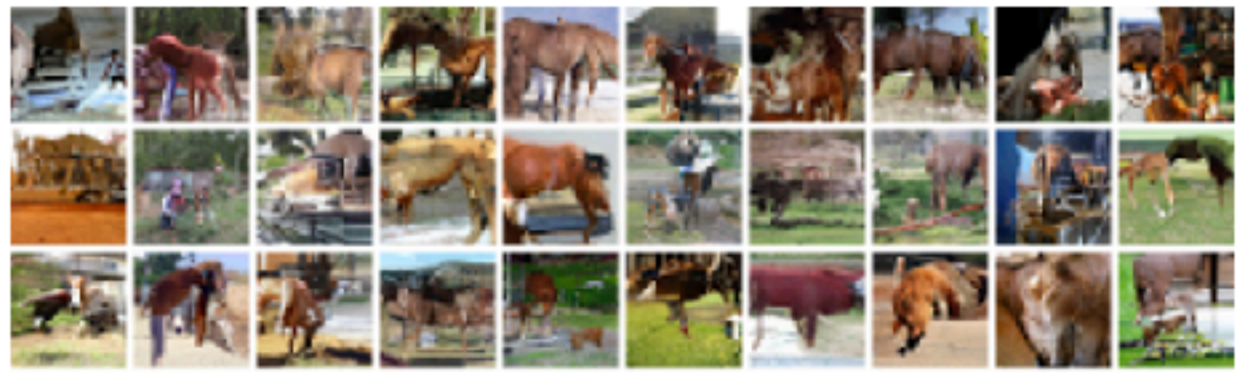
- Input: Raw pixels $\mathbf{x}_{1\ldots t-1}$
- Output: Quantized next color value
  $\mathbf{x}_t \in \{1\ldots256\}$
- Model: $P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \,|\, \mathbf{x}_{1\ldots t-1})$
- Conditioned model:
  $P(\mathbf{x}\,|\,\mathbf{h}) = \prod_{t=1}^{T} P(x_t \,|\, \mathbf{x}_{1\ldots t-1} \,|\, \mathbf{h})$




African elephant


Coral Reef


Sandbar


Sorrel horse

[1] Conditional Image Generation with PixelCNN Decoders. Aaron van den Oord, et al. 2016

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

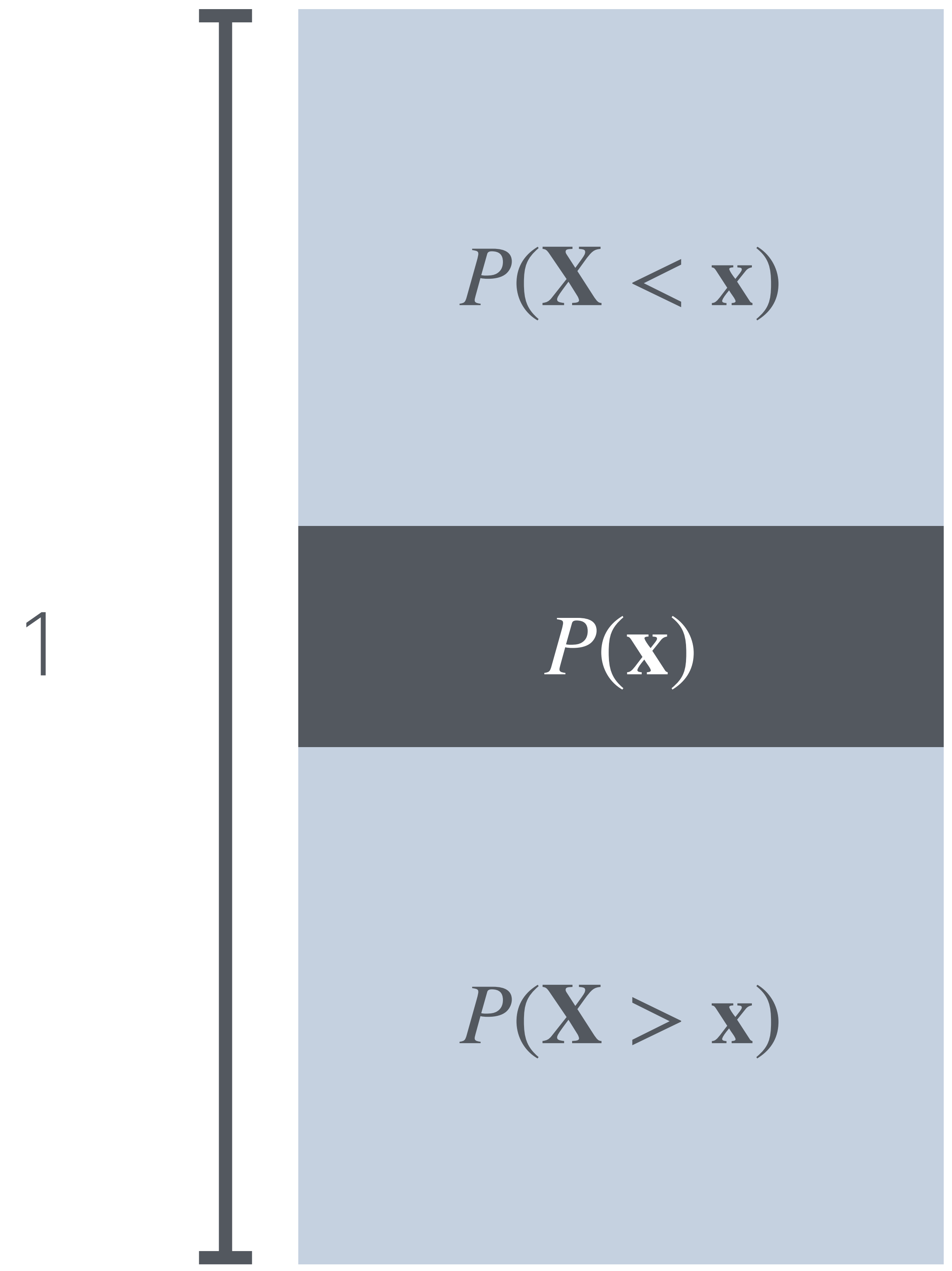
- Knowing $P(\mathbf{x})$ leads to best lossless compression within one bit
    - #bits $= \lfloor -\log_2 P(\mathbf{x}) \rfloor + 1$
- Why?

[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 $P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t | \mathbf{x}_{1...t-1})$
  - $P(\mathbf{X} \leq \mathbf{x}) = \prod_{t=1}^{T} P(X_t \leq x_t | \mathbf{x}_{1...t-1})$
  - Leads to adaptive arithmetic coding

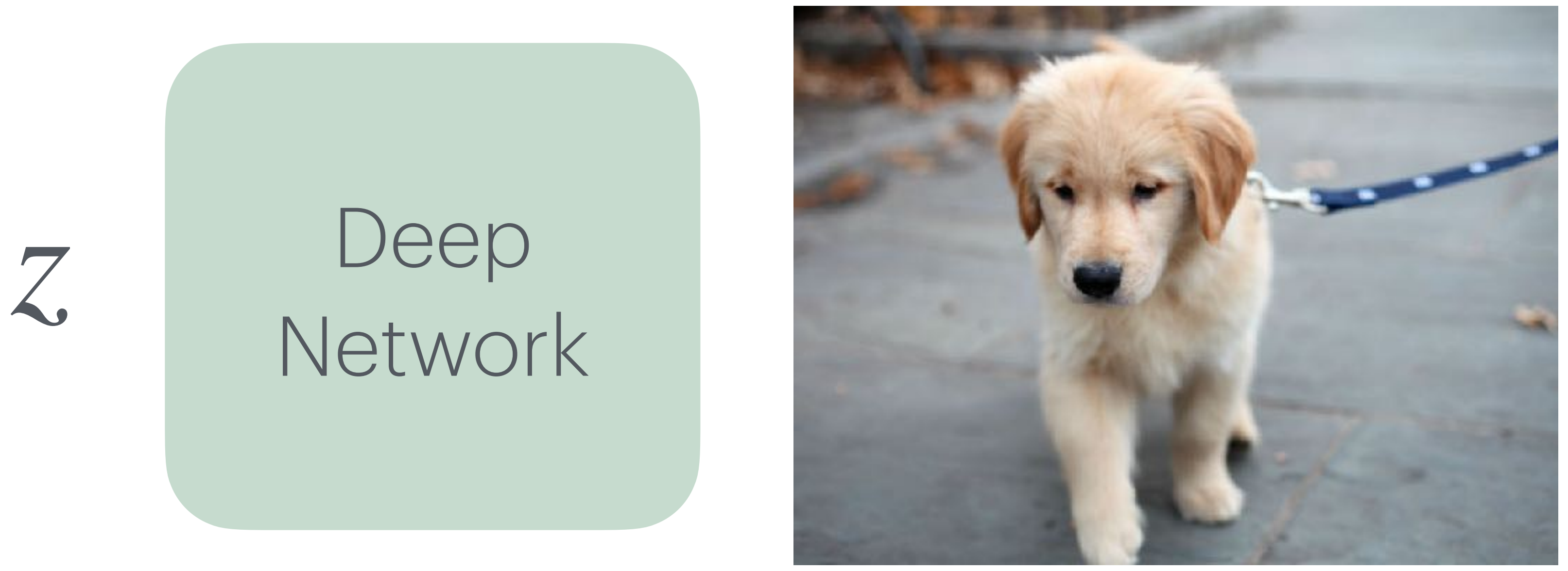
[1] Lossless Image Compression through Super-Resolution. Sheng Cao, et al. 2020
[2] Practical Full Resolution Learned Lossless Image Compression. Fabian Mentzer, et al. 2019

Generative models

Two kinds of models

Sampling based $x \sim P(X)$

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



Density estimation based $P(X)$

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

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