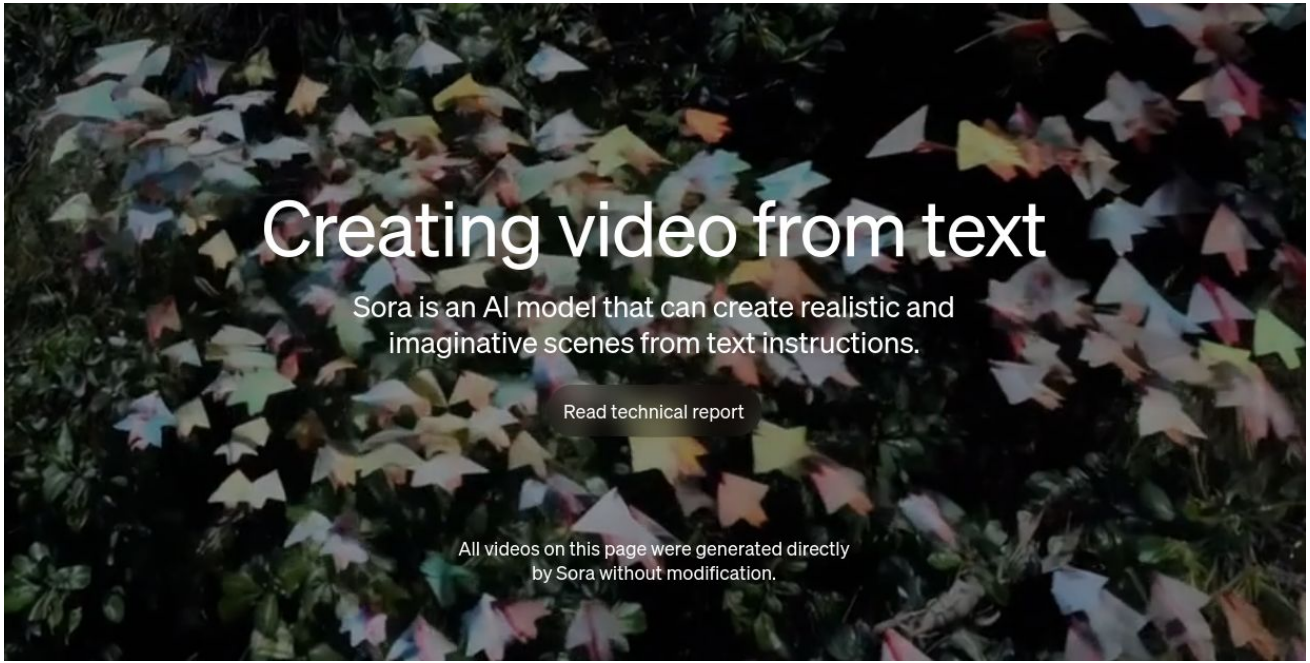
Modelos Generativos en Computer Vision

Eugenio Herrera-Berg
Machine Learning Engineer - CENIA
eugenio.herrera@cenia.cl

“An expressive oil painting of a basketball player dunking, depicted as an explosion of a nebula.” - **Dalle 3**

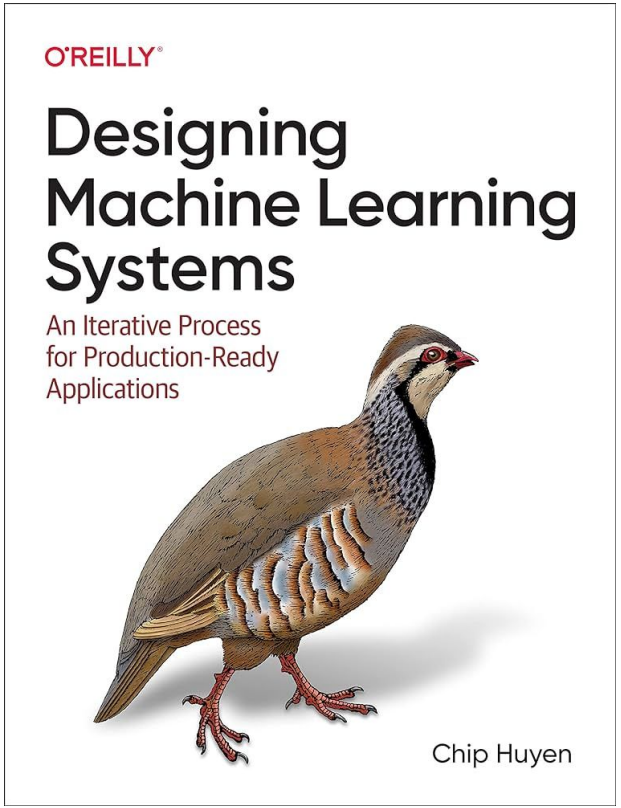

Creating video from text
Sora is an AI model that can create realistic and imaginative scenes from text instructions.
Read technical report
All videos on this page were generated directly by Sora without modification.

Recomendación práctica

Si tu objetivo con el diplomado es aprender a desarrollar proyectos de ML en industria, te recomiendo este libro :)

- Link a amazon.

- Link al repo oficial.



Objetivos de esta clase

- Comprender qué son los modelos generativos.

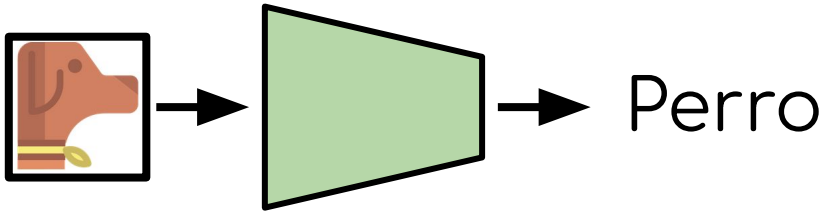
- Familiarizarse con distintas estrategias utilizadas para generar imágenes.

- Ver sus utilidades en la industria.

¿Qué es un Modelo Generativo?

Modelos Clasificadores

Reconocen si algo pertenece a una distribución.



Modelos Clasificadores

Reconocen si algo pertenece a una distribución.

$$f(x) \mid x \in X$$

Modelos Generativos

Aprenden a representar una distribución.

$$f(\cdot) \mid f(\cdot) \approx X$$

Modelos Generativos

Aprenden a representar una distribución.

$$f(x) \longrightarrow$$

Modelos Generativos en CV

- Autoencoders / Variational Autoencoders
- Generative Adversarial Networks
- Modelos de difusión
- Latent Diffusion

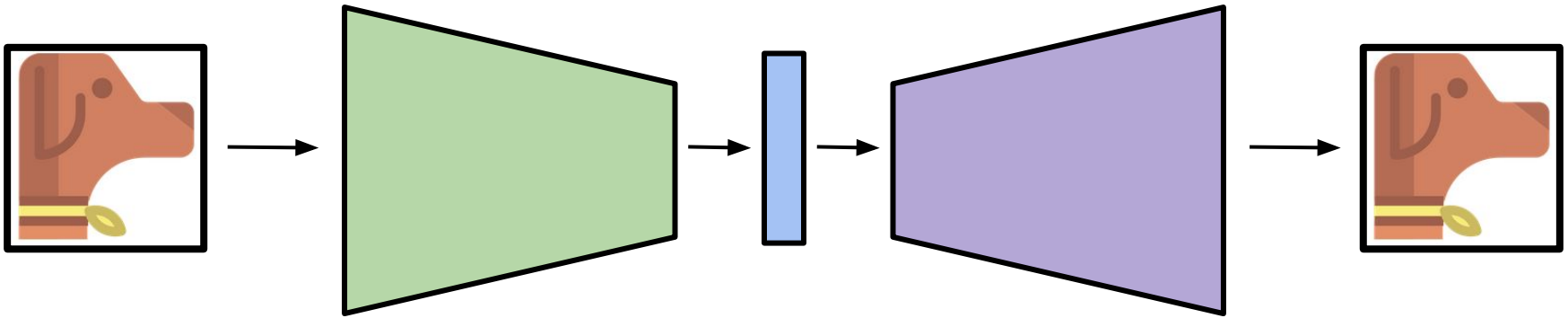
Modelos Generativos en CV

- Autoencoders / Variational Autoencoders
- Generative Adversarial Networks
- Modelos de difusión
- Stable Diffusion

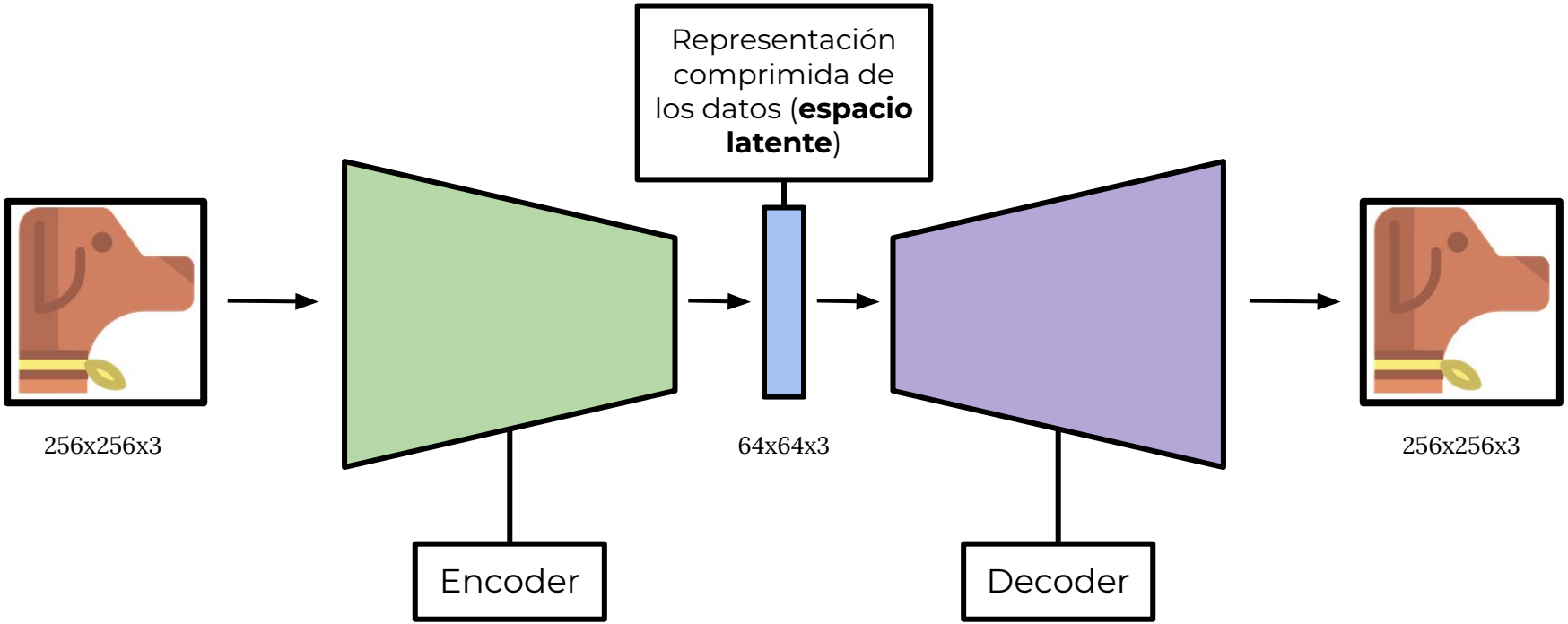
Autoencoders

Arquitectura de redes neuronales donde **se comprime y luego descomprime información**, **generando así una representación intermedia y comprimida de esta (*espacio latente*)**.

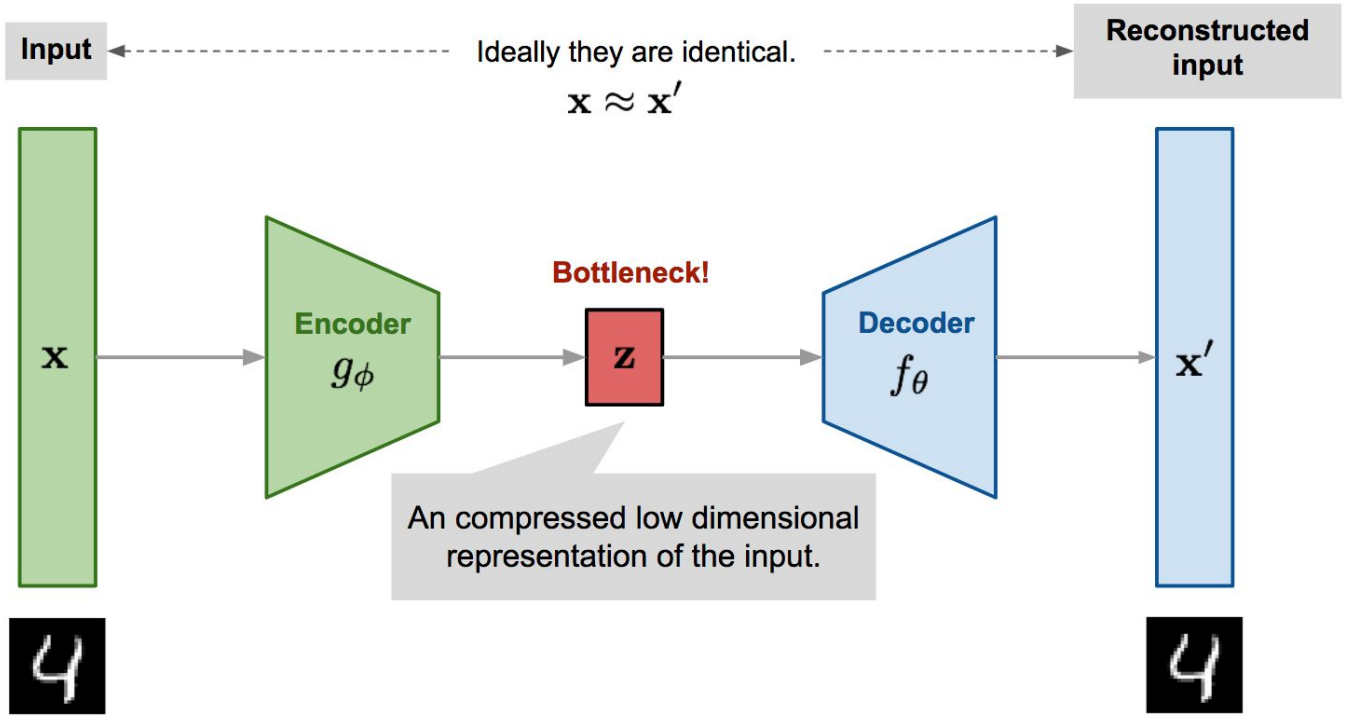
Autoencoders



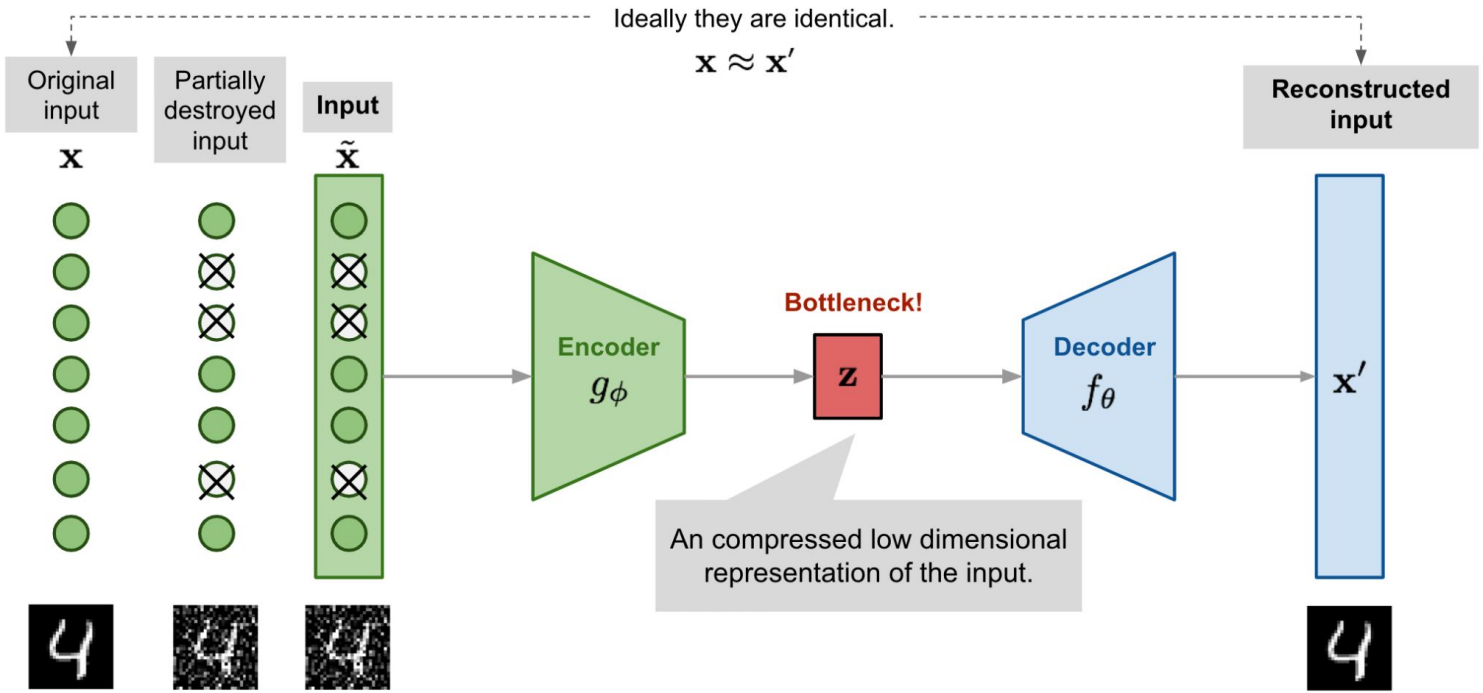
Autoencoders



Autoencoders



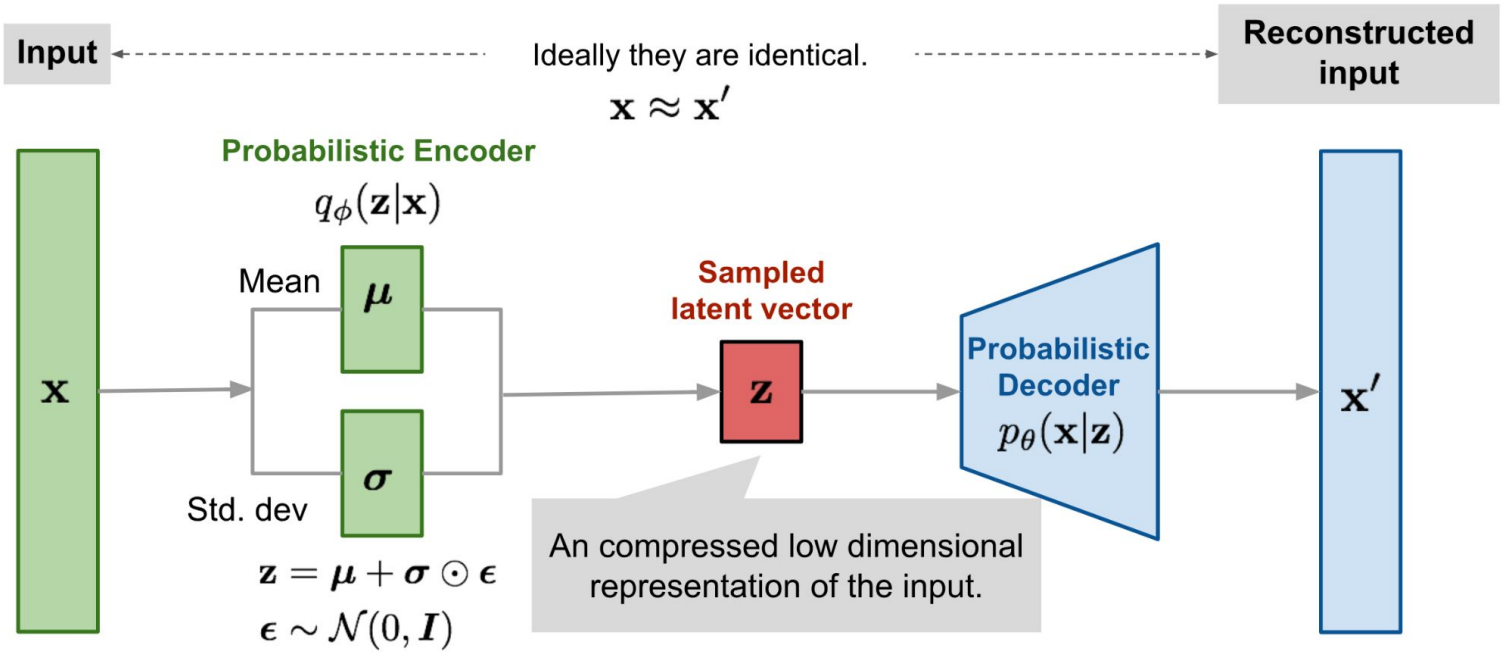
Autoencoders



Variational Autoencoders (VAEs)

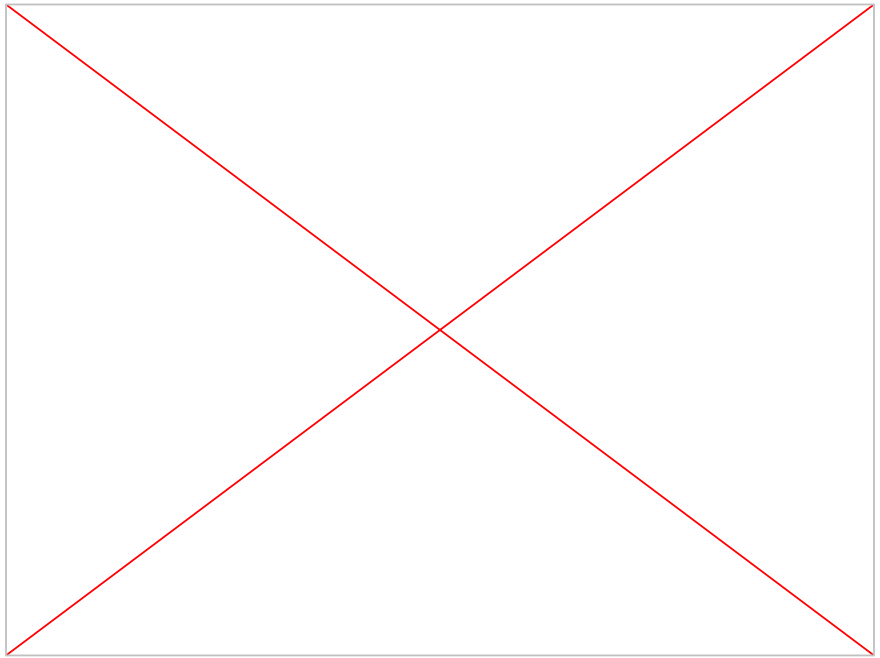
Reformulación de la arquitectura de los autoencoders, **que permite convertirlos en modelos generativos.**

Variational Autoencoders (VAEs)



Espacio latente contínuo (permite interpolación contínua/consistente)

Espacio latente contínuo



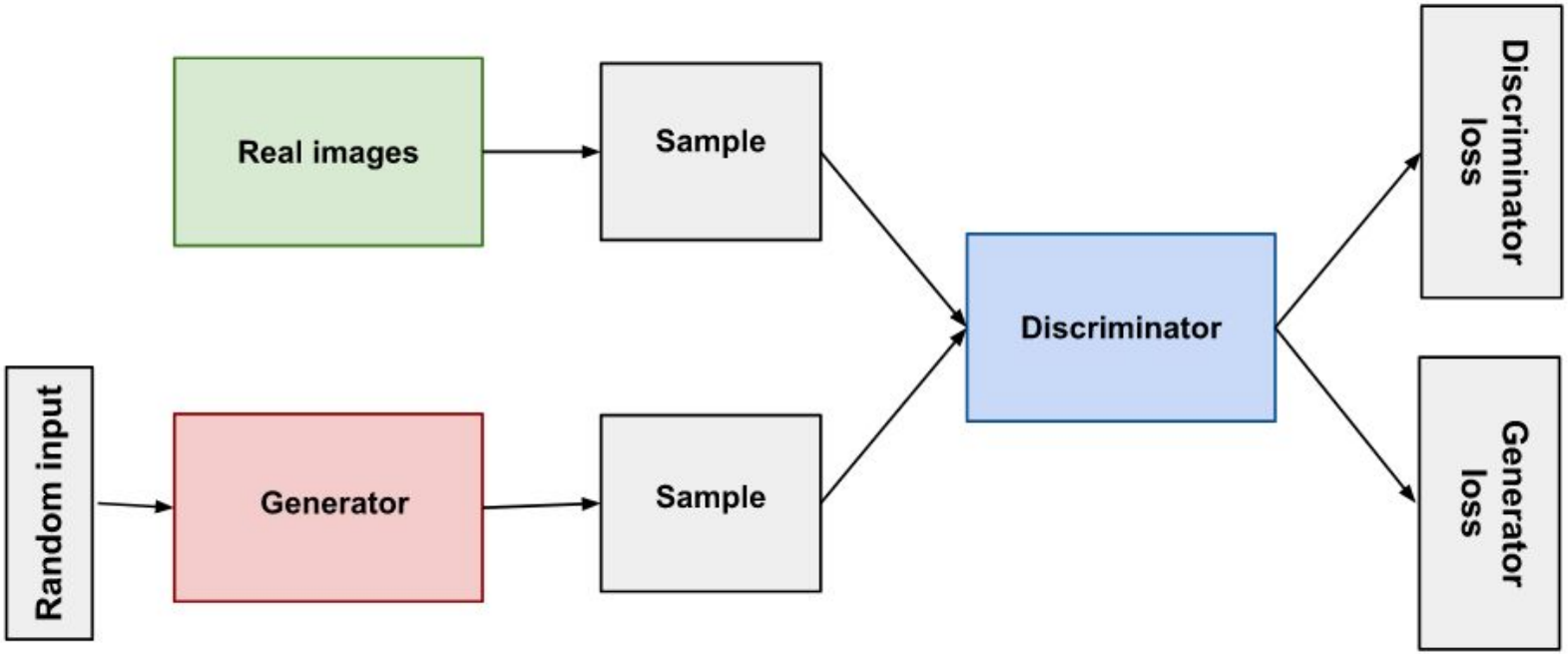
Modelos Generativos en CV

- Autoencoders / Variational Autoencoders
- Generative Adversarial Networks
- Modelos de difusión
- Stable Diffusion

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs)

Arquitectura de modelo generativo donde **se entrenan en conjunto un generador y un clasificador, mediante una competencia entre ambos**.


Real images
Sample
Random input
Generator
Sample
Discriminator
Discriminator loss
Generator loss

Loss

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

z: Ruido / Vector Latente (modelado por una normal)

x : Datos reales

G : Generador

D: Discriminador

G(z): Datos sintéticos

D(x): Discriminador evaluando datos reales

D(G(z)): Discriminador evaluando datos sintéticos

Modelos Generativos en CV

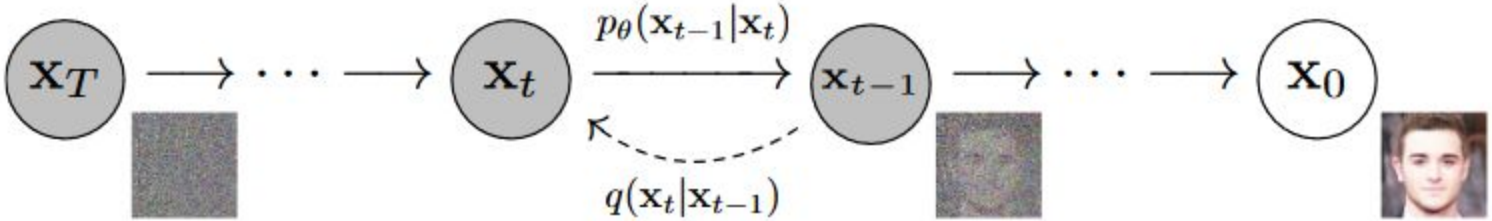
- Autoencoders / Variational Autoencoders
- Generative Adversarial Networks
- Modelos de difusión
- Stable Diffusion

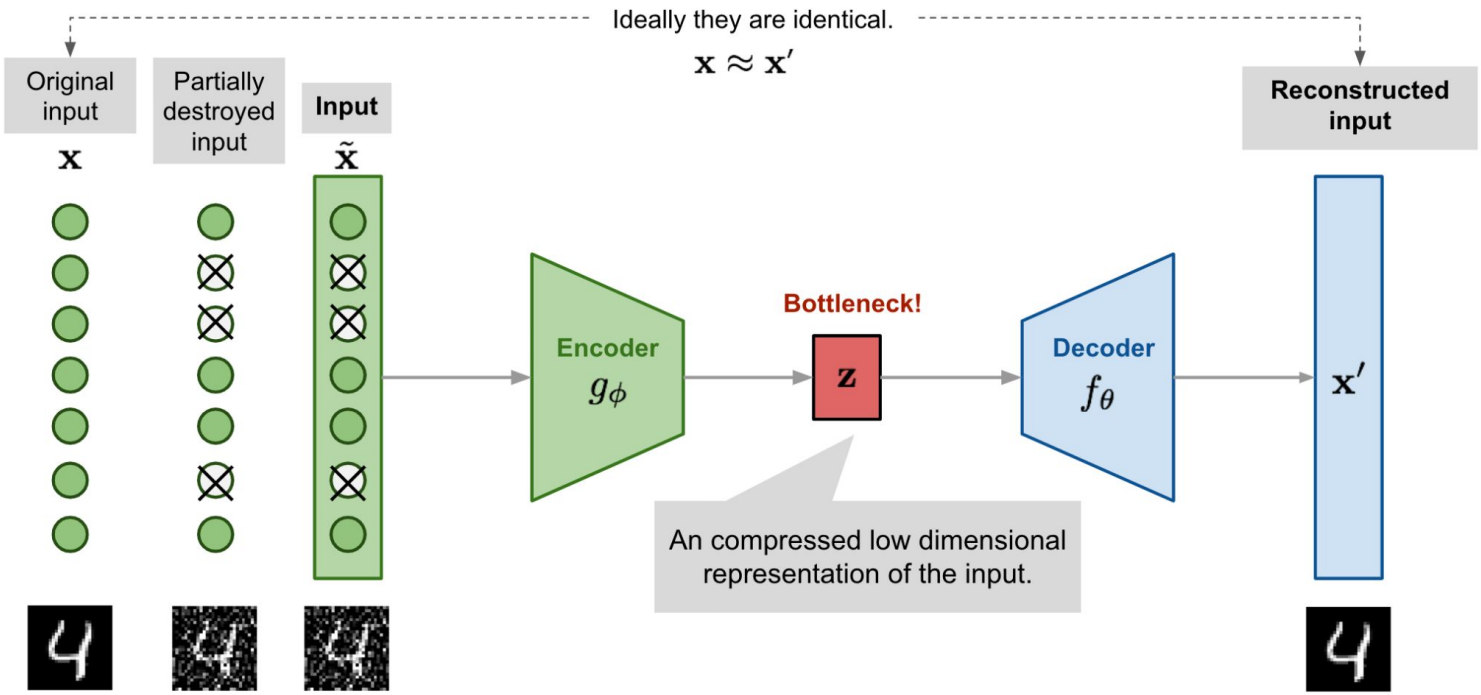
Modelos de Difusión

Modelo generativo que se construye diseñando un **procedimiento para gradualmente convertir datos en ruido, y luego entrenar una red neuronal que aprenda a invertir este proceso paso a paso**.

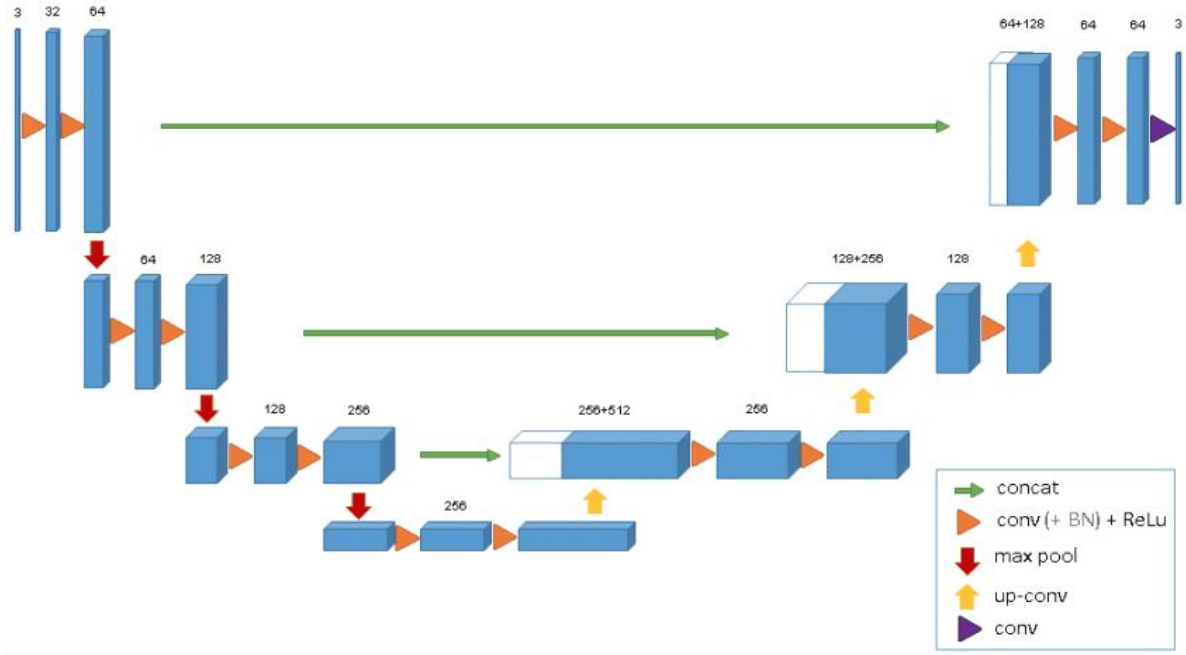
Diffusion models:
Gradually add Gaussian noise and then reverse

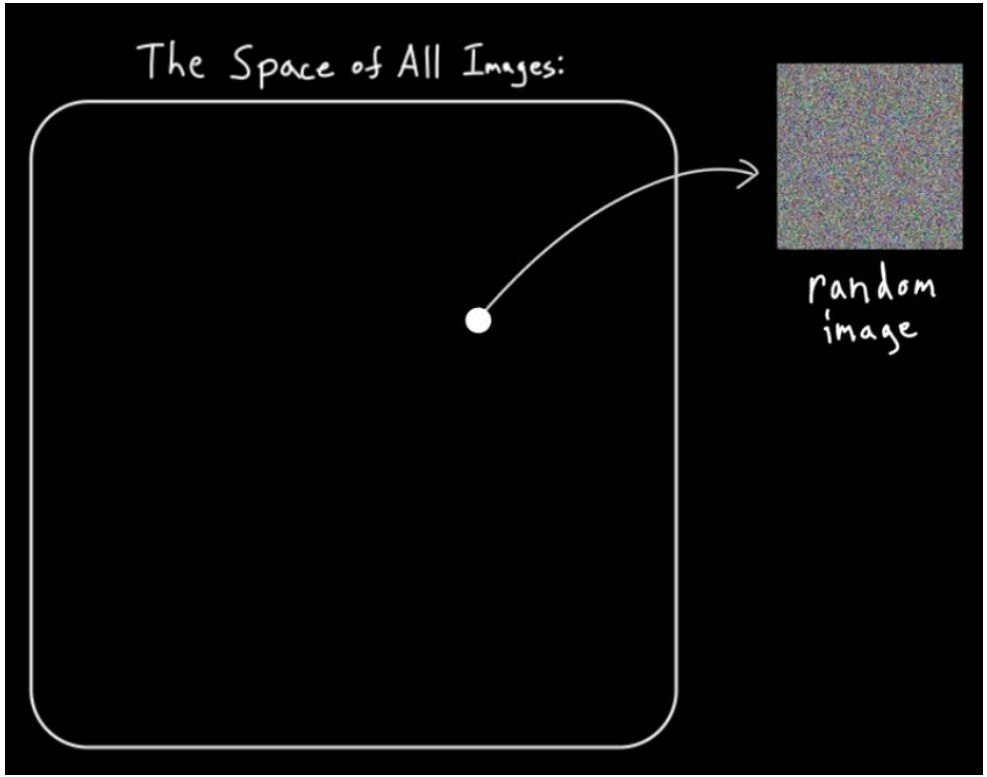
Modelos de Difusión

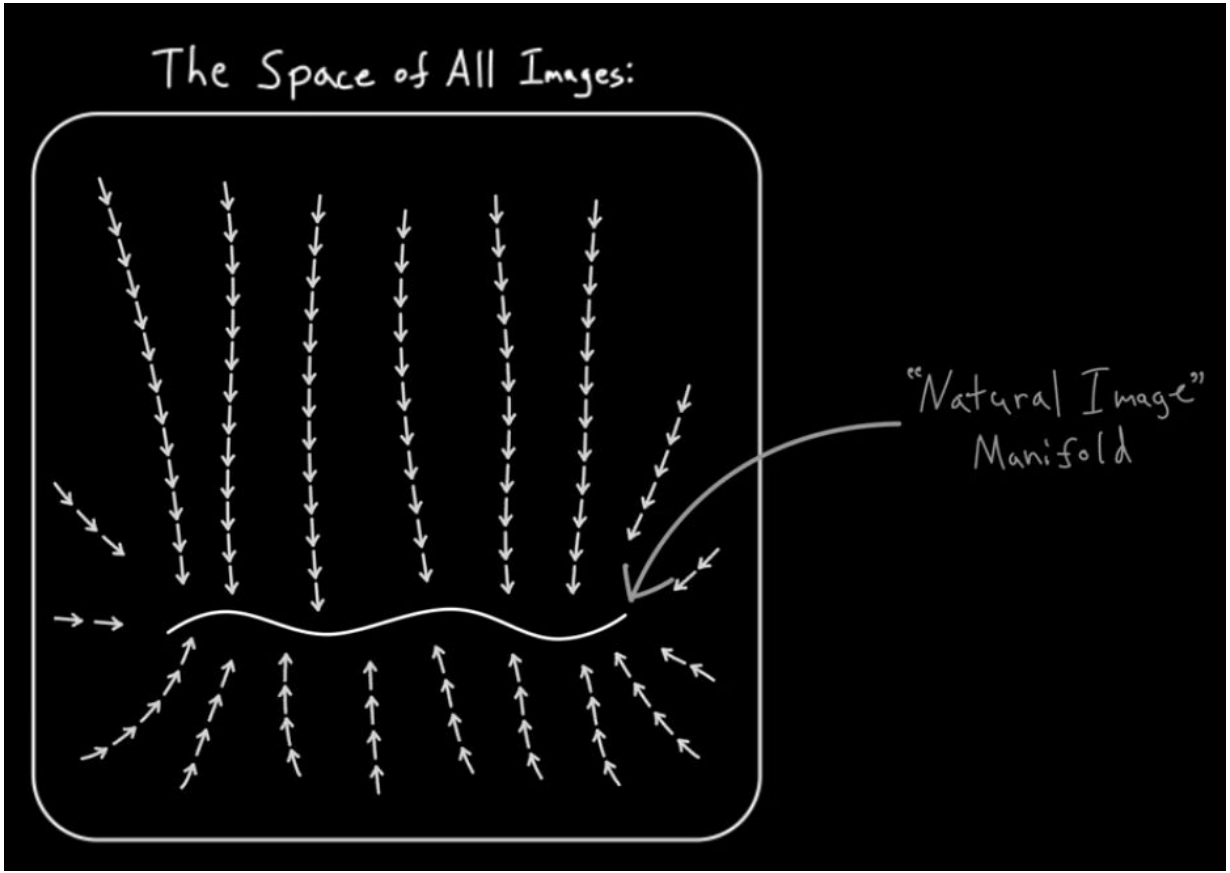



Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Original input
$\mathbf{x}$
Partially destroyed input
Input
$\tilde{\mathbf{x}}$
Encoder
g_ϕ
Bottleneck!
$\mathbf{z}$
Decoder
f_θ
$\mathbf{x}'$
Reconstructed input
An compressed low dimensional representation of the input.

UNet

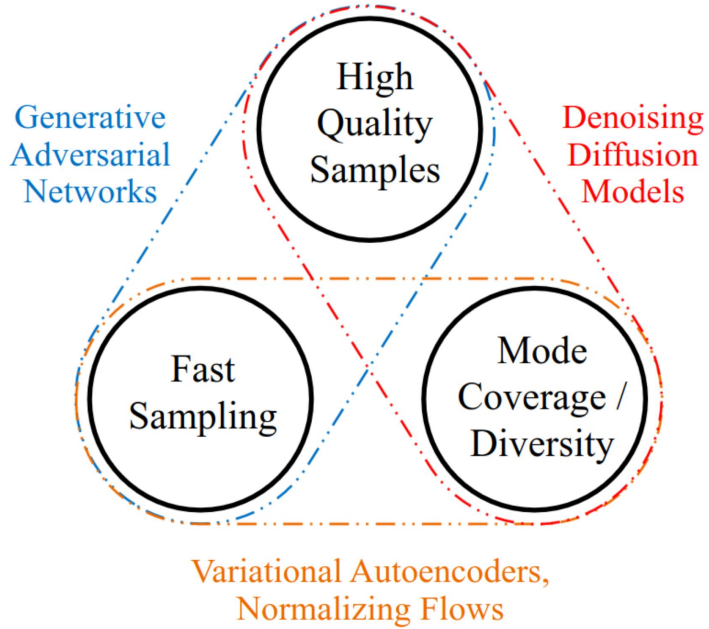

3
32
64
64
128
128
256
256
256+512
256
128+256
128
64+128
64
64
3
concat
conv (+ BN) + ReLu
max pool
up-conv
conv


The Space of All Images:
random image


The Space of All Images:
"Natural Image" Manifold

Comparando Modelos Generativos

Generative learning trilemma



Fréchet inception distance (FID)

Métrica para comparar la calidad de las imágenes creadas por un modelo generativo.

Compara la distribución de las imágenes generadas con las del dataset original utilizando representaciones de un modelo auxiliar (VGG x ejemplo).

Comparación - Fréchet inception distance [Imagenet]

Rank	Model	FID ↓	Inception score	Paper	Code	Result	Year	Tags
1	**EDM2- XXL Autoguidance** (M, T /3.5)	1.25		Guiding a Diffusion Model with a Bad Version of Itself			2024	
2	**EDM2- S Autoguidance** (XS, T /16)	1.34		Guiding a Diffusion Model with a Bad Version of Itself			2024	
3	**EDM2-XXL w/ guidance interval**	1.40		Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models			2024	
4	**EDM2-S w/ guidance interval**	1.68		Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models			2024	
5	**PaGoDA**	1.80		PaGoDA: Progressive Growing of a One-Step Generator from a Low-Resolution Diffusion Teacher			2024	GAN VAE Diffusion
6	**EDM2-XXL**	1.81		Analyzing and Improving the Training Dynamics of Diffusion Models			2023	
7	**EDM2-XL**	1.85		Analyzing and Improving the Training Dynamics of Diffusion Models			2023	
8	**EDM2-L**	1.88		Analyzing and Improving the Training Dynamics of Diffusion Models			2023	
9	**MAGVIT-v2**	1.91	324.3	Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation			2023	
10	**EDM2-M**	2.01		Analyzing and Improving the Training Dynamics of Diffusion Models			2023	

Comparación - Fréchet inception distance [FFHQ]

Rank	Model	FID ↓	FD	Precision	Recall	IS	bits/ dimension	Paper	Code	Result	Year	Tags
1	**StyleSAN-XL**	1.68						SAN: Inducing Metrizability of GAN with Discriminative Normalized Linear Layer			2023	GAN
2	**StyleNAT**	2.05						StyleNAT: Giving Each Head a New Perspective			2022	GAN, Transformer, Neighborhood Attention
3	**StyleGAN-XL**	2.19						StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets			2022	GAN
4	**PDM+CS**	2.57						Compensation Sampling for Improved Convergence in Diffusion Models			2023	
5	**HiT-L**	2.58						Improved Transformer for High-Resolution GANs			2021	Transformer
6	**Poly-INR**	2.72						Polynomial Implicit Neural Representations For Large Diverse Datasets			2023	
7	**StyleSwin**	2.81						StyleSwin: Transformer-based GAN for High-resolution Image Generation			2021	Transformer, Swin-Transformer
8	**HiT-B**	2.95						Improved Transformer for High-Resolution GANs			2021	Transformer
9	**HiT-S**	3.06						Improved Transformer for High-Resolution GANs			2021	Transformer
10	**InsGen**	3.31						Data-Efficient Instance Generation from Instance Discrimination			2021	

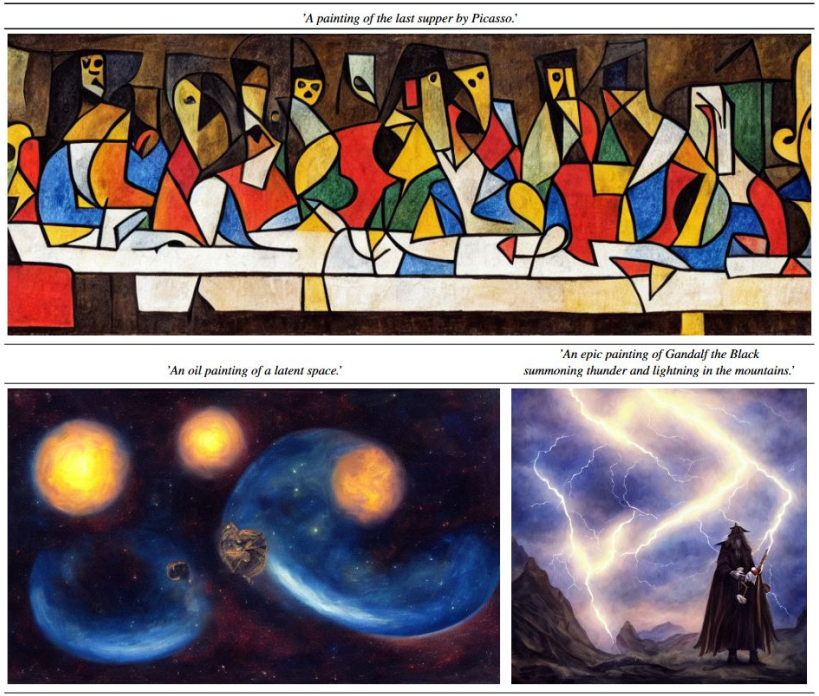
	Velocidad generación imagen	**Calidad imagen**	**Distribución imagenes generadas**
VAEs	Rápida	Baja	Alta
GANs	Rápida	Alta	Baja
Difusión	Lenta	Alta	Alta

Modelos Generativos en CV

- Autoencoders / Variational Autoencoders
- Generative Adversarial Networks
- Modelos de difusión
- Latent Diffusion

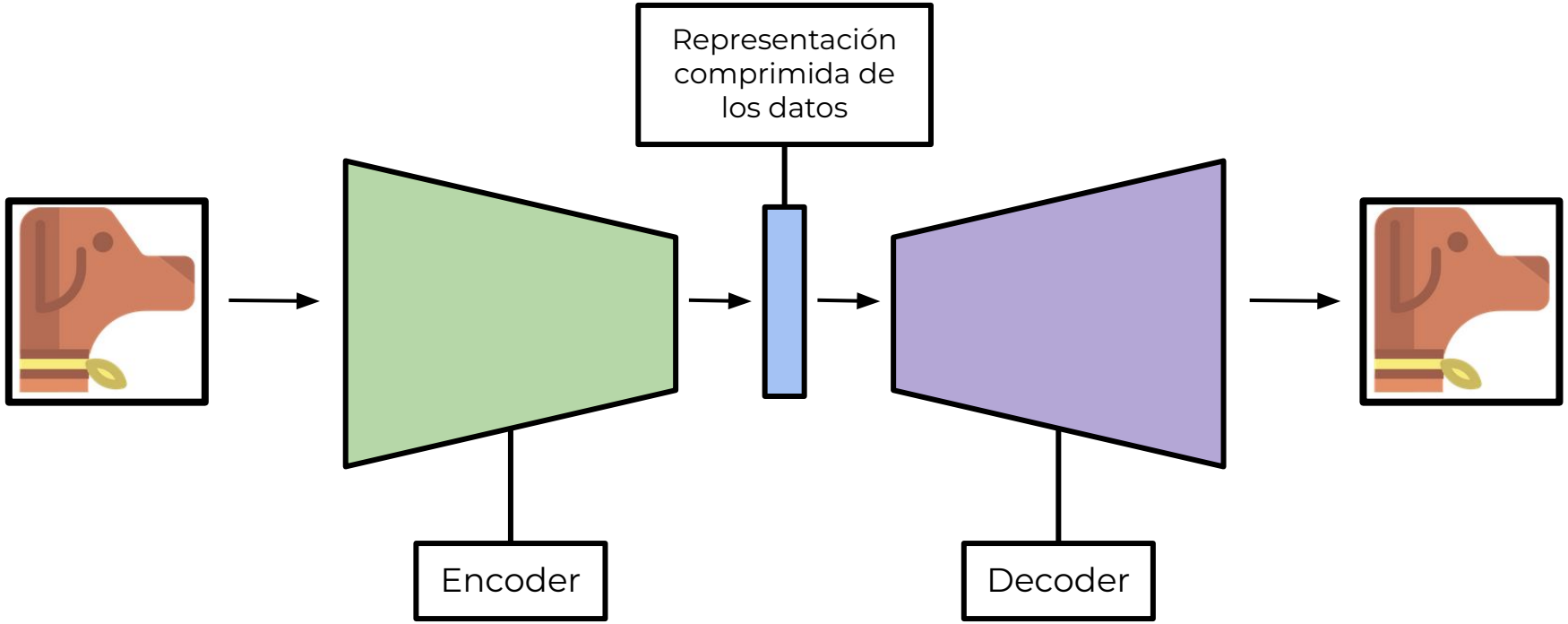
Stable Diffusion

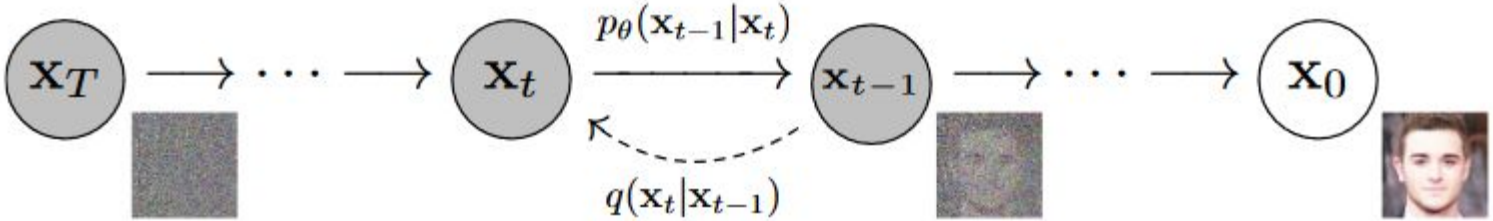
Latent Diffusion (Abril, 2022)

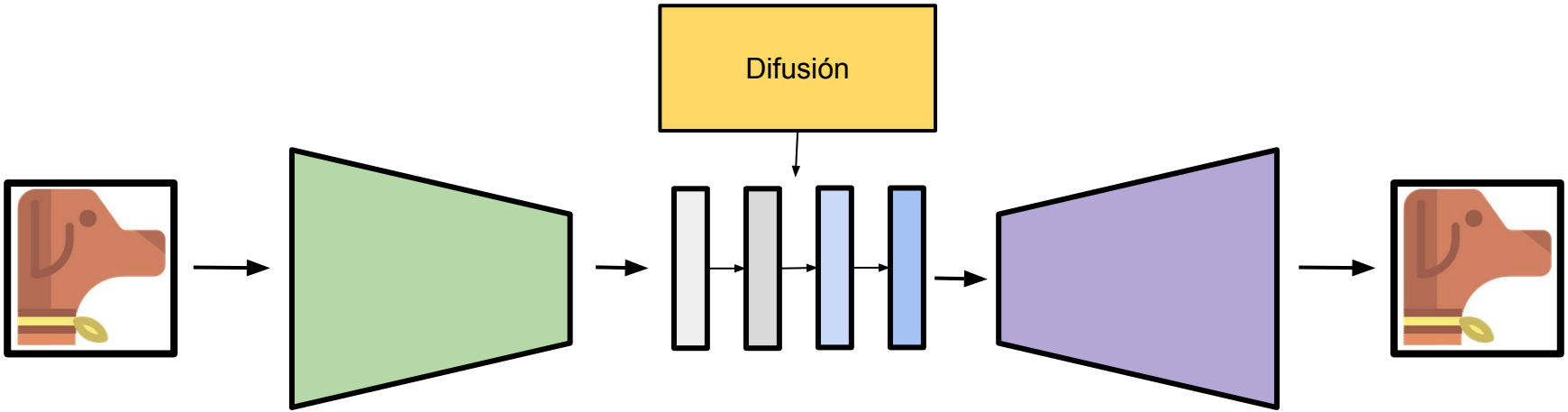


Stable Diffusion (Latent Diffusion)

Arquitectura generativa compuesta, que entrena un **modelo de difusión en el espacio latente de un autoencoder** que fue **pre-entrenado mediante el esquema competitivo de una GAN**.

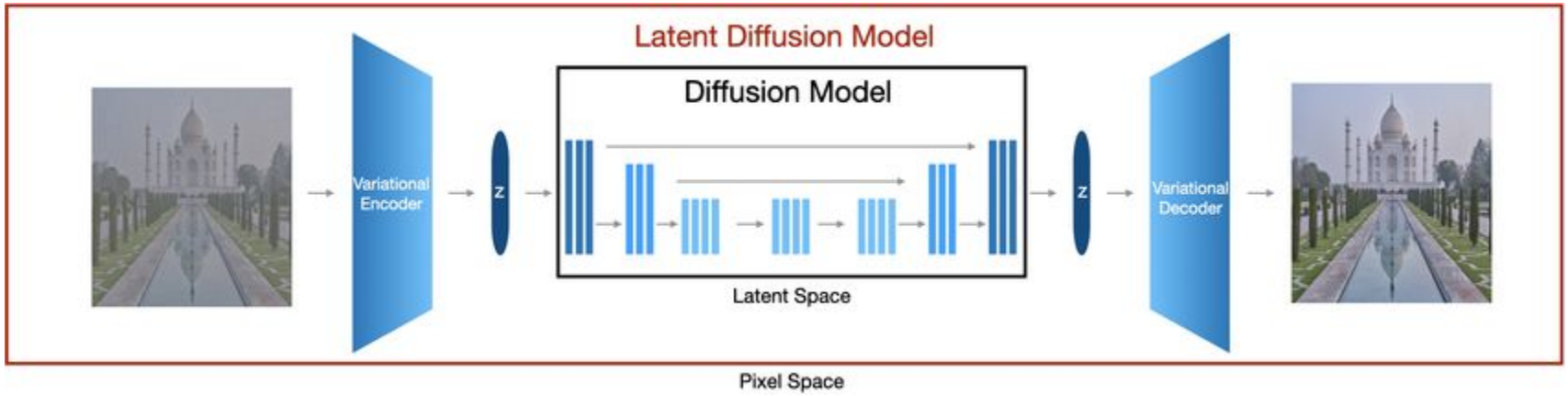

Representación comprimida de los datos
Encoder
Decoder

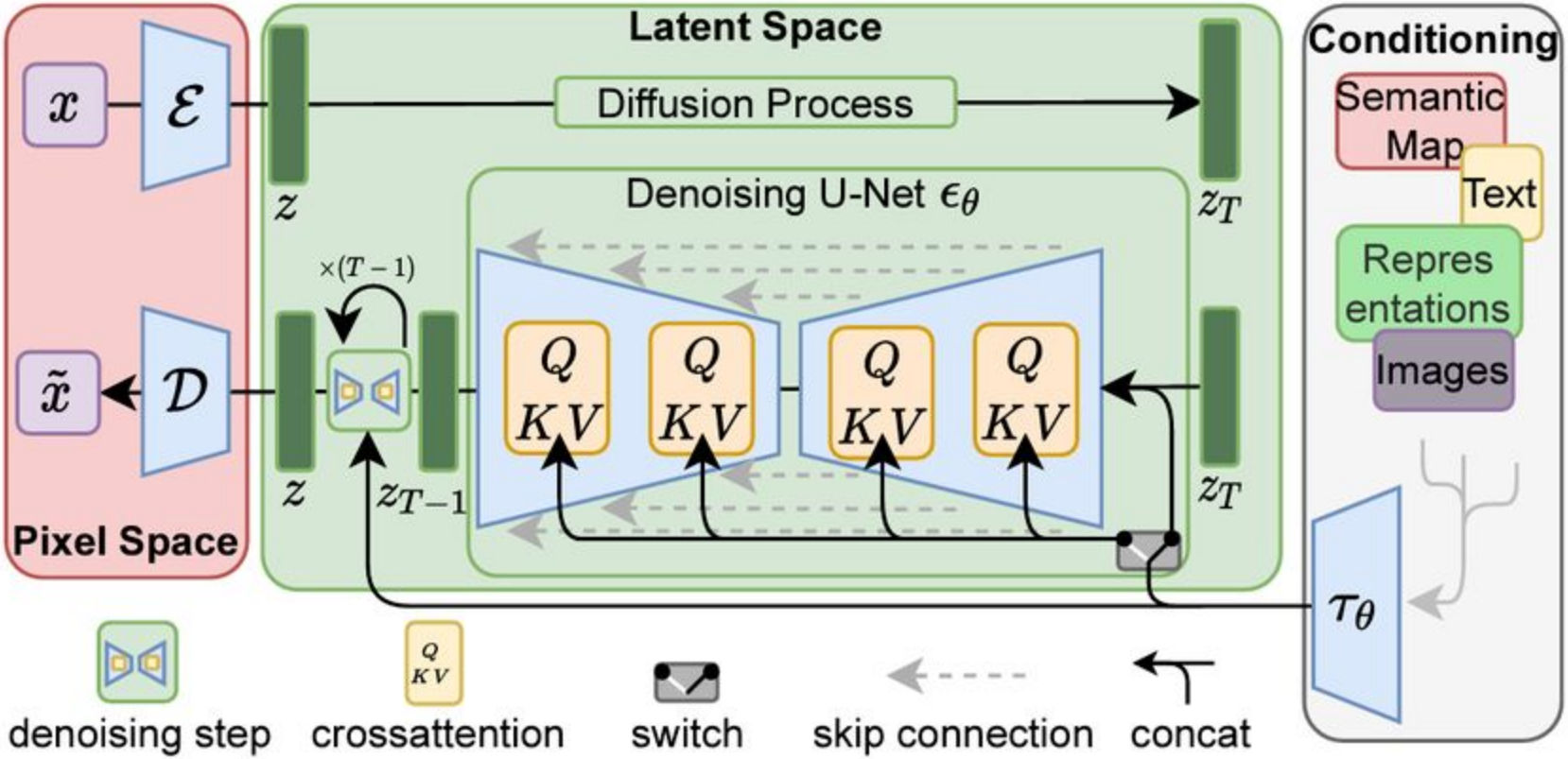

$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$

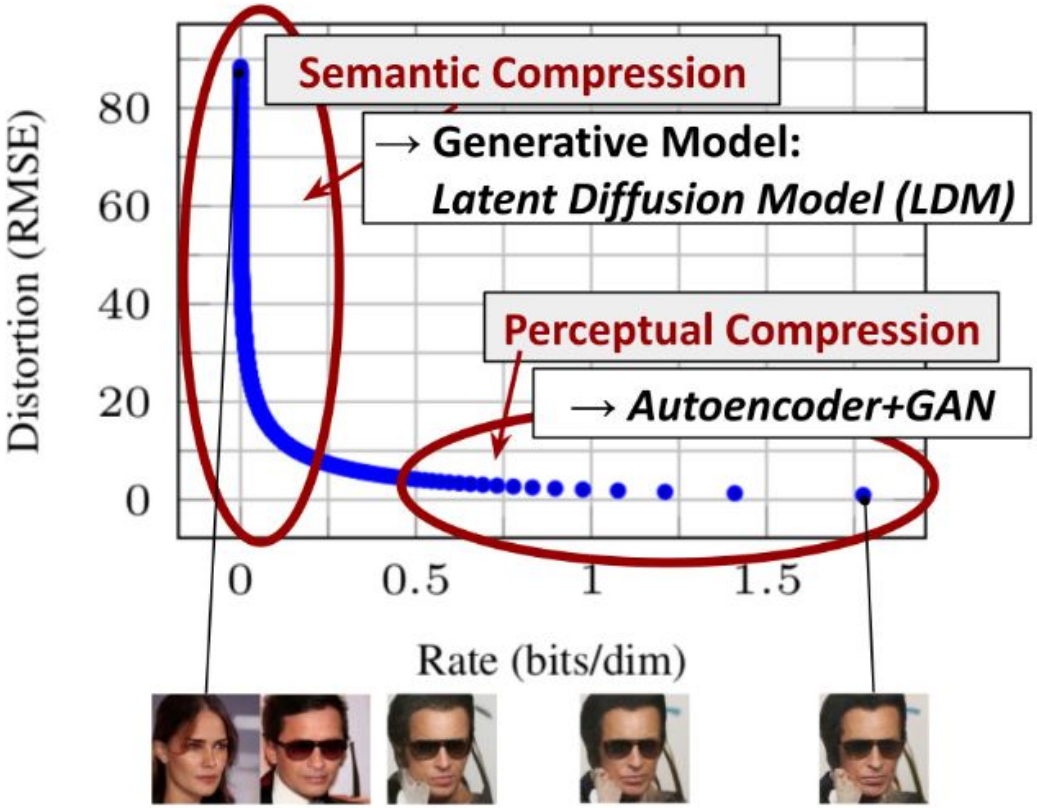


El proceso de difusión ocurre en el espacio intermedio del autoencoder

Latent Diffusion




Latent Space
Diffusion Process
Denoising U-Net ϵ_θ
$\times(T-1)$
x
$\mathcal{E}$
z
z_T
$\tilde{x}$
$\mathcal{D}$
z_{T-1}
Q
KV
Pixel Space
Conditioning
Semantic Map
Text
Repres entations
Images
τ_θ
denoising step
crossattention
switch
skip connection
concat


Semantic Compression
→ Generative Model:
Latent Diffusion Model (LDM)
Perceptual Compression
→ Autoencoder+GAN
Distortion (RMSE)
80
60
40
20
0
0
0.5
1
1.5
Rate (bits/dim)

Ventajas de Difusión Latente

- **“Soluciona” el trilema de los modelos generativos.**

- **Permite entrenar modelos de difusión** competentes **con bajos recursos**.

- **Genera imágenes de cualquier dimensión**, sin necesidad de reentrenamiento.

- Desacopla la lógica de condicionamiento, por lo que **puede ser ajustado para nuevas tareas, como generación de imágenes a partir de texto, repintar secciones de imágenes**, etc.

Usos en la industria

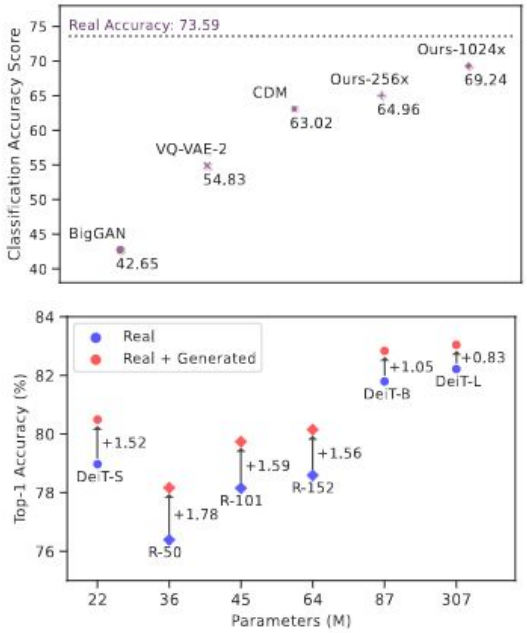
¿cuándo me va a servir esto?

Synthetic Data from Diffusion Models Improves ImageNet Classification

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia*, Mohammad Norouzi*, David J. Fleet
Google Research, Brain Team †

Abstract

Deep generative models are becoming increasingly powerful, now generating diverse high fidelity photo-realistic samples given text prompts. Have they reached the point where models of natural images can be used for generative data augmentation, helping to improve challenging discriminative tasks? We show that large-scale text-to-image diffusion models can be fine-tuned to produce class-conditional models with SOTA FID (1.76 at 256×256 resolution) and Inception Score (239 at 256 × 256). The model also yields a new SOTA in Classification Accuracy Scores (64.96 for 256×256 generative samples, improving to 69.24 for 1024×1024 samples). Augmenting the ImageNet training set with samples from the resulting models yields significant improvements in ImageNet classification accuracy over strong ResNet and Vision Transformer baselines.



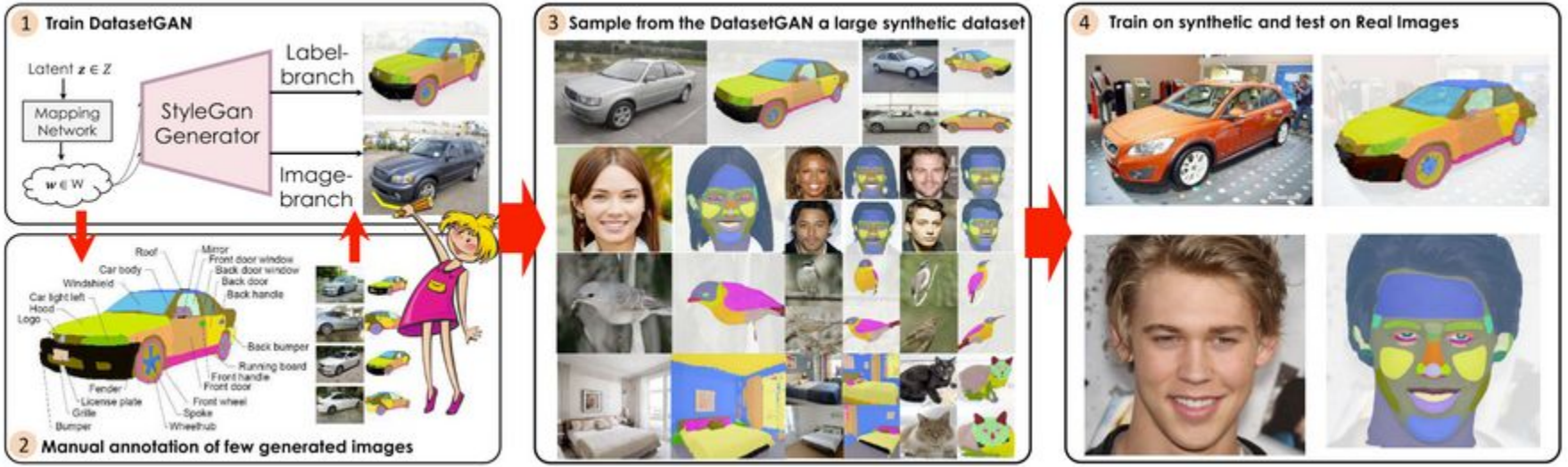
1. Introduction

Deep generative models are becoming increasingly mature to the point that they can generate high fidelity photo-

Data augmentation

Caso: quiero entrenar un clasificador, pero tengo pocos datos, o son de mala calidad.

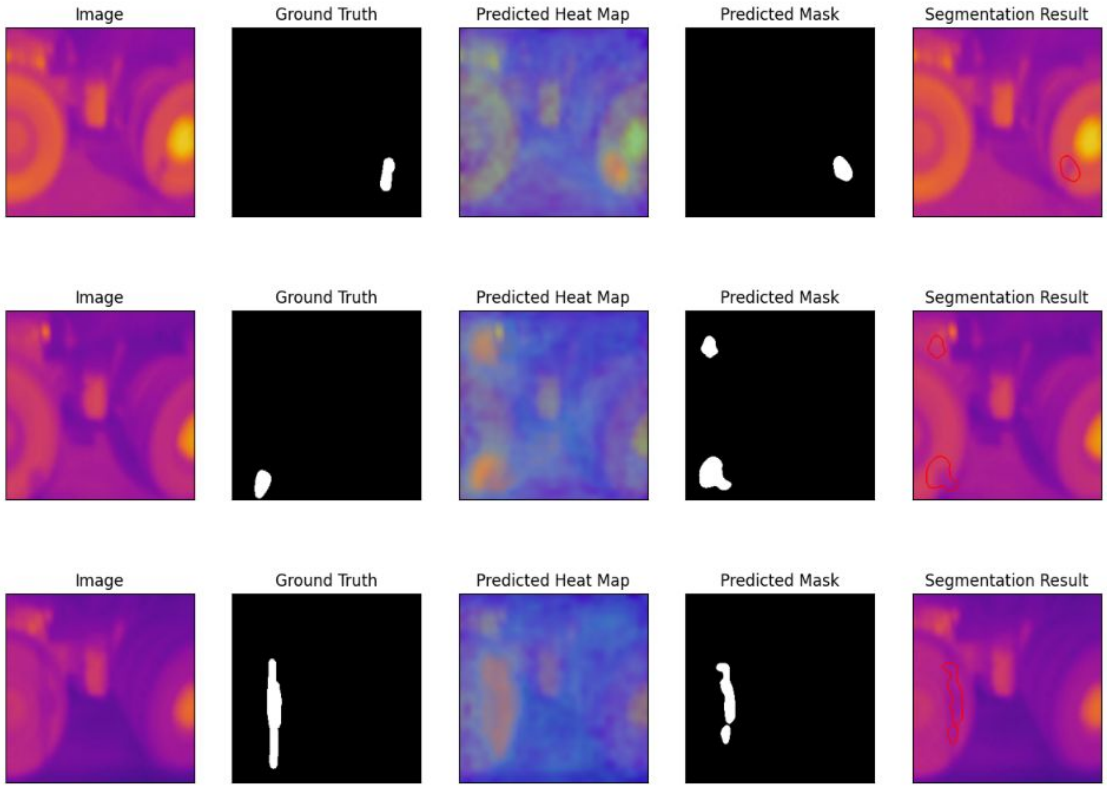
Un modelo generativo pre-entrenado puede ser *fine-tuneado* para generar datos sintéticos de alta calidad, permitiendo así entrenar un modelo competente.



DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort

https://nv-tlabs.github.io/datasetGAN/

Ejemplo real en CENIA :)








TL:DR
(Resumen)

Modelos Generativos

- Los modelos generativos **aprenden a representar distribuciones**.
- Es un área de constante estudio **sin una arquitectura “ganadora”**. Cada arquitectura tiene sus pro y contras.
- Es útil para escenarios en los que quiero entrenar un modelo y tengo pocos datos. **Permite aumentar mi dataset con datos sintéticos de alta calidad.**

Problemas de Modelos Generativos en CV

- No son útiles para todo tipo de datos; **por ejemplo, datos tabulares**.
- **Requieren muchos recursos para ser entrenados desde 0**.
- **Heredan los sesgos** de los datos de entrenamiento.

Fin ✌️

Avances Recientes

ControlNet

Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala
Stanford University
{lvmin, anyirao, maneesh}@cs.stanford.edu

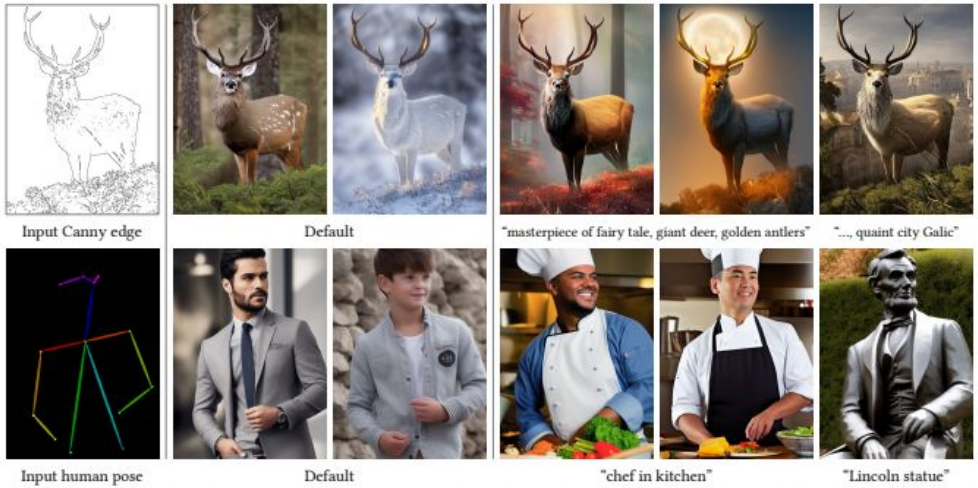
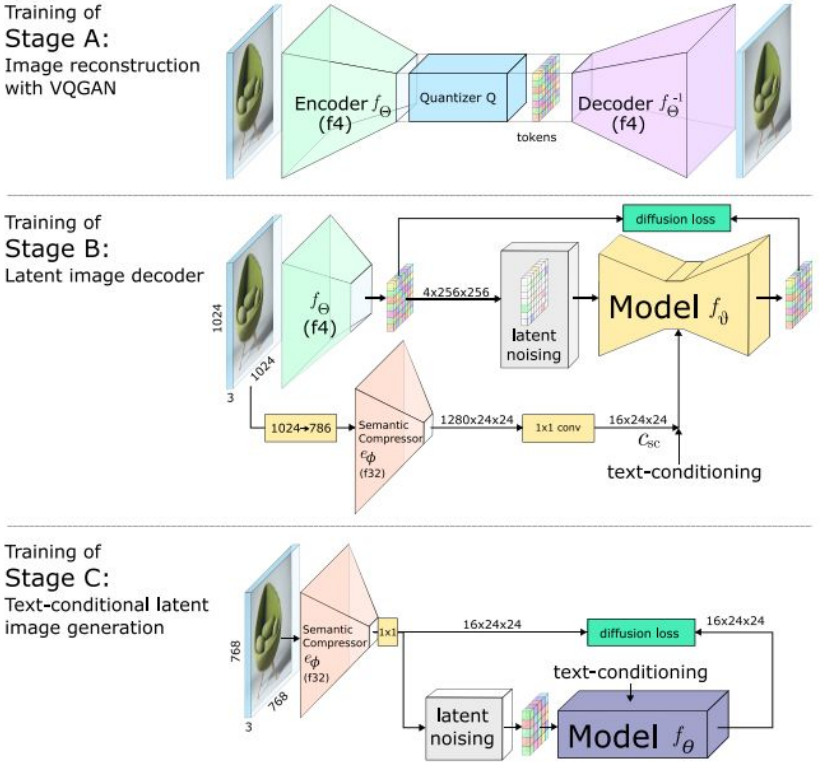


Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc*., to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen".

Würstchen (Stable Cascade)



Diffusion Transformer (DiT)

Scalable Diffusion Models with Transformers

William Peebles*
UC Berkeley

Saining Xie
New York University

Figure 1. **Diffusion models with transformer backbones achieve state-of-the-art image quality.** We show selected samples from two of our class-conditional DiT-XL/2 models trained on ImageNet at 512×512 and 256×256 resolution, respectively.