# CS-503 Visual Intelligence: Machines and Minds

Roman Bachmann

04.03.2025

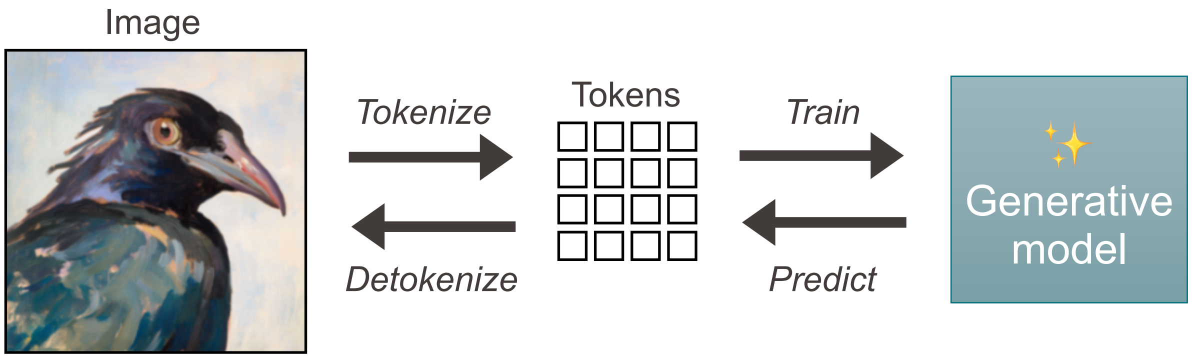
École polytechnique fédérale de Lausanne

Bachmann

# Image Tokenization Crash Course

# Image Tokenization

**Goal**: Project images into a sequence of tokens to model with a generative model



# Image Tokenization





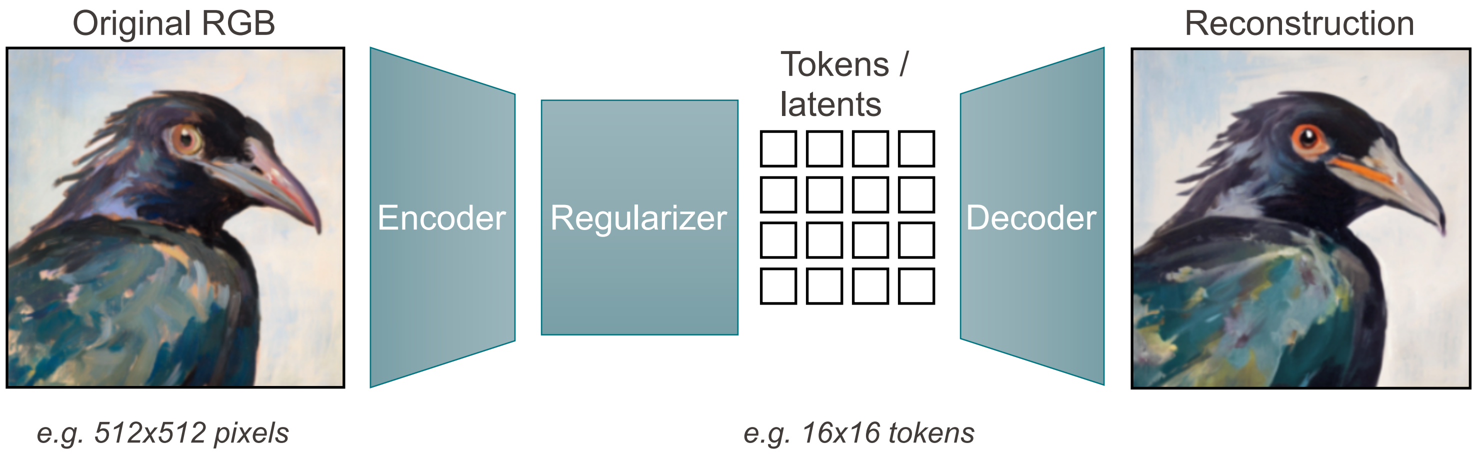
**Goal**: Project images into a sequence of tokens to model with a generative model

- **Why tokenize**:
  - Reduce sequence length
  - Abstract away imperceptible details (lossy compression)
- **Modeling-dependent properties of tokens**:
  - Regularized latent space
  - For autoregressive models: Provide a prediction target that can be sampled from
  - Semantic latent space
  - Ordering
  - ...

# Image Tokenization

## Sequence length reduction

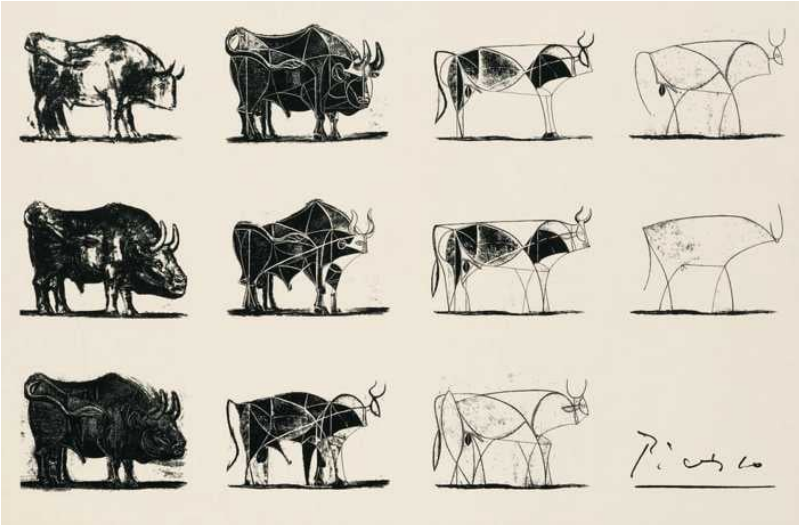
- **Before tokenization:**
  512 x 512 image: 512*512 = 262'144 tokens
- **After tokenization**:
  Downsample with patch size 16 x 16: (512/16)*(512/16) = 1'024 tokens



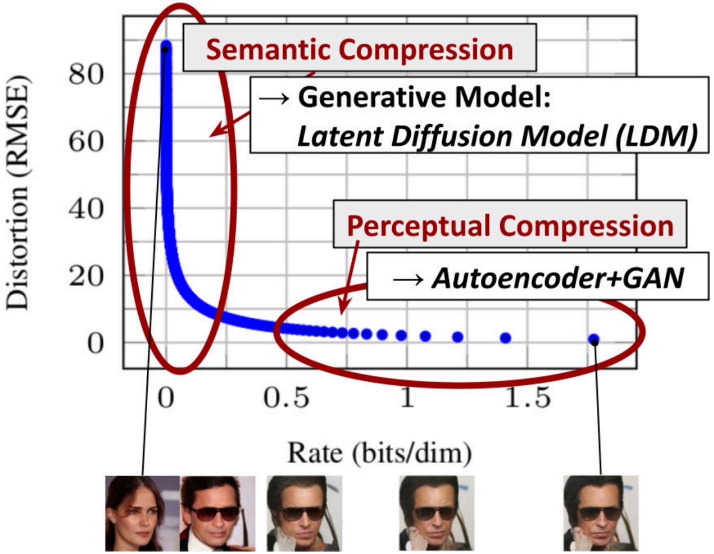
# Image Tokenization

## Abstract away fine-grained (imperceptible) details

- We want to spend model capacity to predict aspects that matter




[The Bull, Pablo Picasso]




[Rombach 2021, Ho 2020]

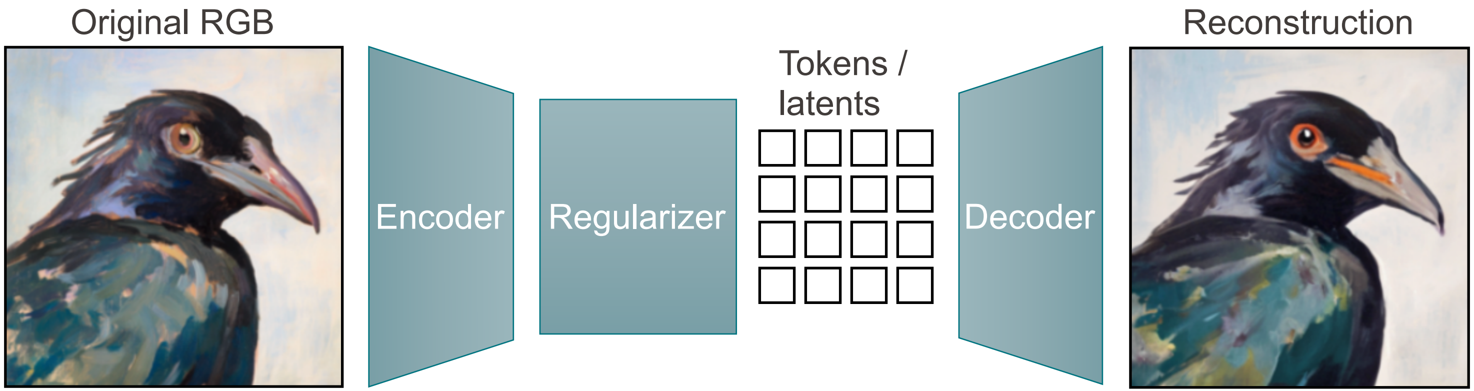
# How to train an image tokenizer?





## Overview

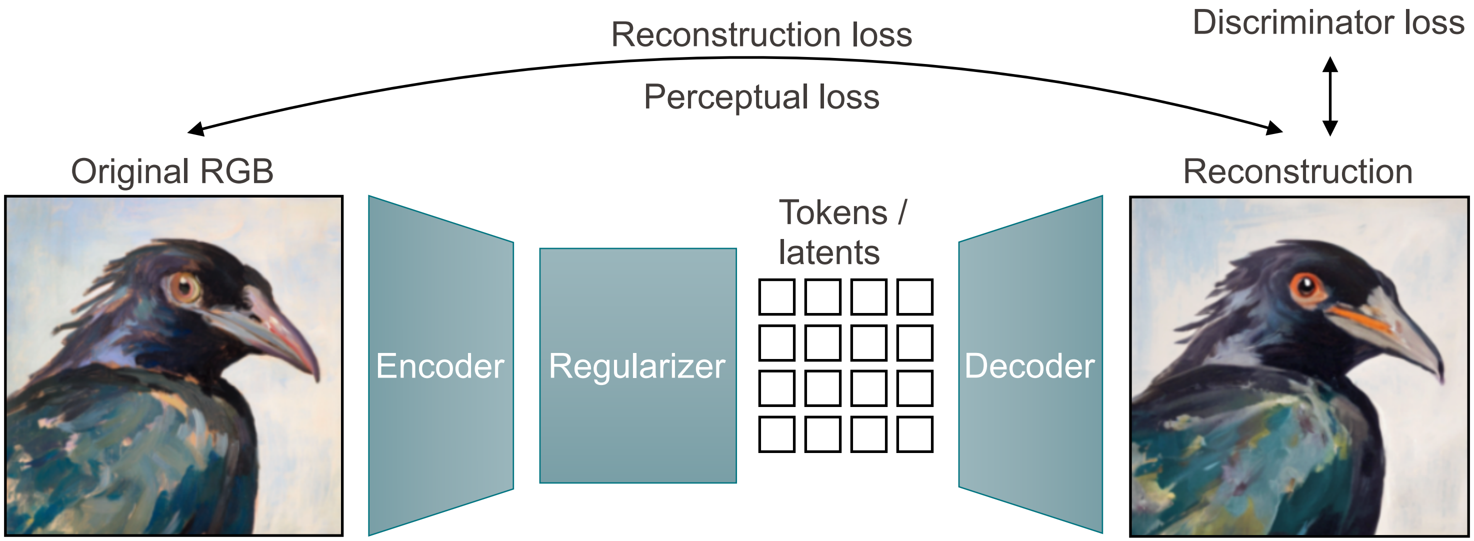
- **Architecture**: Bottleneck autoencoder
- **Bottleneck**: Discrete or continuous regularization
- **Objective**: Mostly autoencoding (reconstruction)



# How to train an image tokenizer?

## Objective

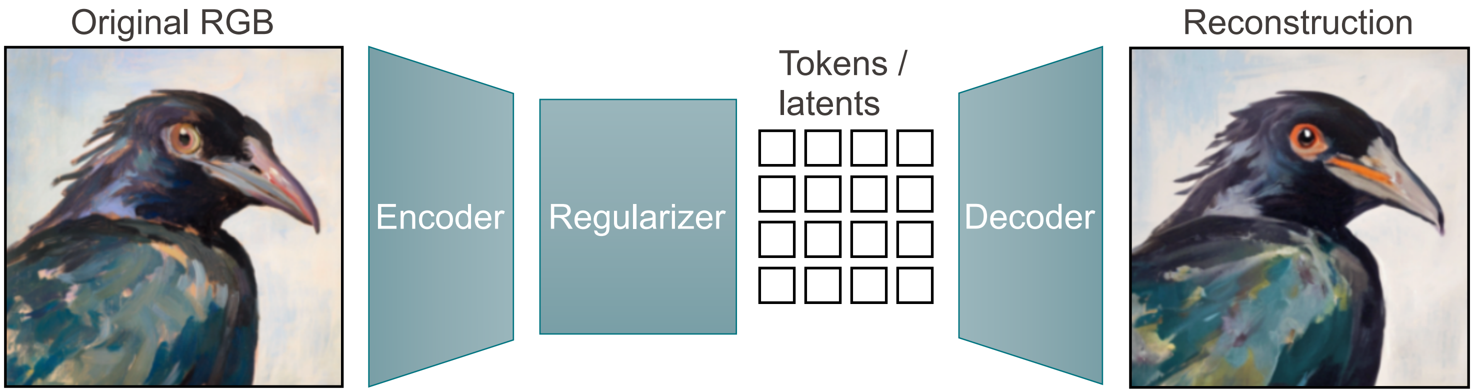
- **Main objective**: Autoencoding (i.e. reconstruction loss)
- **Auxiliary objectives**: Perceptual loss, Discriminator loss, etc...



# How to train an image tokenizer?

## Regularizer / Bottleneck

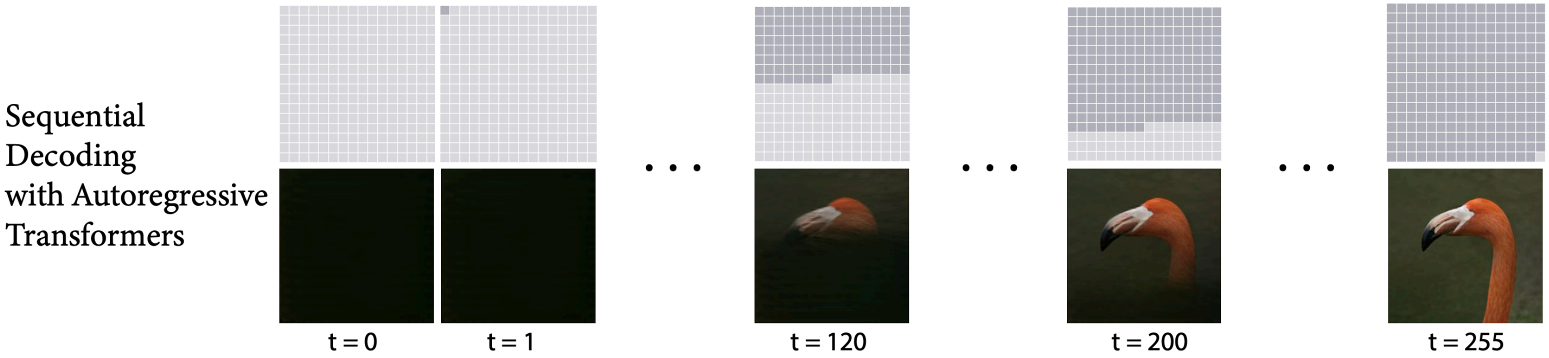
- **Discrete:**
  - Each token can be one of K classes. Commonly K = 4k, 16k, 64k, ...
  - Train with a discrete bottleneck (e.g. FSQ, vector quantization, ...)
- **Continuous:**
  - Each token is a d-dimensional continuous latent. Commonly d = 4, 8, 16, ...
  - Train with KL-regularizer to keep latent space well-behaved



# Autoregressive image modeling

## Parti, LlamaGen, ...

- Train a "GPT" on discrete image tokens

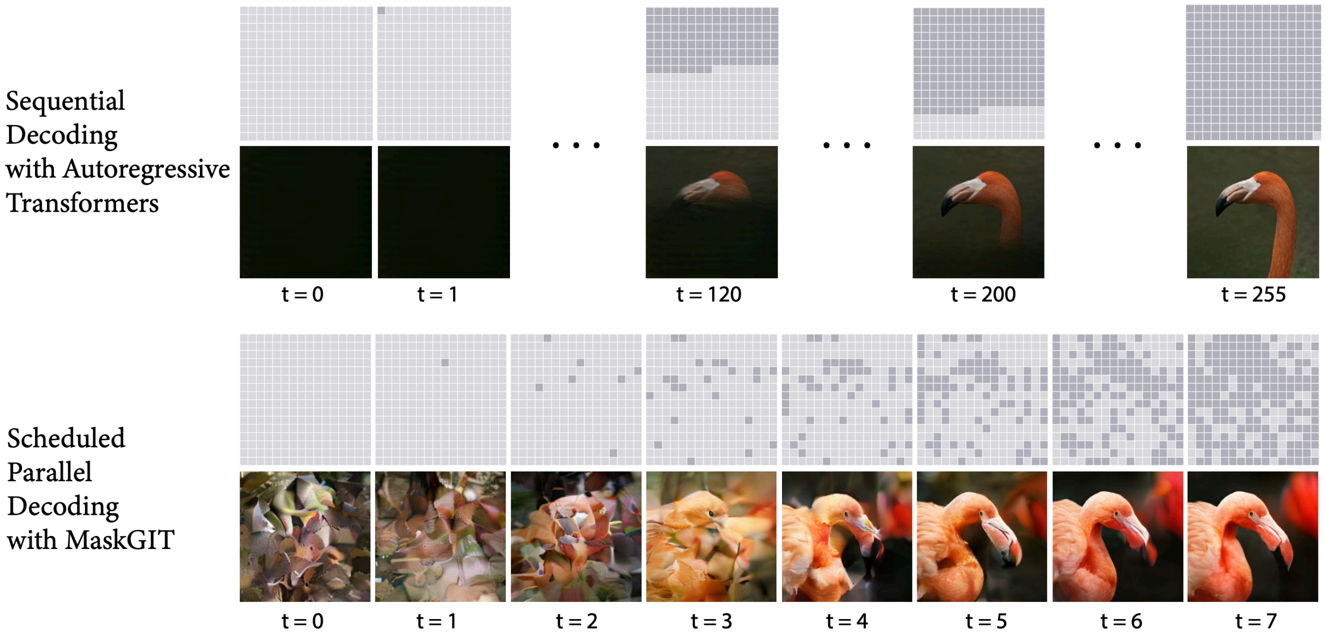



Credit: MaskGIT: Masked Generative Image Transformer, Chang et al. 2022

# Discrete diffusion / Parallel decoding

## MaskGIT, Muse, MAGE, …

- Discrete token-based masked image models are fast generative models

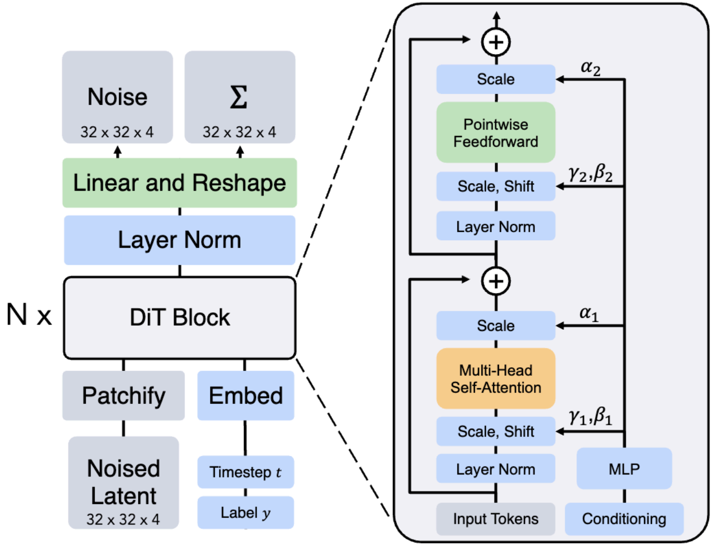



Credit: MaskGIT: Masked Generative Image Transformer, Chang et al. 2022

# Diffusion / Flow Matching

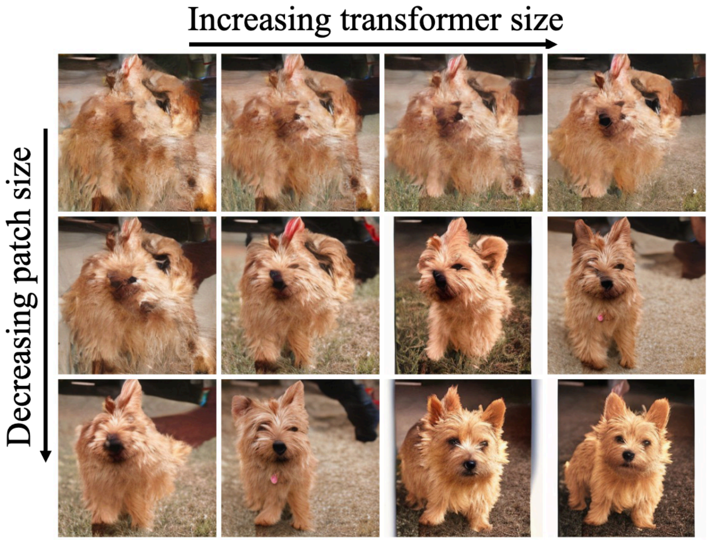
## DiT, SiT, ...

- Train a diffusion / flow Transformer on continuous latents




Latent Diffusion Transformer

DiT Block with adaLN-Zero




Credit: Scalable Diffusion Models with Transformers, Peebles et al. 2022

# Hot topics in tokenization

## Some pointers

- Autoregressive modeling with continuous tokens (instead of discrete)
  [GIVT, MAR, Fluid, ...]
- Next-scale prediction
  [VAR, Infinity, ...]
- Flexible-length tokenization
  [FlexTok, ALIT, ElasticTok, ...]
- Jointly train tokenizer with an AR model
  [JetFormer]

# Enjoy the Course!

Amir Zamir (amir.zamir@epfl.ch)

Rishubh Singh (rishubh.singh@epfl.ch head TA)
Zhitong Gao (zhitong.gao@epfl.ch)
Roman Bachmann (roman.bachmann@epfl.ch)

https://vilab.epfl.ch/