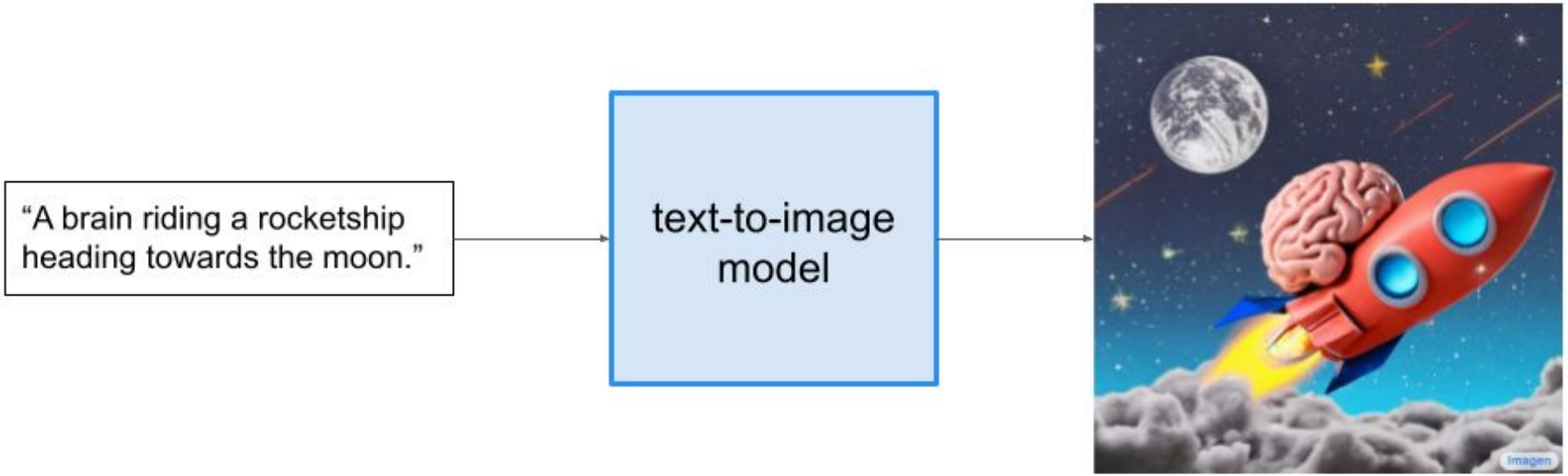
Text-to-image models

Enrico Benedetti, Filippos Papandreou & Simona Scala

Intro

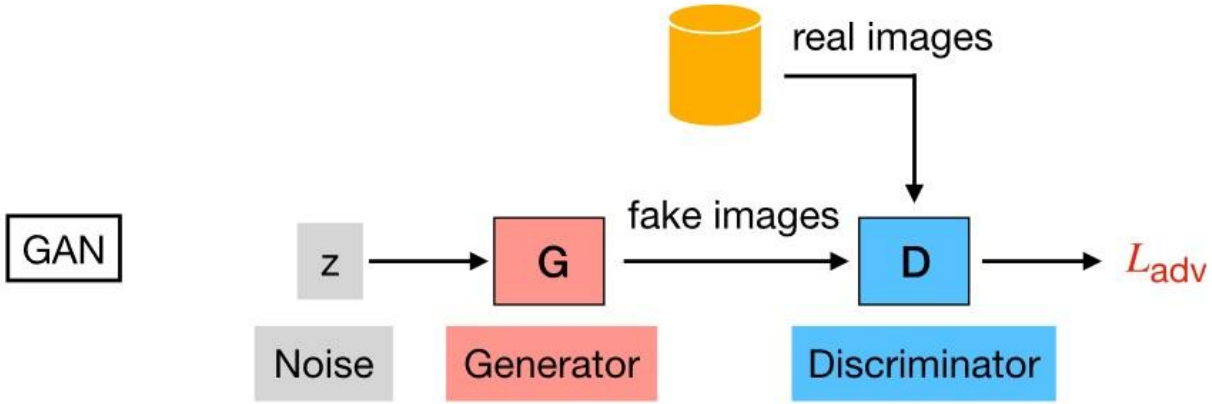


From "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", Saharia et al.

Generative Adversarial Networks (GANs)

Interplay between two neural networks

- Generator
- Discriminator



From “Improved techniques for training GANs”, Salimans et al.

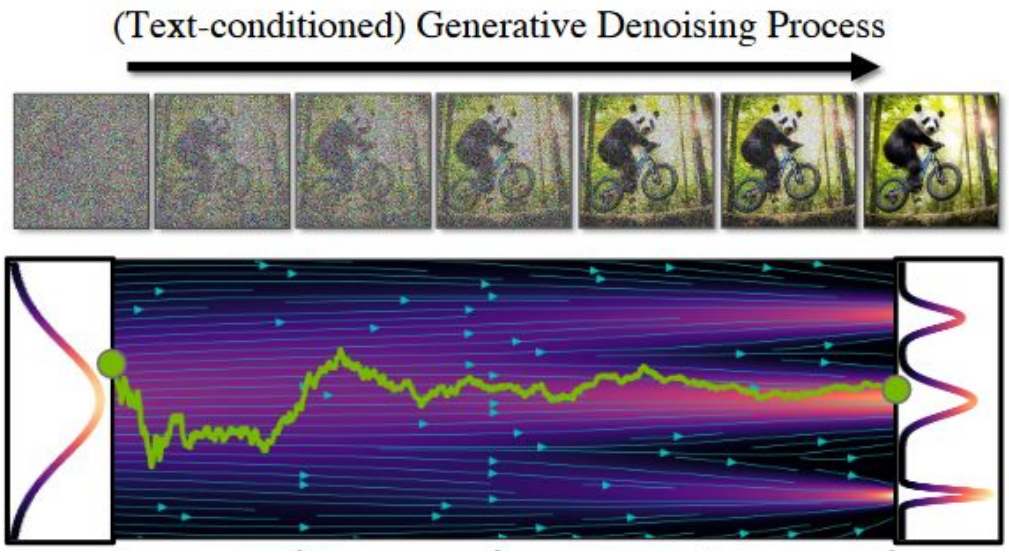
Conditional approach for text-to-image synthesis using GANs

The typical architecture for text-to-image synthesis using GANs consists of

- Text encoder
- Image generator
- Discriminator
- Training Process
- Datasets

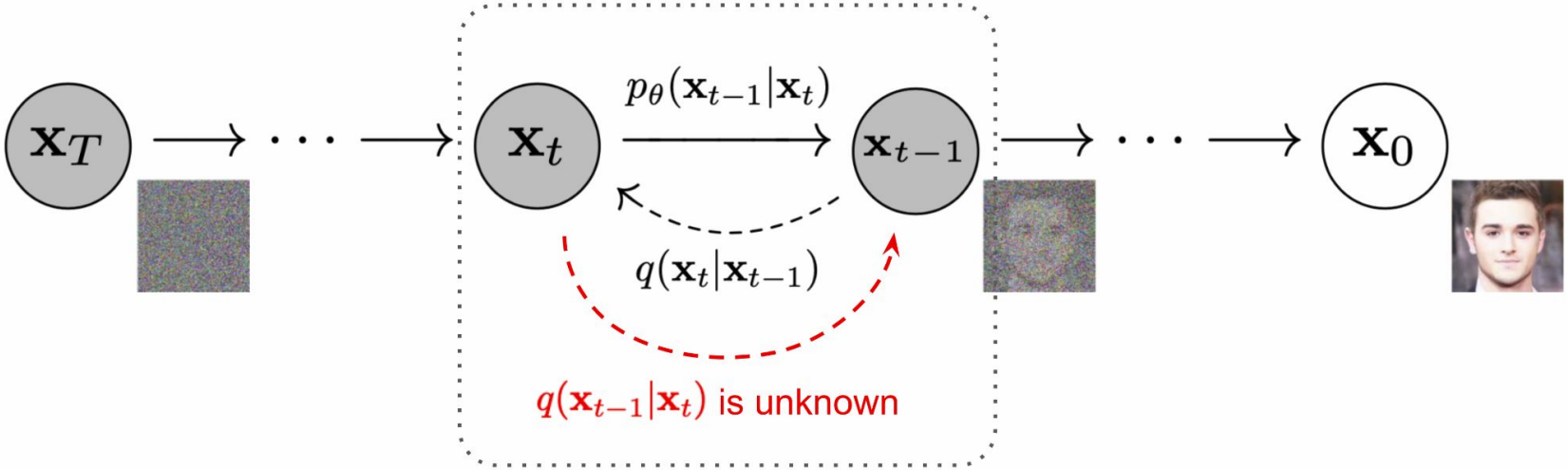
Diffusion models

- Start with white noise
- Learn how to remove some at each step
- End with an image from the data distribution



from “eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers”, Balaji et al.

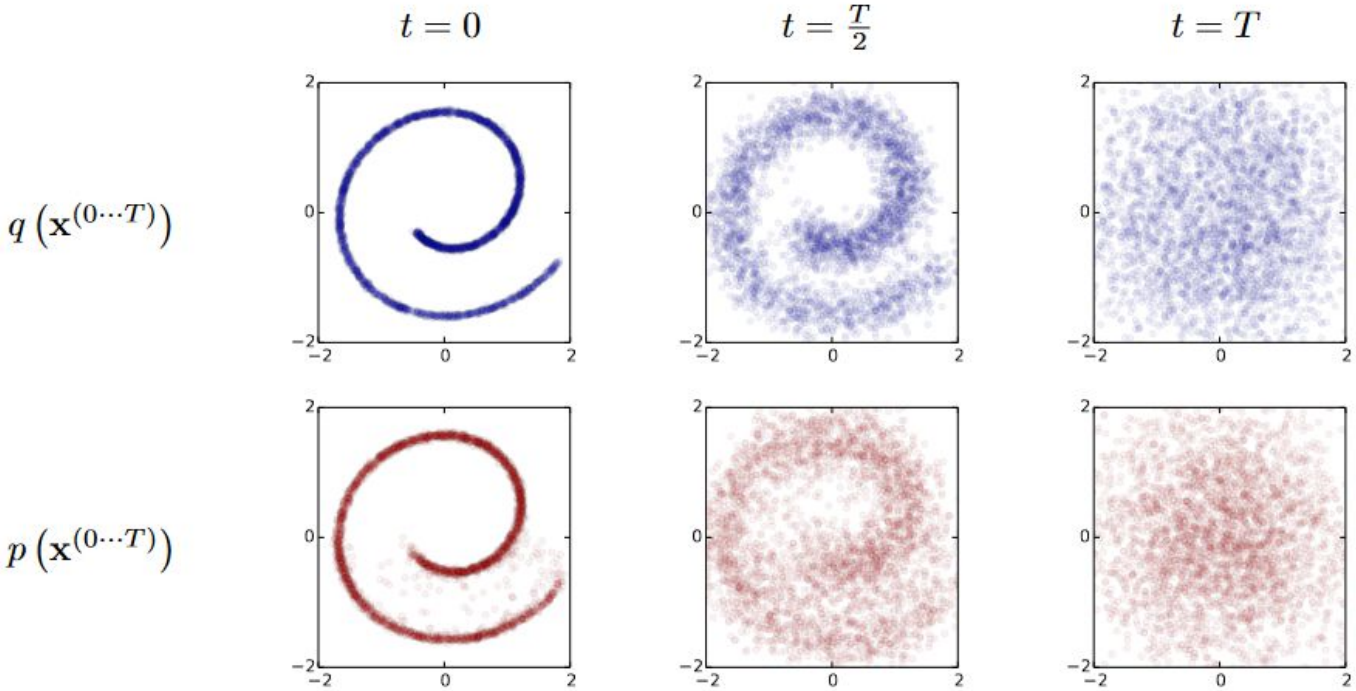
Forward & backward processes



$$L_{\text{simple}} = E_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right]$$

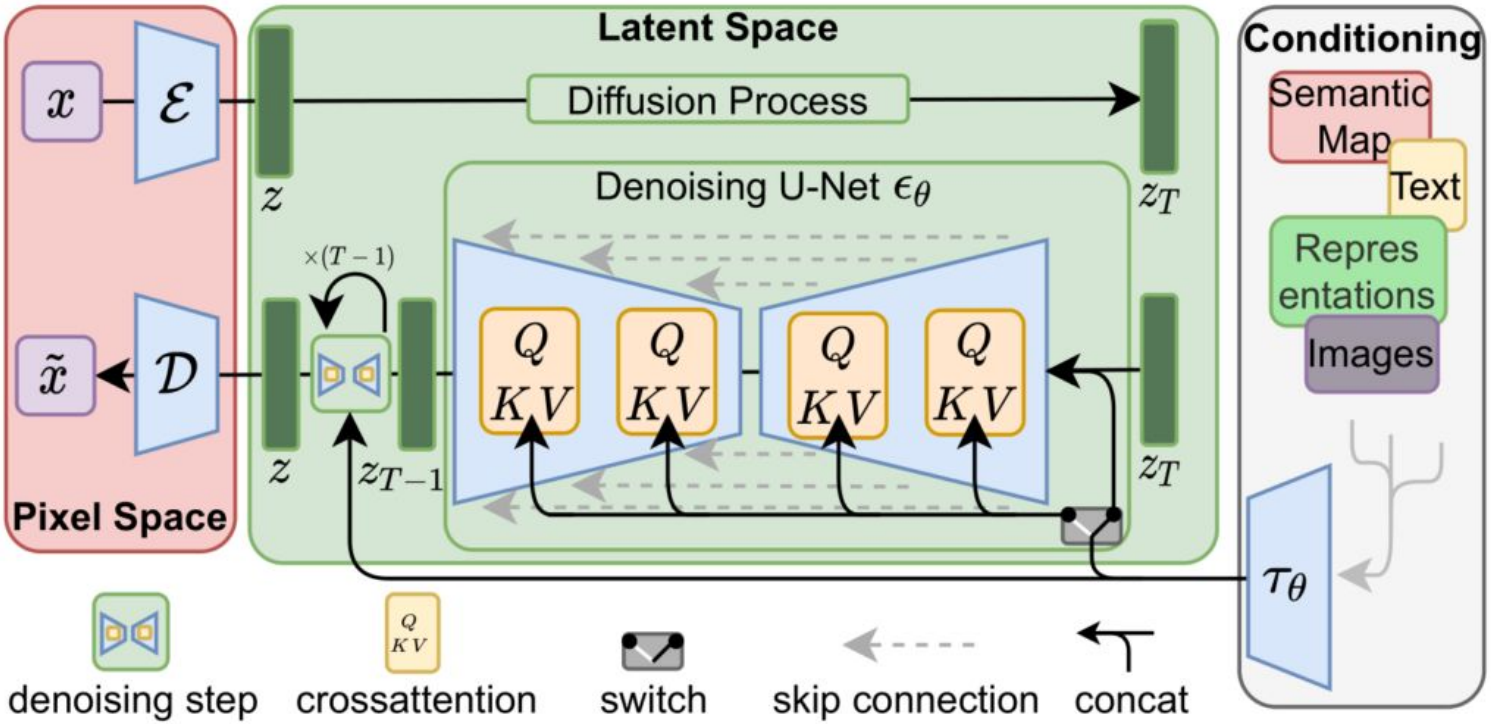
from “Denoising Diffusion Probabilistic Models”, Ho et al.

Forward & backward processes



from “Deep Unsupervised Learning using Nonequilibrium Thermodynamics” by Sohl-Dickstein et al.

Example: latent diffusion models



from “High-Resolution Image Synthesis with Latent Diffusion Models”, Rombach et al.

In short, diffusion models…

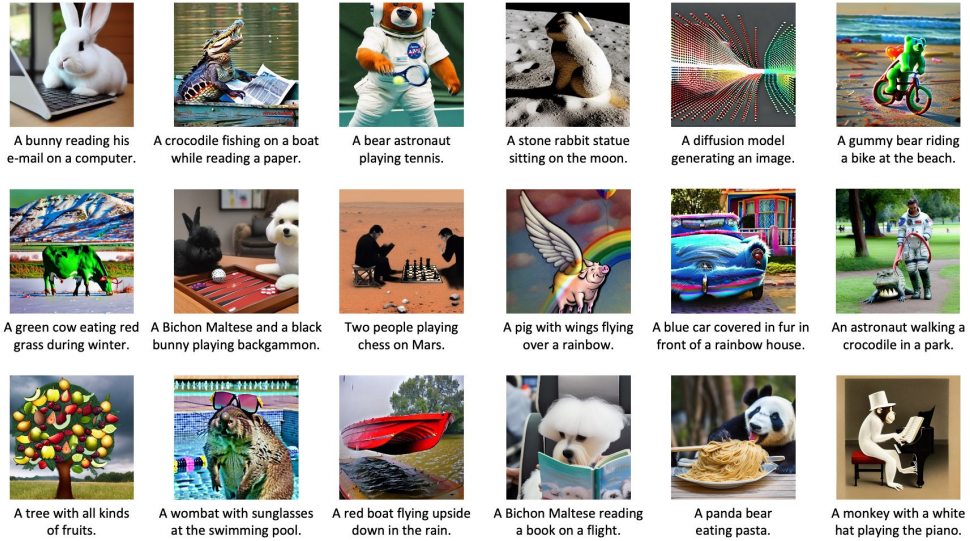
- Are a hot way to do conditional image generation
- Achieve high quality and diversity, but the iterative steps make them slow
- Are based on noising and denoising processes
- The main ML foundations are the U-net and the attention mechanism

Evaluation metrics

- Inception Score (IS)
  - Leverages a pre-trained Inception-v3 network: classification
  - Measures both recognizability and diversity of generated images
- Fréchet Inception Distance (FID)
  - Leverages a pre-trained Inception-v3 network: extraction
  - Measures the similarity between model-generated images and real images
- R-Precision (RP)
  - Compares the encoding vectors of the image and the caption using cosine similarity
  - Assesses the performance in matching image queries with textual descriptions
- Human evaluation
  - List of prompts to test image quality and fidelity to the caption

GAN-based vs. Diffusion-based methods

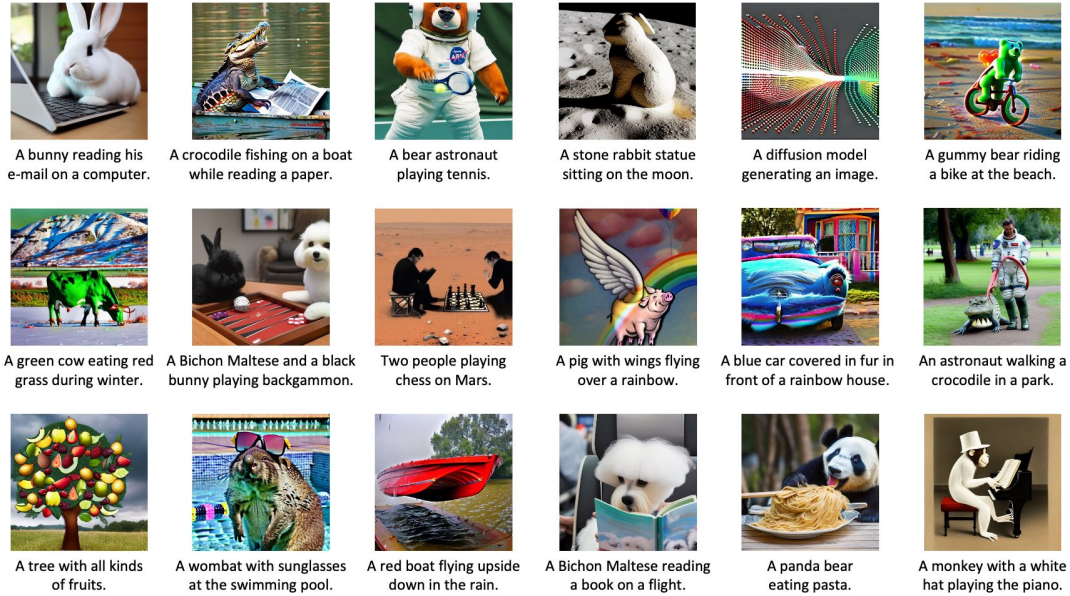
from “Adversarial text-to-image synthesis: A review”, Frolov et al.



from “Diffusion Models in Vision: A Survey”, Croitoru et al.

Comparisons

- GAN-based methods
- Diffusion-based methods

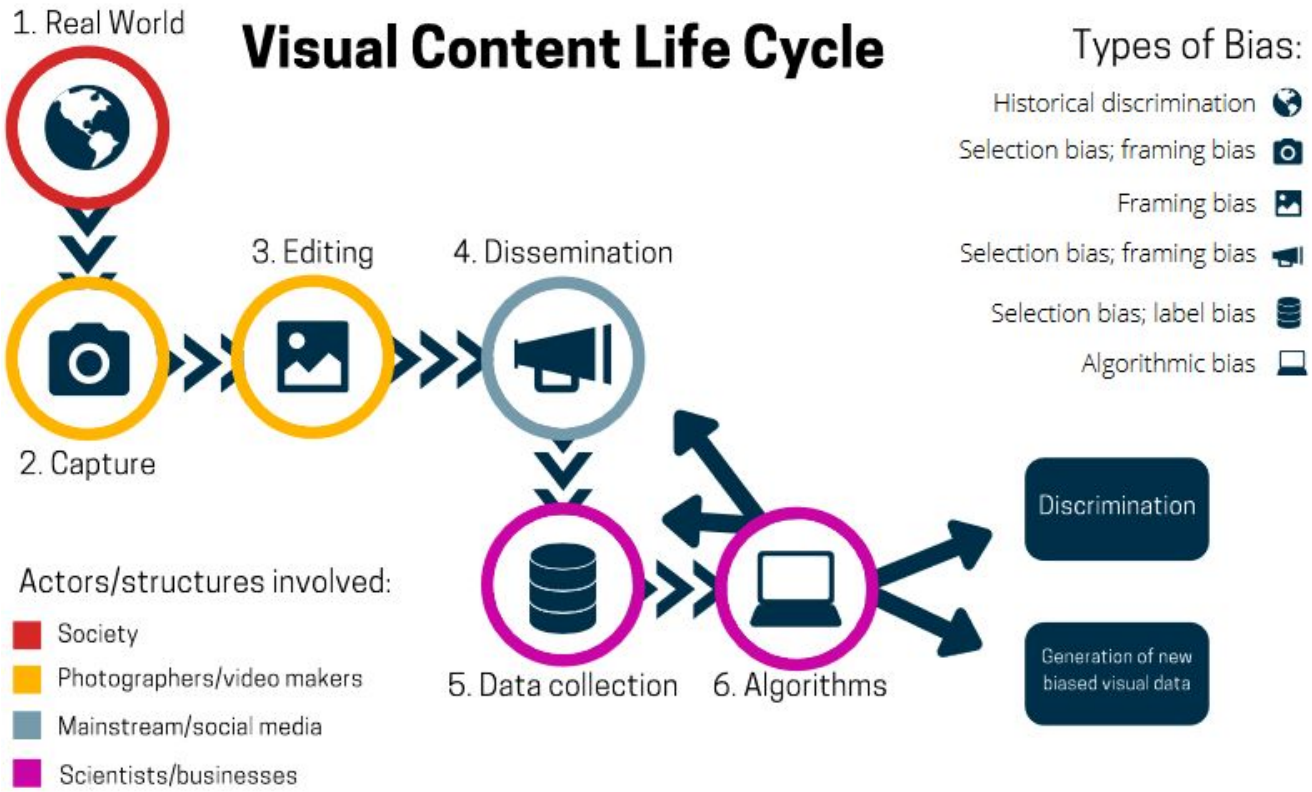


Ethics

We should explore some ideas and issues…

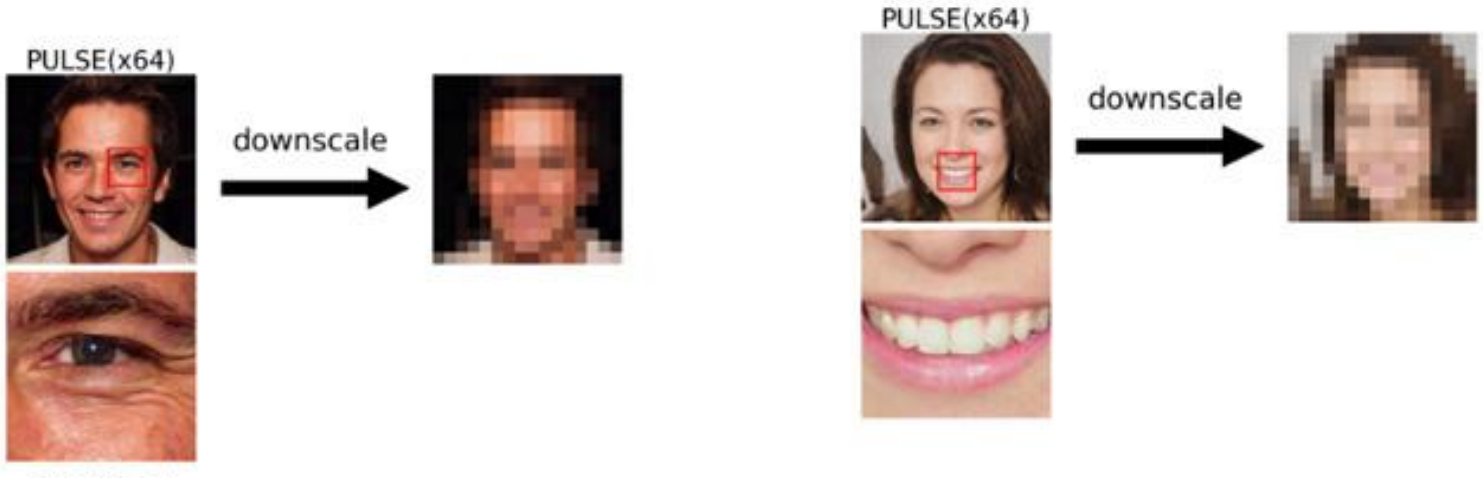
- Bias
- Deepfakes & fake content
- Creativity, Art, Copyright

Bias



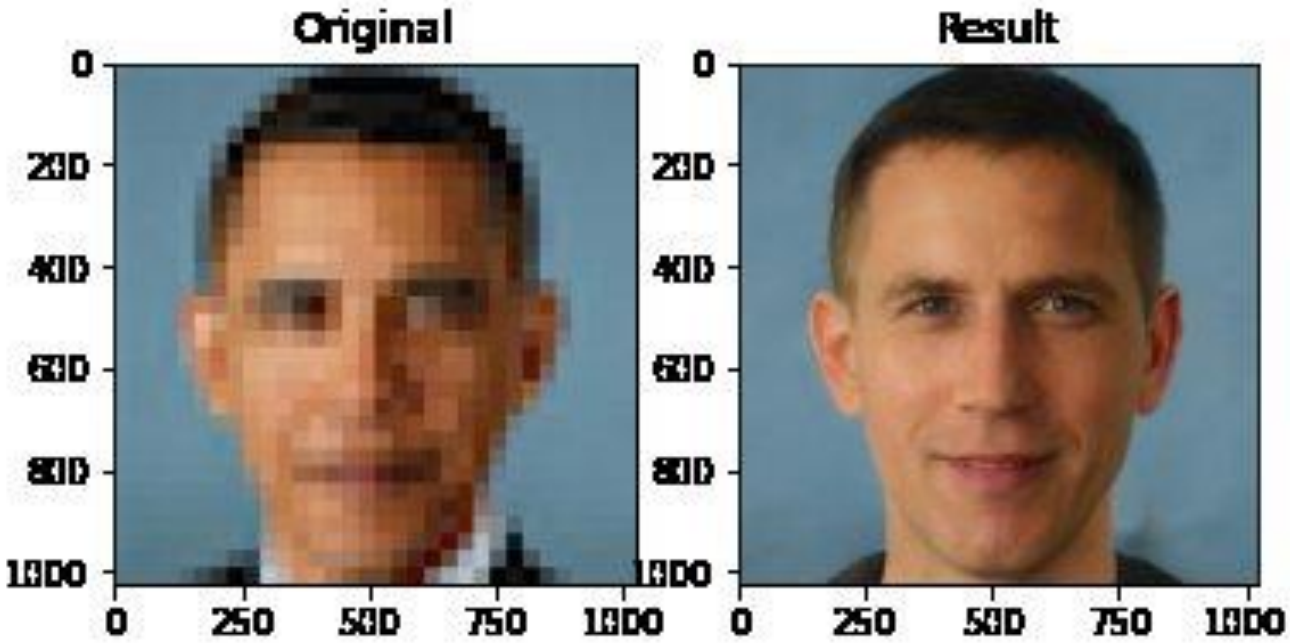
from “A Survey on Bias in Visual Datasets”, Fabbrizzi et al.

Bias: PULSE



from the PULSE github repo (link)

Bias: PULSE



from Twitter (link)

Bias: ethical interventions

from “How well can text-to-image generative models understand ethical natural language interventions?”, Bansal et al.

Fake content & drip

(from the internet)

Fake content & drip

(also from the internet)

Generative AI for fake news?

Some considerations

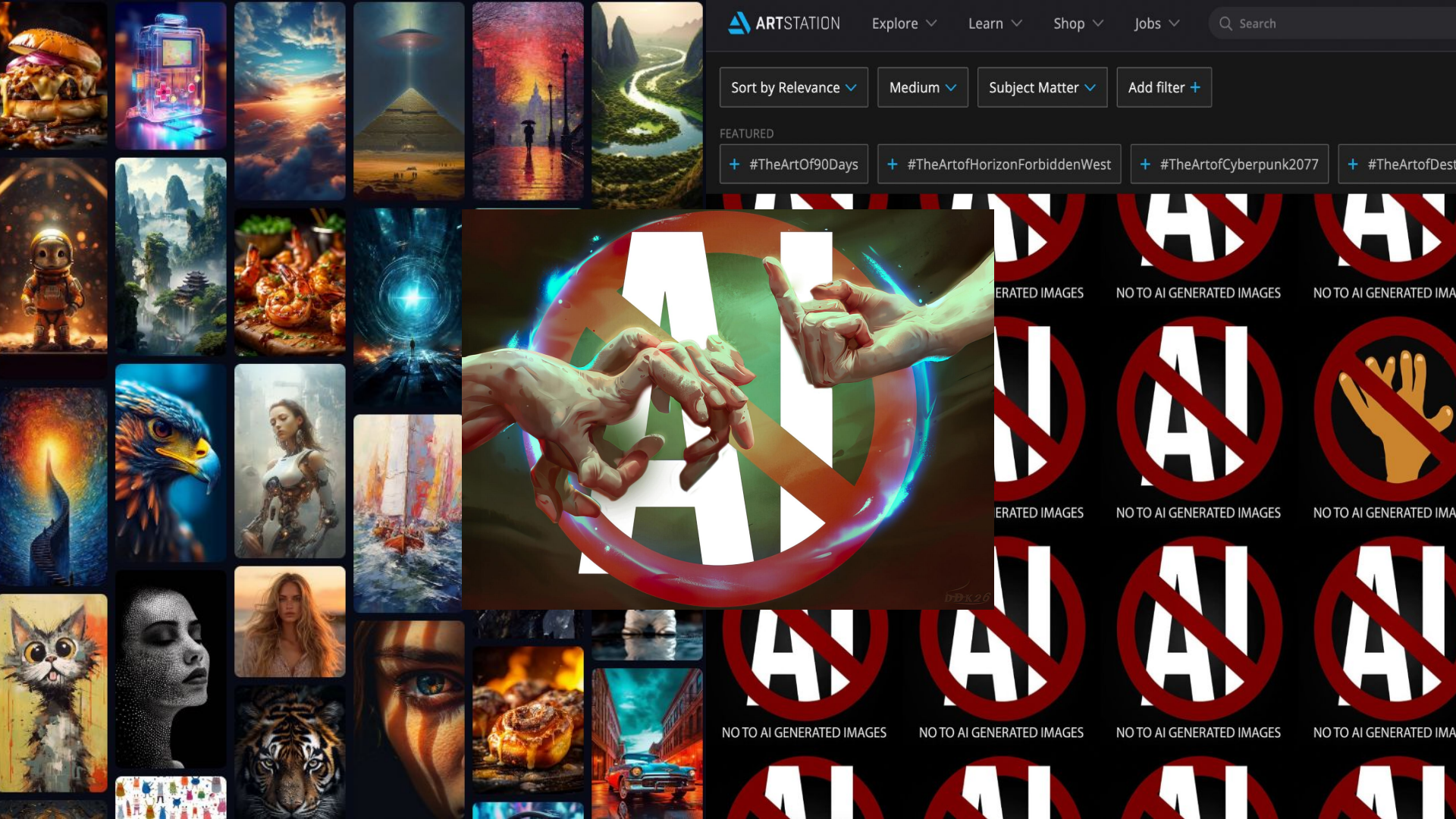
- Cost/Time efficiency
- Ease of access to models & resources
- Model filtering
- Social media platforms’ self-regulation
- Detection mechanisms
- General level of digital literacy

Legal frameworks

- GDPR
- AI Act
- Copyright law

Creativity, Copyright & other issues

- Artists' work has been used in crawled datasets (e.g. LAION) and can be reproduced by models
- Their style can be reproduced as well
- Datasets contain personal and/or sensitive data + "inappropriate" content: automatic filtering is needed
- Is everything on the internet public domain?
- Commercial platforms keep their datasets private, though many open source projects exist


ARTSTATION
Explore
Learn
Shop
Jobs
Search
Sort by Relevance
Medium
Subject Matter
Add filter +
FEATURED
+ #TheArtOf90Days
+ #TheArtofHorizonForbiddenWest
+ #TheArtofCyberpunk2077
NO TO AI GENERATED IMAGES

Hands-on Generation

How to generate images

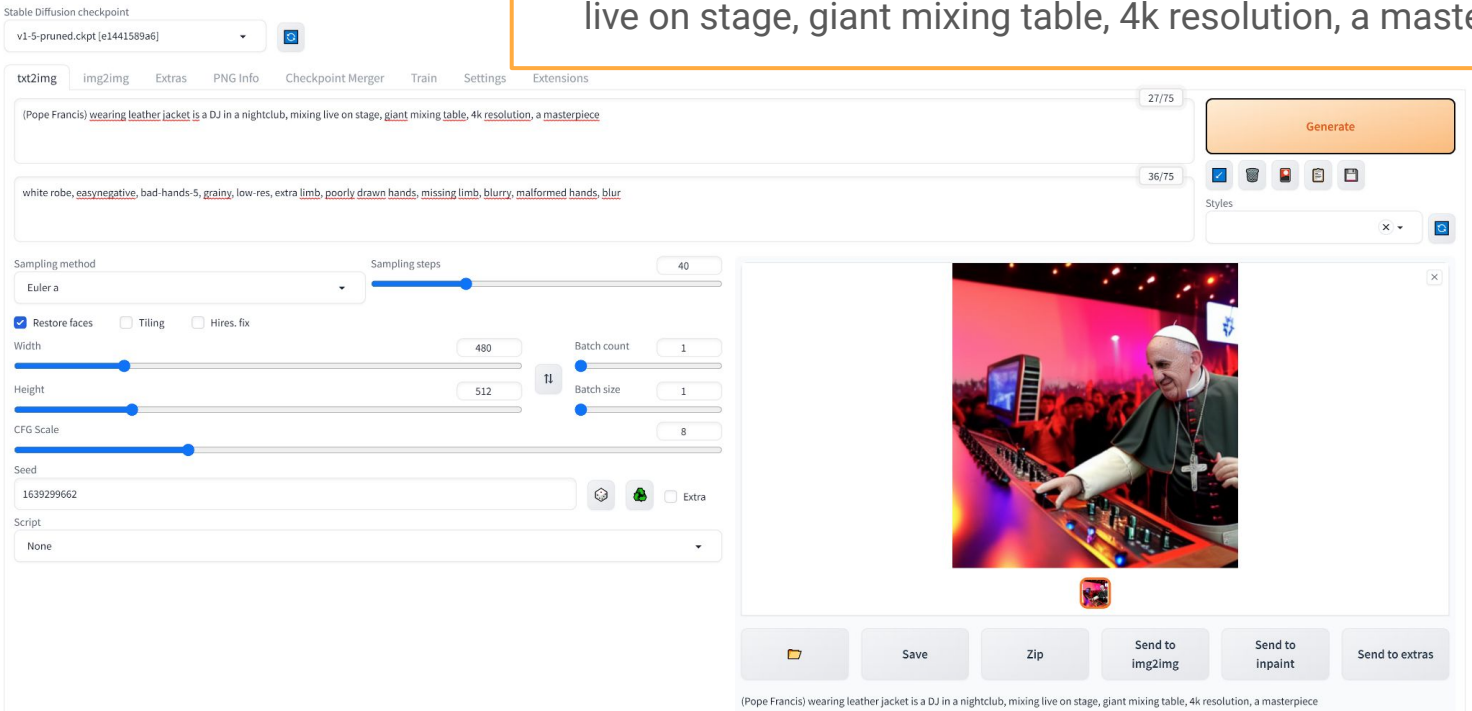
- Coding notebooks
- Using text-to-image models locally (WebUI)
- Using online websites (e.g. Midjourney)

Coding notebooks

An example: [Stable Diffusion guide by Hugging Face]

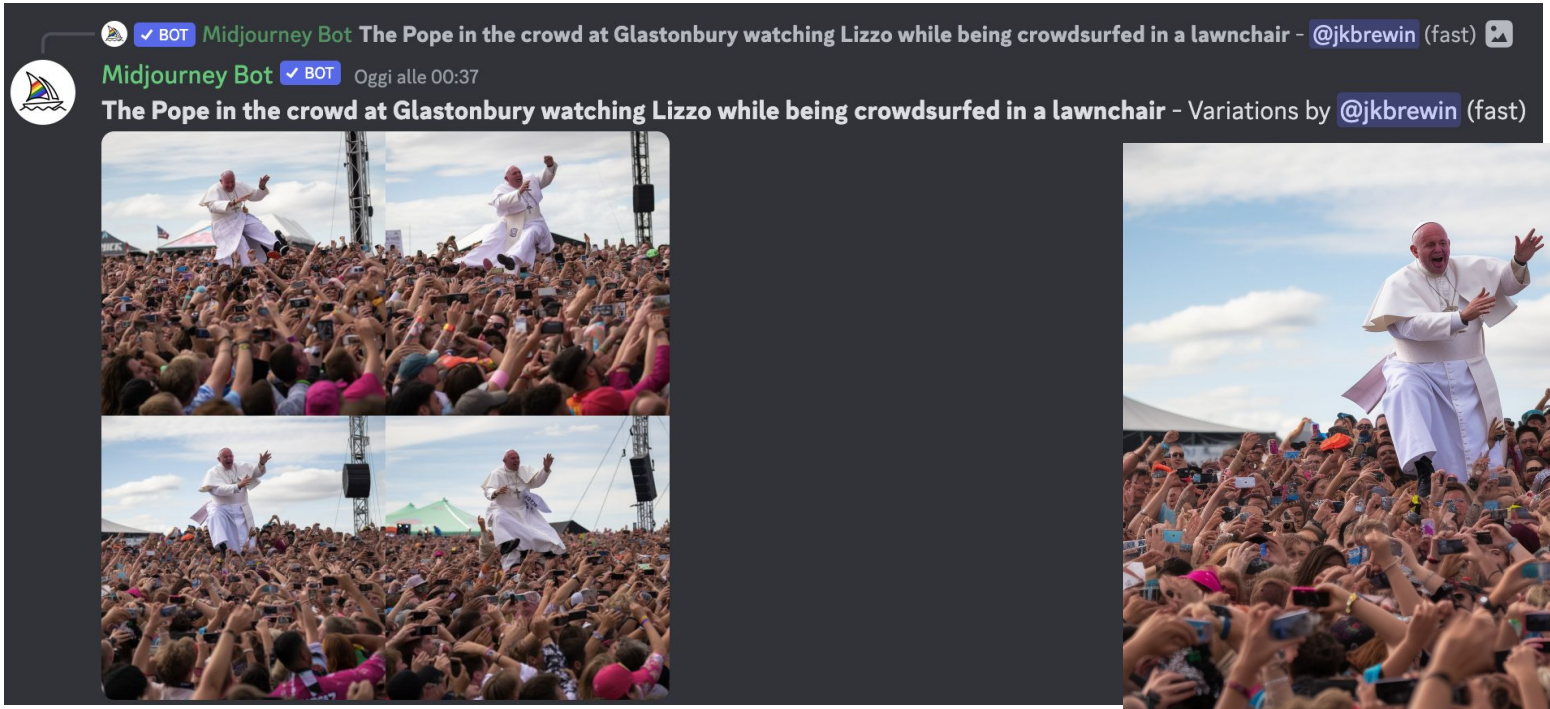
Using text-to-image models locally

"(Pope Francis) wearing leather jacket is a DJ in a nightclub, mixing live on stage, giant mixing table, 4k resolution, a masterpiece"



From stable-diffusion-webui (link)

Using online websites (e.g. Midjourney)




The End