Deep Generative Models

Jan Čech

Deep Generative Models

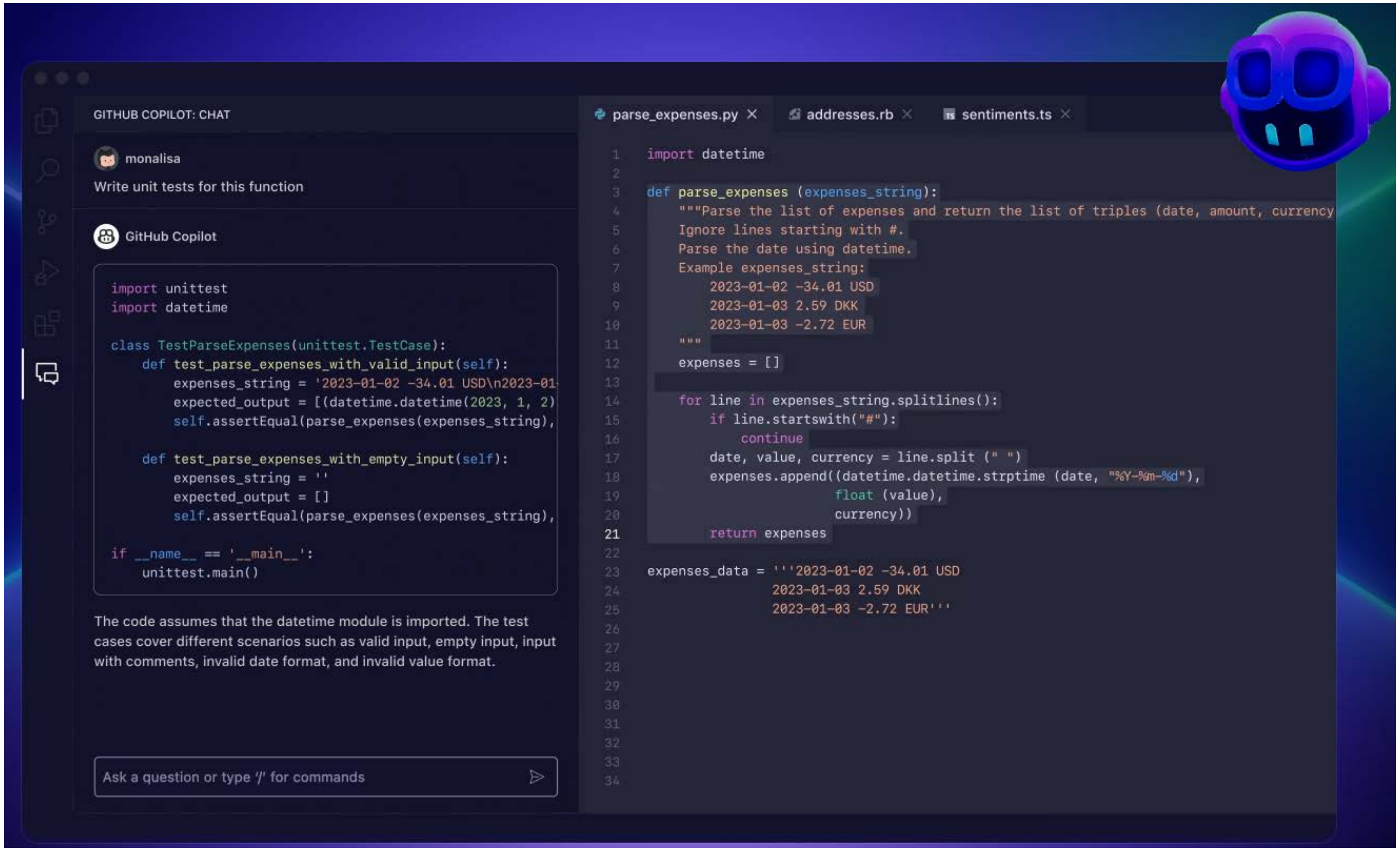
- Models generating samples from a given distribution
  - The target distribution can be very complex (high-dimensionality)
  - Synthesis of **realistic** samples of, e.g. Text, Code, Audio, Music, **Images**, Video, 3D models

- Unconditioned / Conditioned models
  - Conditioning by a “prompt” (text, image, semantic maps, …)

- **Lecture outline**
  1. Examples
  2. Brief taxonomy and Evaluation
  3. GANs
  4. Diffusion Models

Examples of Deep Generative Models

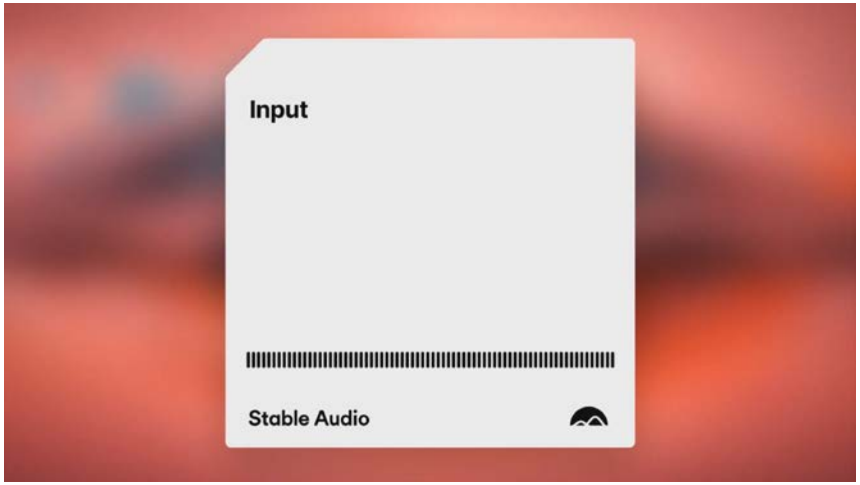
- Text
  - Large Language models
  - GPT 4 (OpenAI), Llama 3 (Meta) – open source
- Code
  - GitHub Copilot – free for students



Examples of Deep Generative Models

- Audio
  - text2audio, audio2audio
    - [Stable Audio 2.0](#) (Stability.AI)

      Prompt: Epic music, big drums, strong melody



    - [Suno AI](#) (Cambridge, MA, USA)

      Lyrics: Computer Vision Method course focuses on the following computer vision problems: finding correspondences between images using image features and their robust invariant descriptors, image retrieval, object detection and recognition, and visual tracking.
      Style: Large symphonic orchestra and children chorus, epic melody.

  - text2speech
    - Many models, emotional speech, voice cloning, very realistic…

Examples of Deep Generative Models

- Unconditioned Image generators
  – Generating photo-realistic samples from image distributions

(Images synthetized by a random sampling)

Examples of Deep Generative Models

- Text2image
  - DALL-E (OpenAI), Imagen (Google), Midjourney, …
  - **StableDiffusion** (Stability.AI, free open source)

panda mad scientist mixing sparkling chemicals, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

Examples of Deep Generative Models

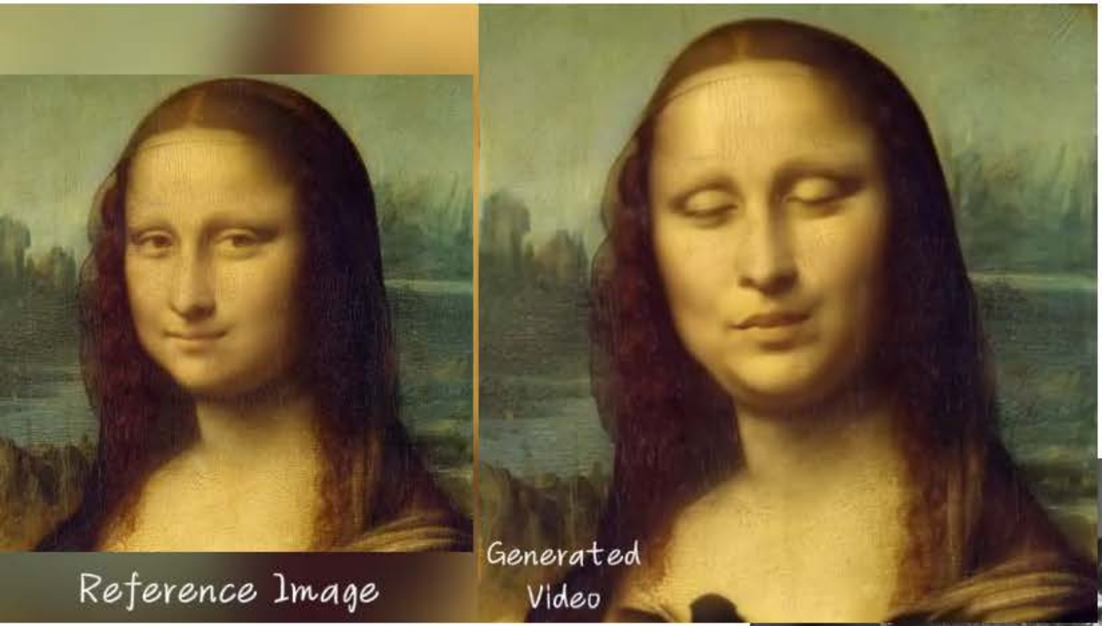
- Video
  - Text2video, image2video, video2video
  - SORA (OpenAI)

A litter of golden retriever puppies playing in the snow. Their heads pop out of snow

Beautiful, snowy Tokyo city is bustling. The camera moves through the bustling city street, following several people enjoying the beautiful snowy weather and shopping at nearby stalls. Gorgeous sakura petals are flying through the wind along with snowflakes

Examples of Deep Generative Models

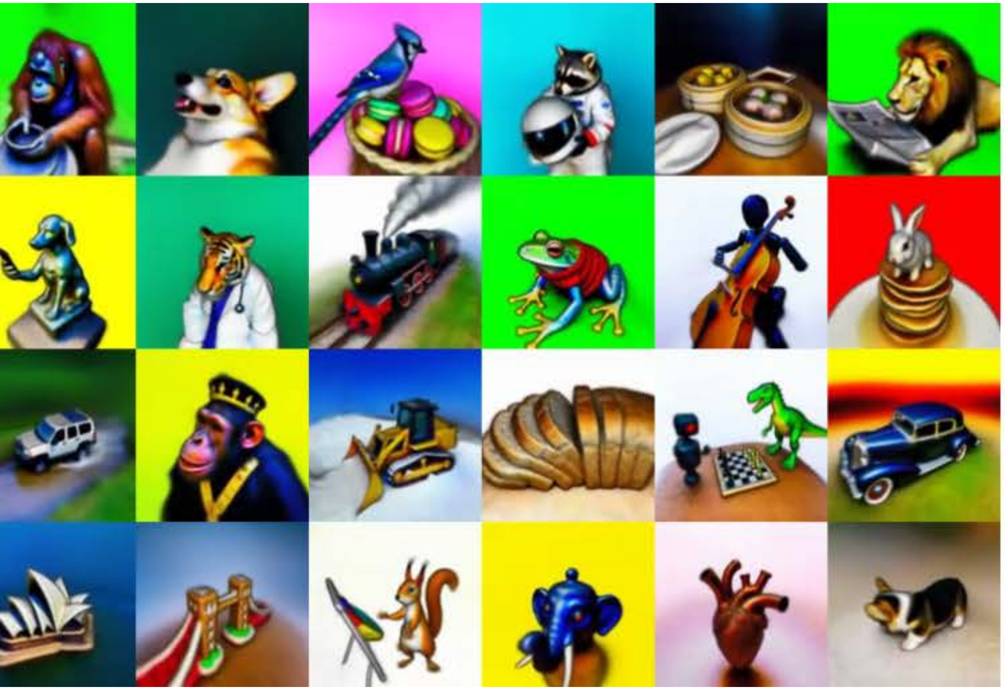
- "Talking/Singing head" (audio2video)
  - EMO: Emote Portrait Alive [Tian-2024] (Alibaba group)



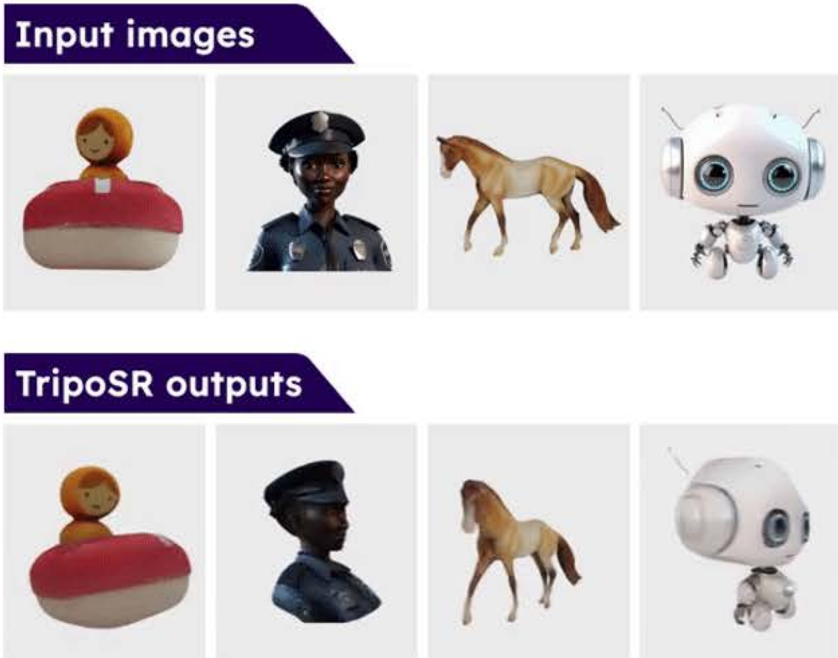
Examples of Deep Generative Models

- 3D models (audio2video)
  - Text to 3D [DreamFusion-2022] (Google), [Magic3D-2023] (NVIDIA)
  - Image to 3D (TripoSR, Stability.AI)

DreamFusion



TripoSR



Brief Taxonomy and Evaluation

Taxonomy of Deep Generative Models

- Several approaches:
  1. Normalizing flow models [Dinh-2017]
  2. Autoregressive models [Oord-2016]
  3. Variational Autoencoders [Kingma-2014]
  4. **Generative Adversarial Networks (GANs)** [Goodfellow-2014]
  5. **Diffusion models** [Sohl-Dickstein-2015, Rombach-2022]

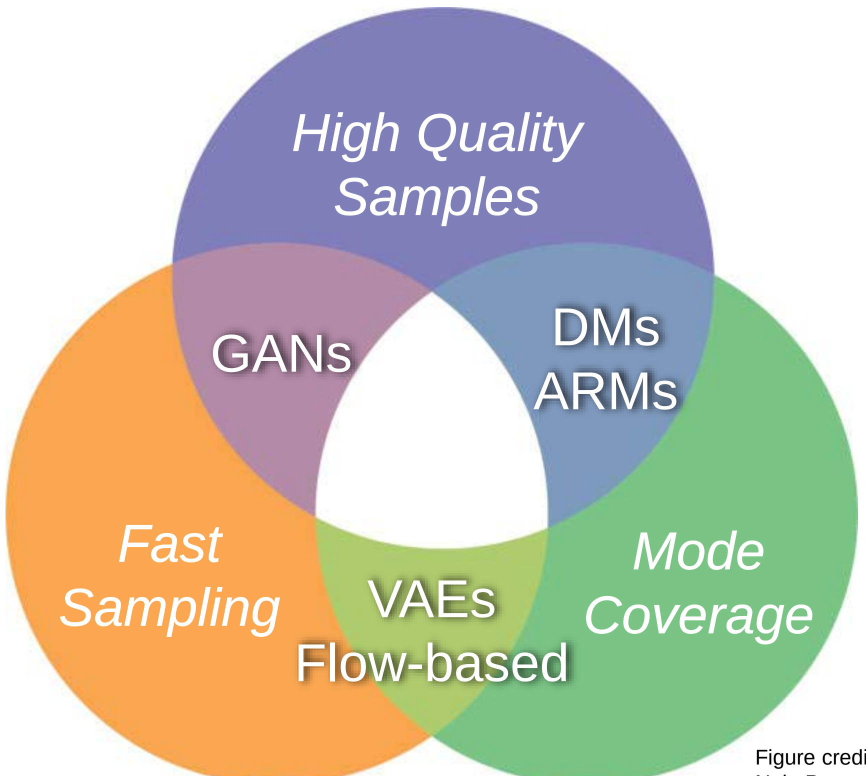


Figure credit
Nela Petrzelkova

How to Measure Quality of Generative Models?

- Synthetic samples should resemble samples of real distribution in the sense of:
  - Fidelity (no obvious artifacts visible)
  - Diversity (enough variability respecting the original distribution)
    - Mode collapse – model always generate the same sample
- All methods use classifier trained on ImageNet (usually Inception v3)

1. Inception score (IS) [Salimans-2016]
   - Only synthetic dataset
   - Output softmax score $p(y|x)$

low entropy

high entropy

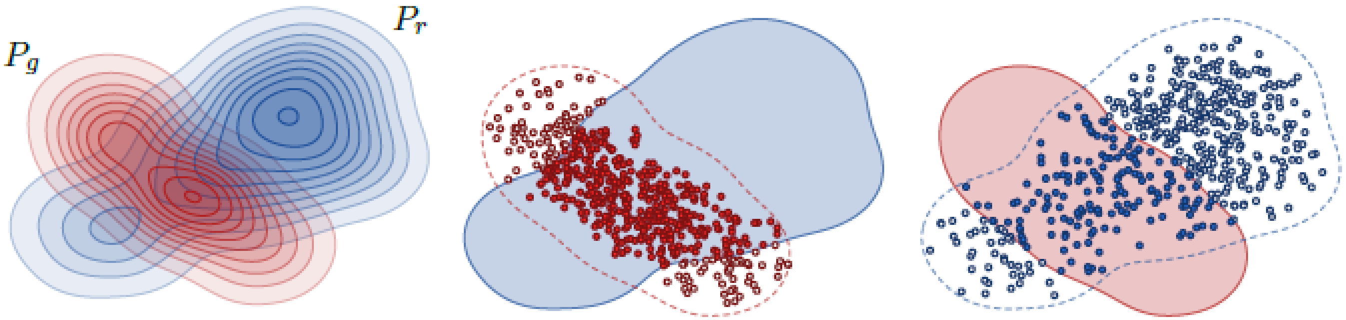
$$\mathrm{IS} = \exp\Big(\mathbb{E}_x \mathrm{KL}(p(y|x) || p(y)\Big)$$

2. Fréchet Inception Distance (FID) [Heusel-2017]
   - Two datasets – synthetic, real
   - Each sample is “embedded” (features of the penultimate layer)
   - Fit Gaussians $(\mu, \Sigma), (\mu', \Sigma')$

$$\mathrm{FID} = ||\mu - \mu'||_2^2 + \mathrm{Tr}\Big(C + C' - 2(CC')^{1/2}\Big)$$

How to Measure Quality of Generative Models?

3. Precision-Recall for distributions [Kynkäänniemi-2019]



(a) Example distributions (b) Precision (c) Recall

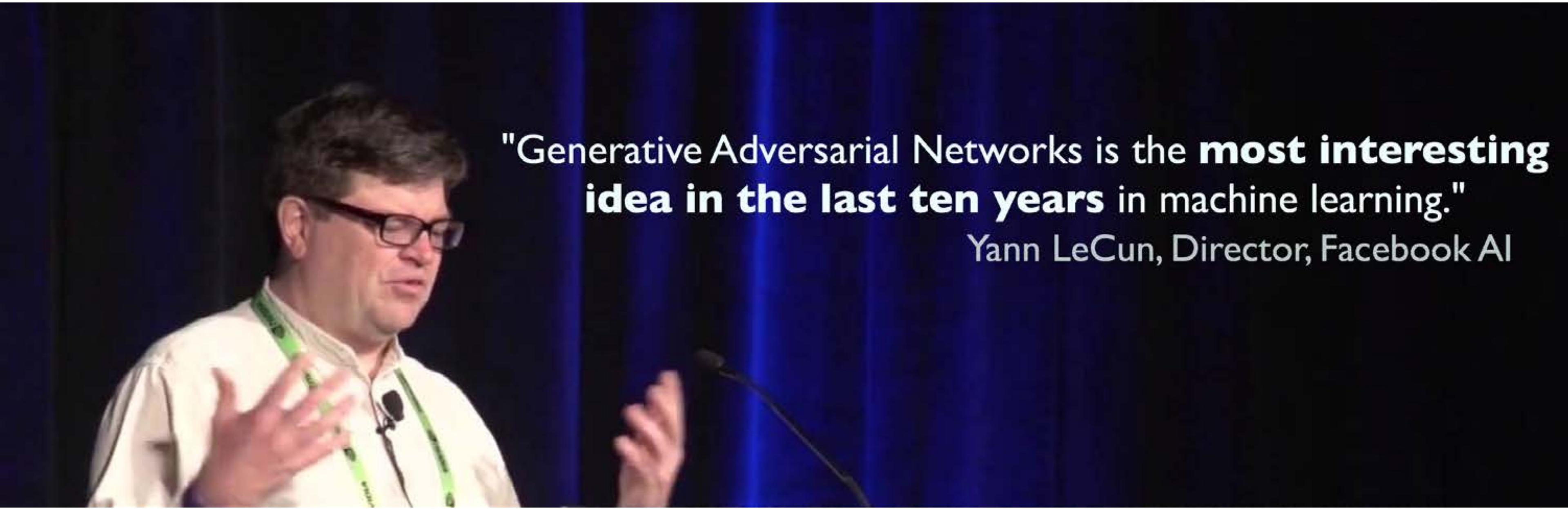
$$f(\phi, \Phi) = \begin{cases} 1, & \text{if } \left\|\phi - \phi'\right\|_2 \leq \left\|\phi' - \mathrm{NN}_k\left(\phi', \Phi\right)\right\|_2 \text{ for at least one } \phi' \in \Phi \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{precision}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f(\phi_g, \Phi_r)$$

$$\text{recall}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f(\phi_r, \Phi_g)$$

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs)



Generative Adversarial Networks (GANs)

[Goodfellow-2014]

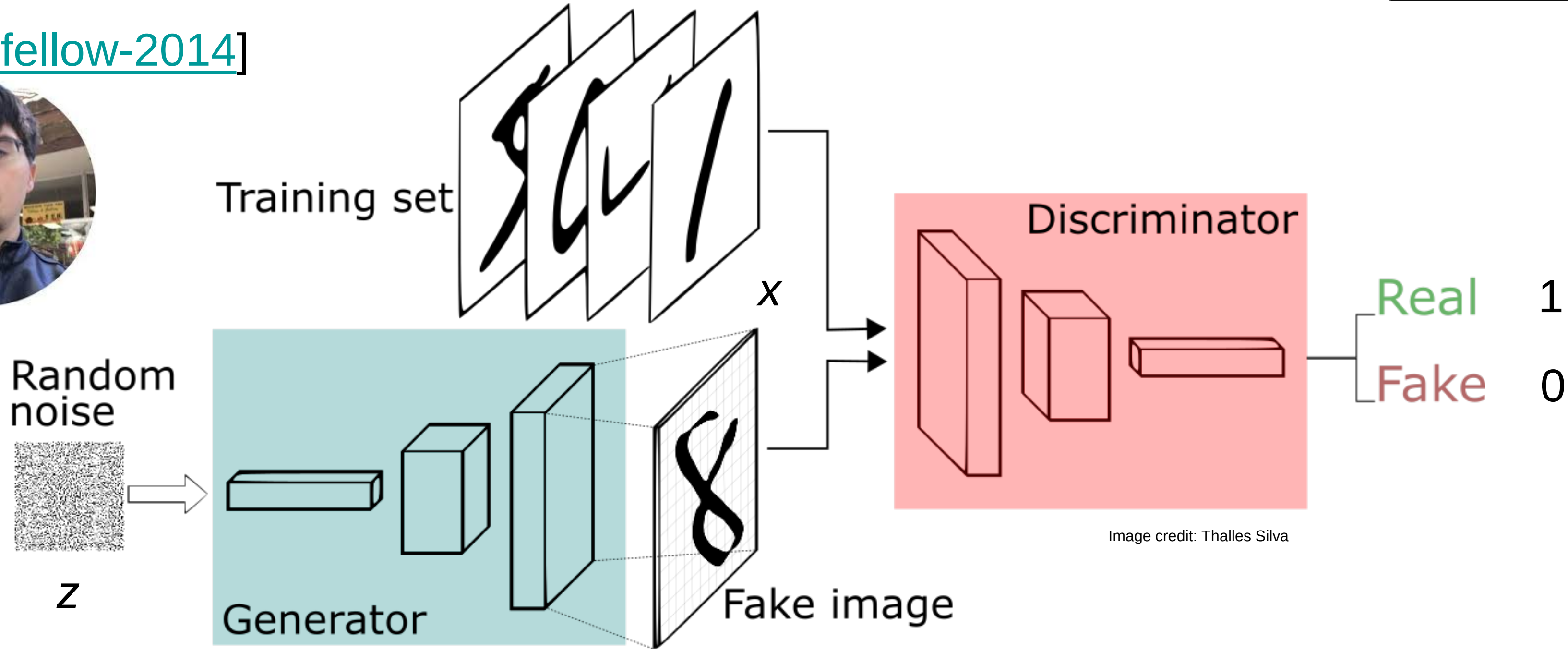
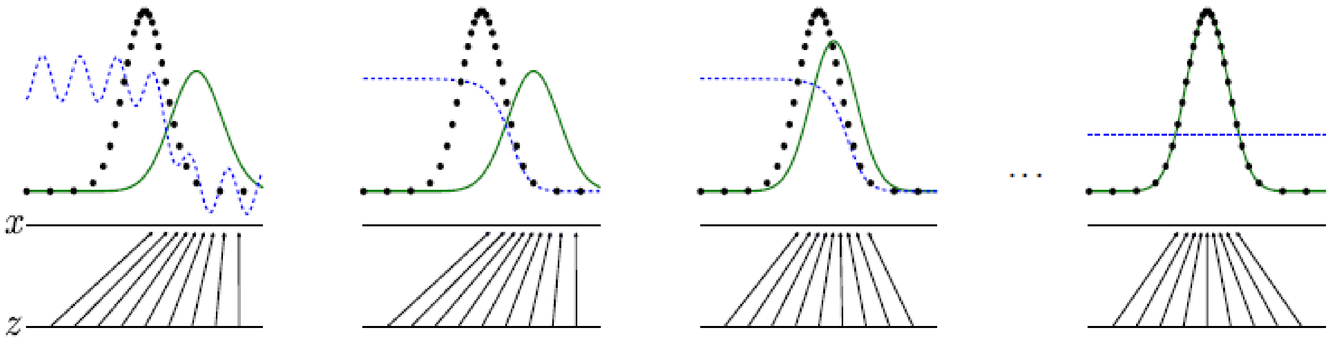


Image credit: Thalles Silva

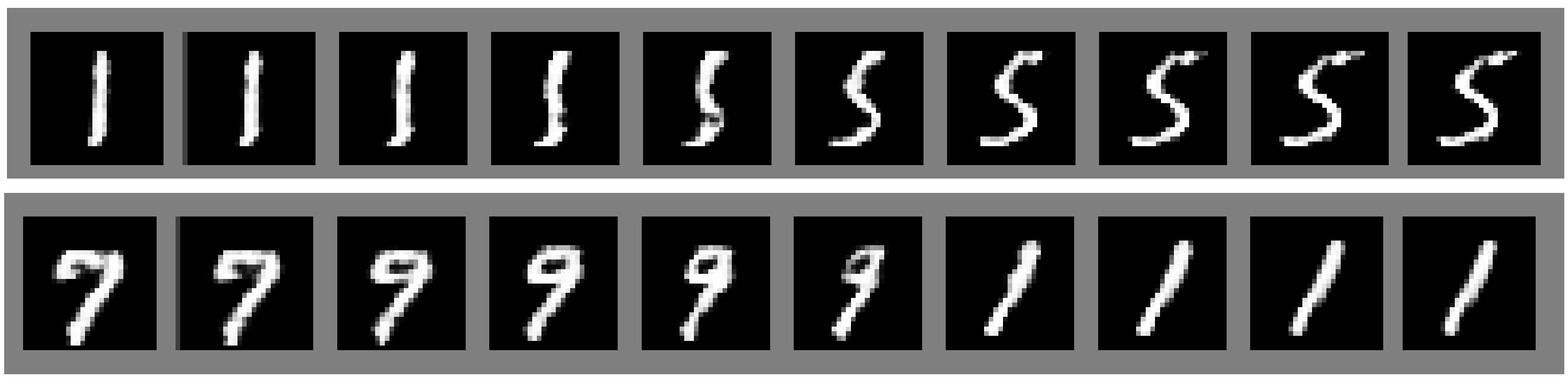
- Two networks: Generator G: $N(0,1)^k \rightarrow X$, Discriminator D: $X \rightarrow [0,1]$
- Min max game between G and D when training
  - The discriminator tries to distinguish generated and real samples
  - The generator tries to fool the discriminator

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

Generative Adversarial Networks (GANs)



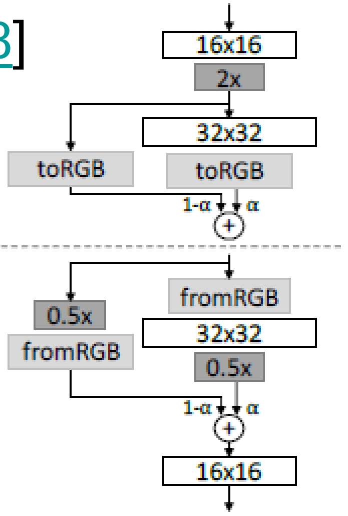
- Seems to capture the image manifold
  - Smooth transitions when interpolating in the latent space

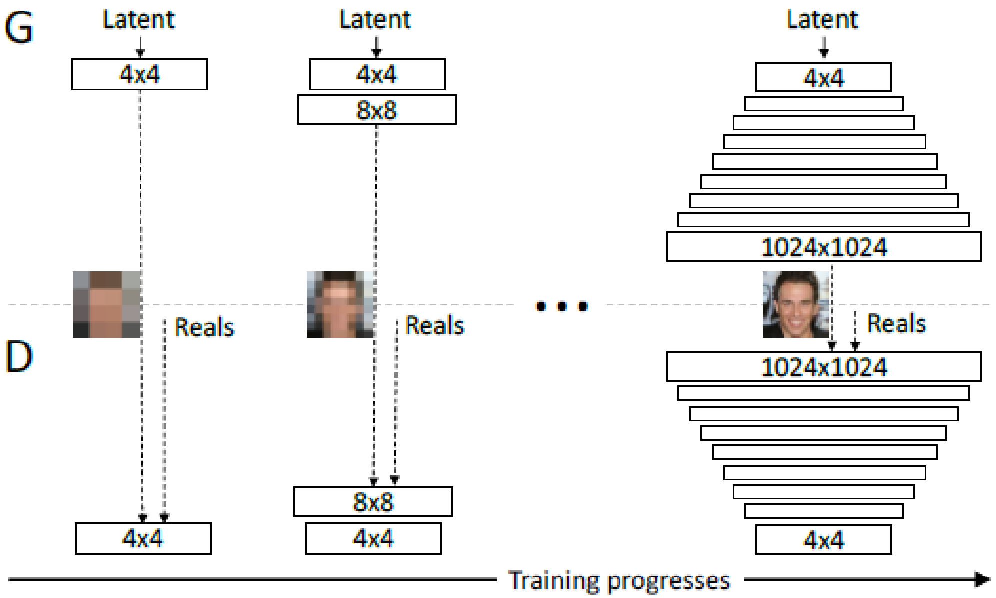


- However:
  - The training is fragile (alternating optimization), mode collapse
  - Did not work well for high-resolution (until recently)

High resolution GANs

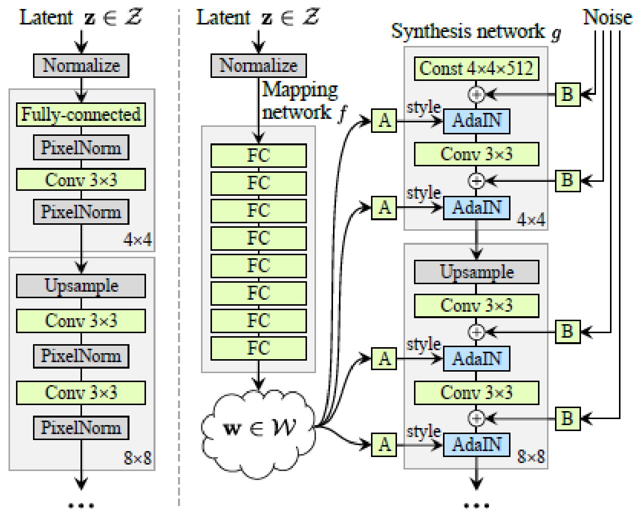
- Synthesis of 1024x1024 face images [Nvidia-ProGAN-2018]
- Trained from CelebA-HQ dataset 30k images
- Progressive training
  - Complete GAN for low-resolution (4x4)
  - Upsample, concatenate with res-net connections
  - Train everything end-to-end



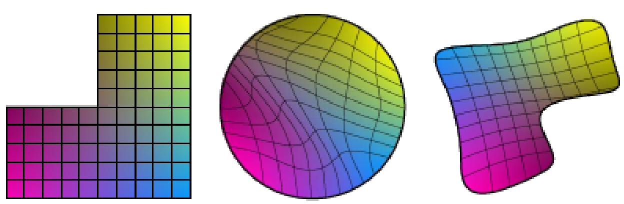


StyleGAN

- StyleGAN [Karras-2019] (NVidia)



(a) Traditional (b) Style-based generator

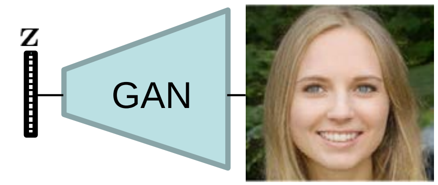


(a) Distribution of features in training set (b) Mapping from $\mathcal{Z}$ to features (c) Mapping from $\mathcal{W}$ to features

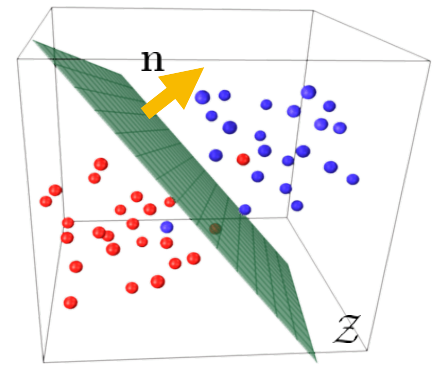
  – Multi-layer style transfer, training from 70k Flicker dataset, “hyper-realistic”

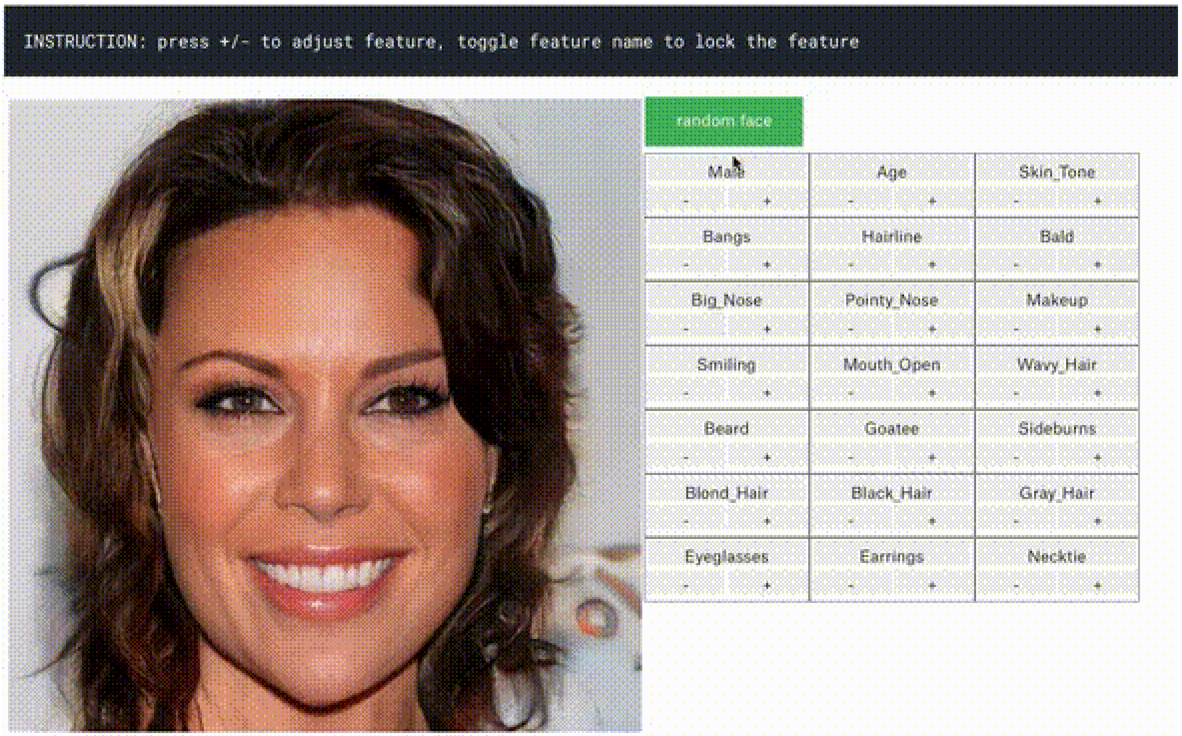
- Follow-up paper [Nvidia-2020, Nvidia-2021, Nvidia-2022]

GAN – latent space manipulation



- Every $\mathbf{z}$ from input distribution gives a realistic image
- Finding semantic direction in the latent vector space
  - Train a linear binary classifier on labeled set $(\mathbf{z}_i, y_i)$
  - Normal of the discriminative hyperplane is the semantic direction
- Semantic Editing / "Manipulation" $\mathbf{z} = \mathbf{z}_0 + \alpha\mathbf{n}$



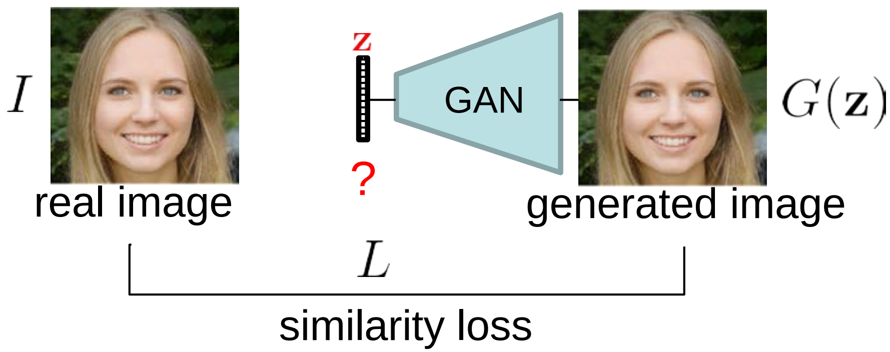


[demo]

[Abdal-SIGGRAF-2021]

GAN Inversion (projection)

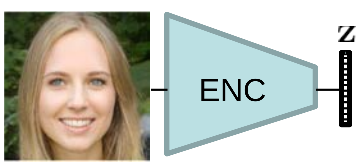
- Given an input real image *I*, find the latent code that generate the image



- Many approaches:
  - Direct optimization
    - Works, but slow (~10s)

$$\min_{\mathbf{z}} L(I, G(\mathbf{z}))$$

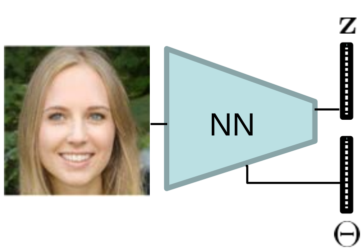
  - Encoders [Alaluf-2022, Tov-2021]
    - Very fast, but less accurate



  - Pivotal tuning [Roich-2021]
    - Optimizes the GAN model parameters

$$\min_{\mathbf{z},\Theta} L(I, G(\mathbf{z}, \Theta))$$

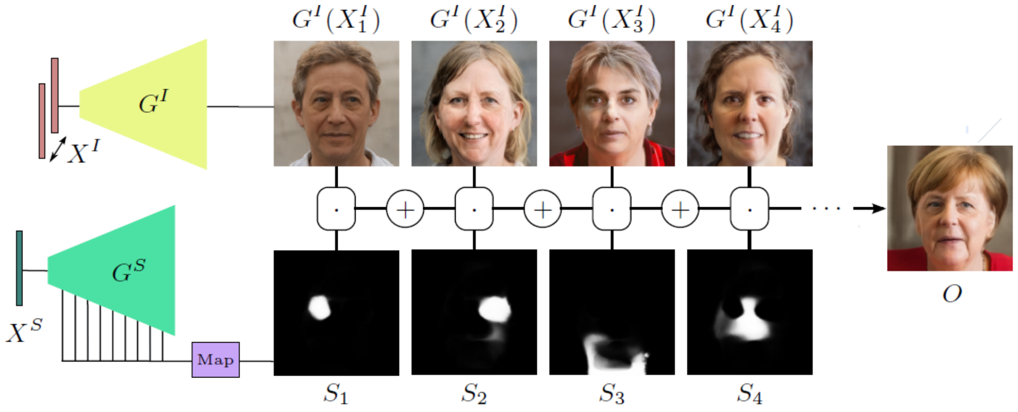
  - Hyper-networks [Alaluf-2022]
    - NN which adjust GAN model parameters



  - Piece-wise inversion [Šubrtová-2022]
    - More degrees of freedom, very accurate, but slow

GAN Inversion (projection)

- Chunky GAN – piecewise image inversion [Šubrtová-2022]



$$O(X^I, S) = \sum_{i=1}^{n} G^I(X_i^I) \cdot S_i.$$

$$\min_{X^S, X^I} \mathcal{L}_{\text{LPIPS}}\Big(I, \sum_{i=1}^{n} G^I(X_i^I) \cdot G^S(X^S)_i\Big) + \\ + \lambda_{reg} \sum_{i=1}^{n} \|X_i^I - X_\mu^I\|_2^2.$$

  – Partial inversion, Interactive editing

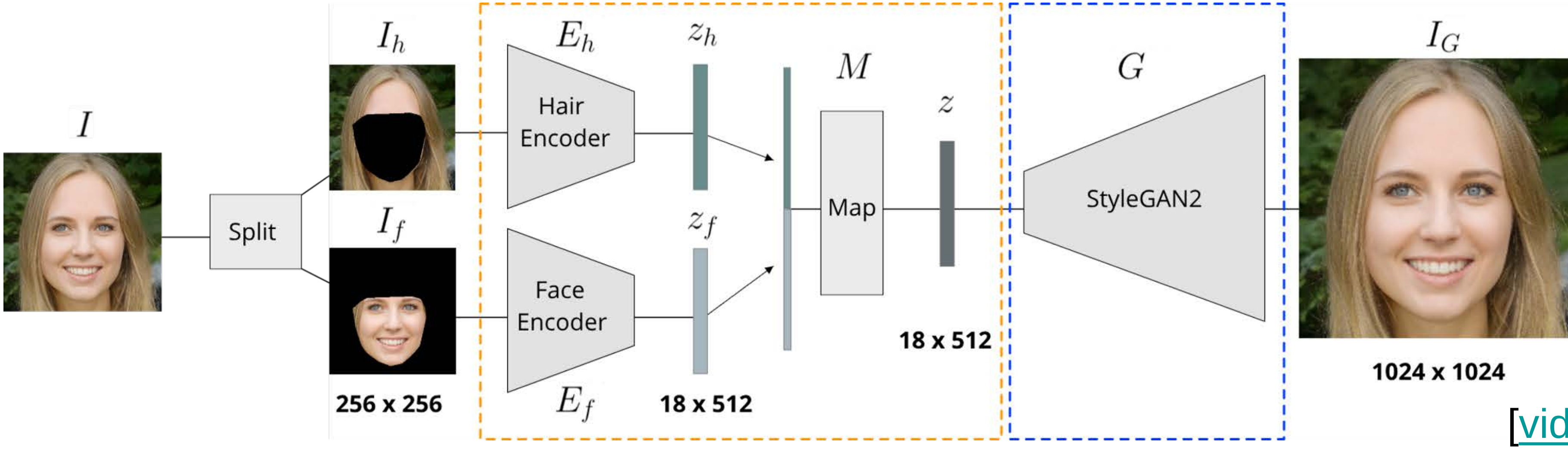
ChunkyGAN: Real Image Inversion via Segments

1 CTU in Prague, FEE 2 Adobe

Adéla Šubrtová*[1] David Futschik*[1] Jan Čech[1]
Michal Lukáč[2] Eli Shechtman[2] Daniel Sýkora[1]
* joint first authors

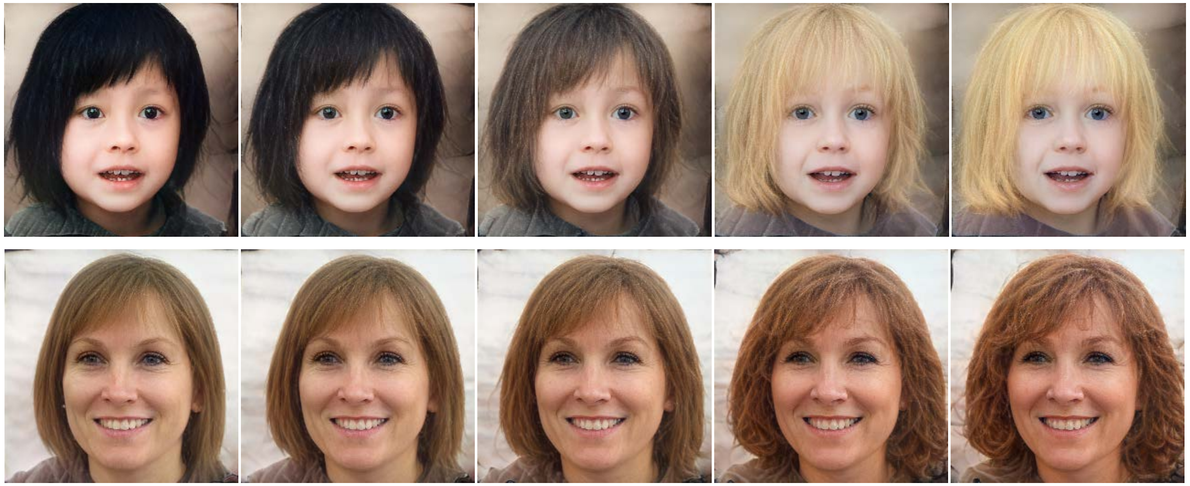
Hairstyle Transfer using StyleGAN

- Fully automatic hairstyle transfer, unaligned portraits [Šubrtová-FG-2021]



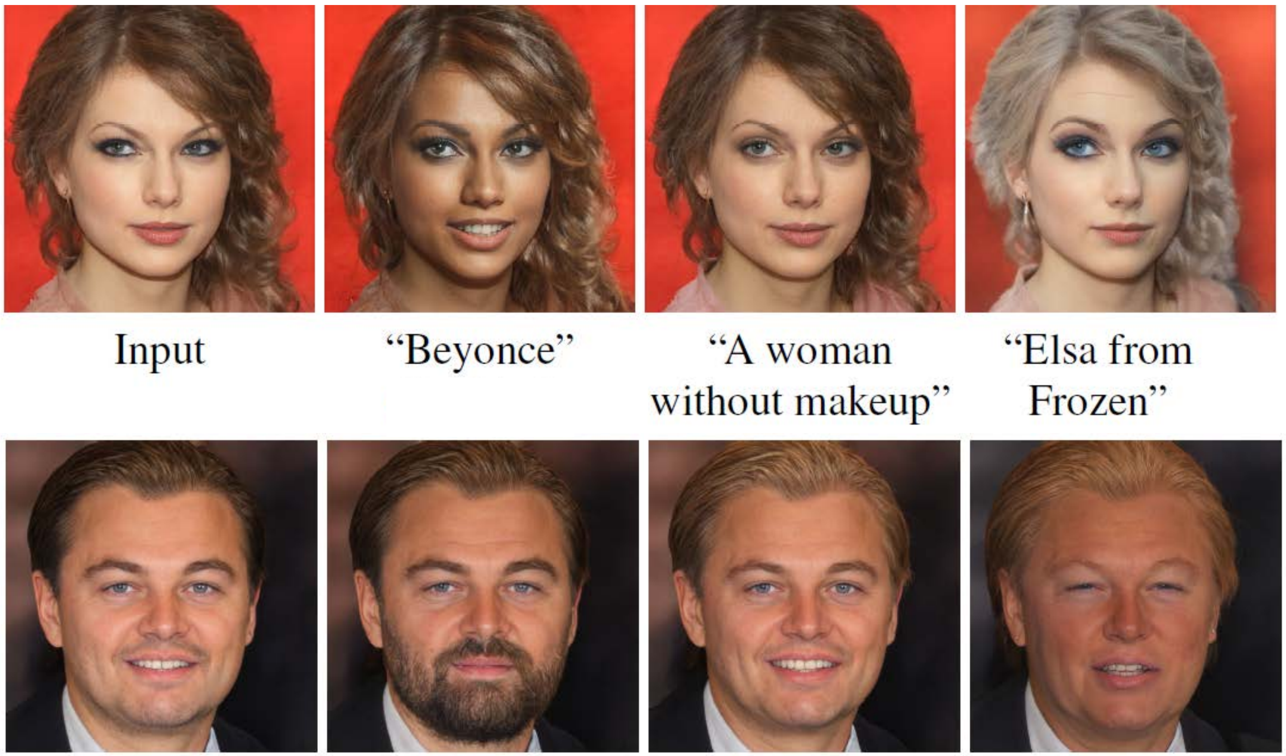
[video]

- Basic idea: Train two encoders (Hair, face) + fixed StyleGAN decoder
- Hairstyle interpolation, Editing in hairstyle latent space



Text-based Image Manipulation

- StyleCLIP [Patashnik-2021]
  – Text-Driven Manipulation of StyleGAN Imagery
  – Latent code manipulation driven by CLIP text-image similarity



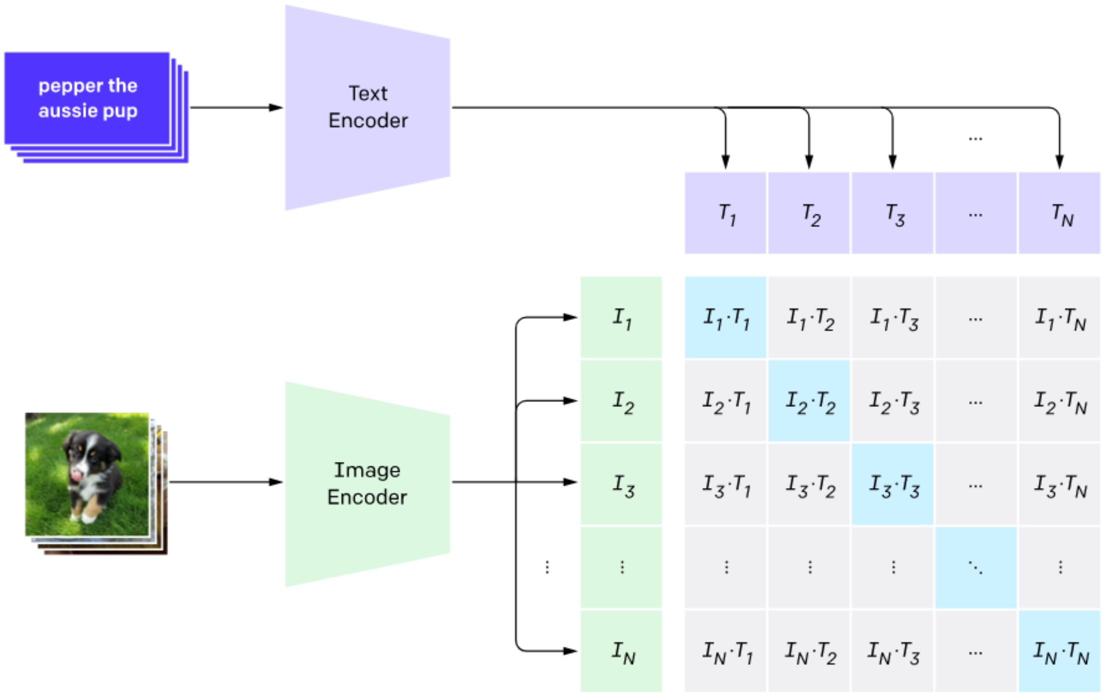
Input | "Beyonce" | "A woman without makeup" | "Elsa from Frozen"

Input | "A man with a beard" | "A blonde man" | "Donald Trump"

$$\arg\min_{w \in \mathcal{W}+} D_{\mathrm{CLIP}}(G(w), t) + \lambda_{\mathrm{L2}} \left\| w - w_s \right\|_2 + \lambda_{\mathrm{ID}} \mathcal{L}_{\mathrm{ID}}(w)$$

CLIP – Connecting Text and Images (recap)

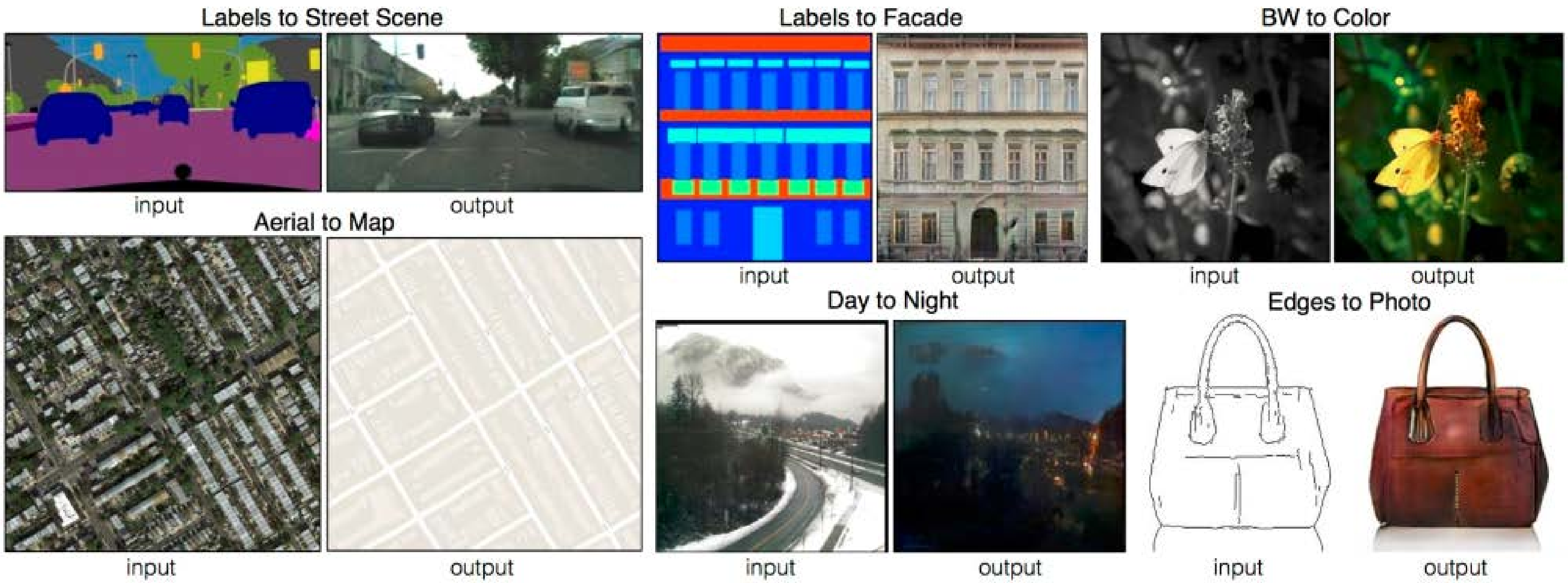
- CLIP [Radford-2021] by OpenAI
  - *"Contrastive Language–Image Pre-training"*
  - Learn joint text-image embedding => Text-image (cosine) similarity
  - Learned from 400M WebImageText (WIT) dataset



  - Zero-shot prediction (on par with Resnet on ImageNET benchmark)
    - Loop over ImageNET-classes: *max* CLIP(E_T(*"A photo of a* <*class*>*"*), $E_I(I)$)
  - Trained model publicly available

Image to Image Translation

- Transfer image between domains [Isola-2017]

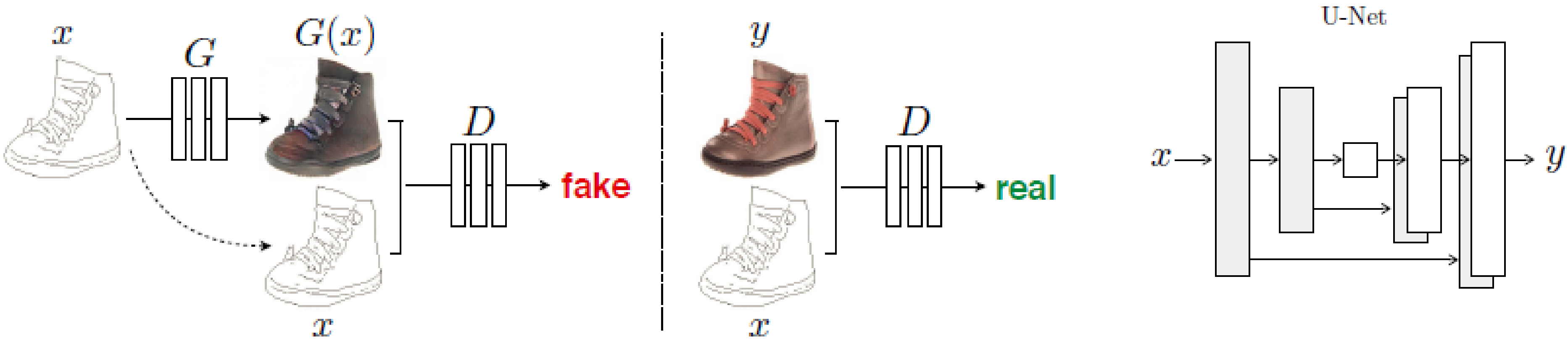


- Many applications [pix2pix], Super-resolution [Šubrtová-2018]

16x16 256x256 (predicted) 256x256 (ground-truth)

Image to Image Translation

- Combines fully convolutional net training with (conditional) GAN



$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

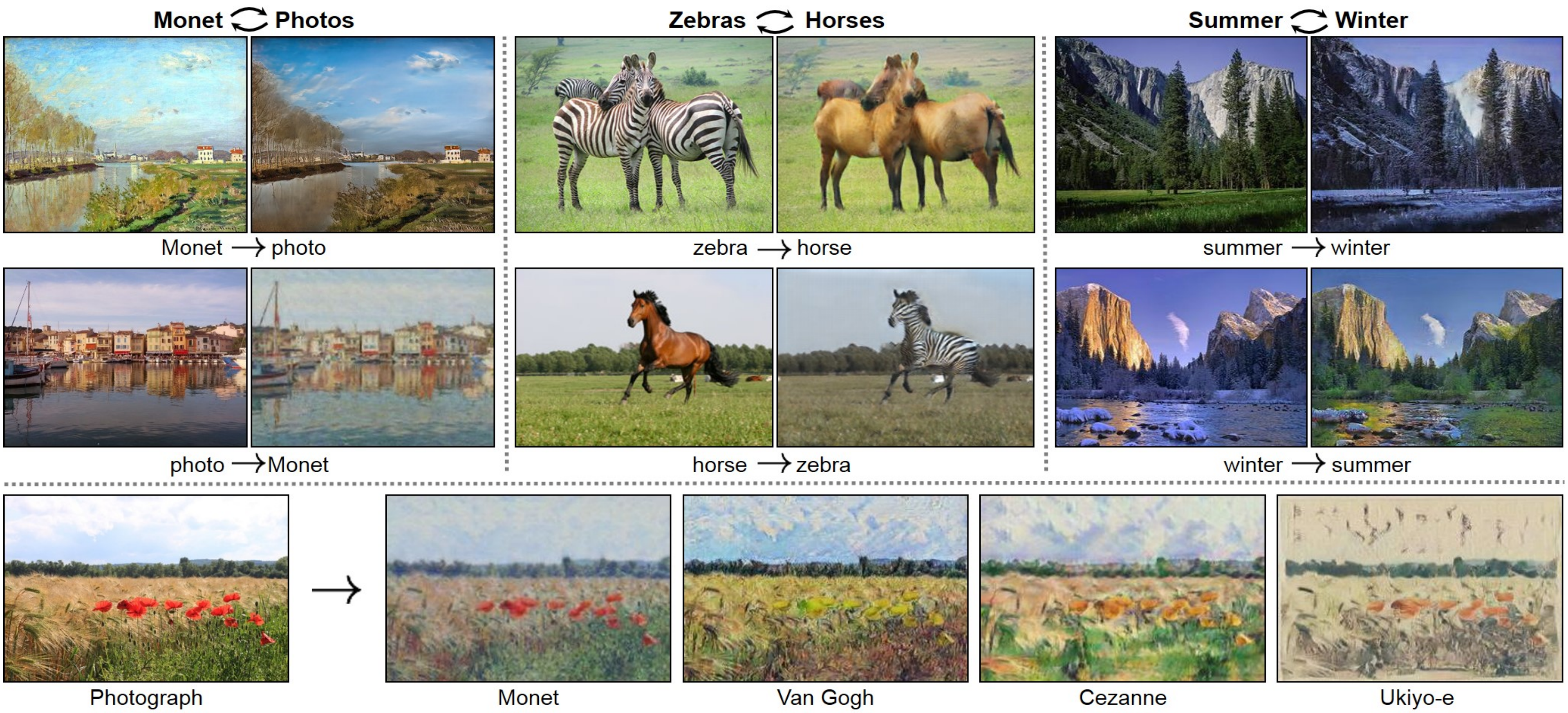
$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1]$$

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))]$$

- Difficulties with imposing variability (only via dropout when testing)
- Training needs pixel-to-pixel source and target image correspondences

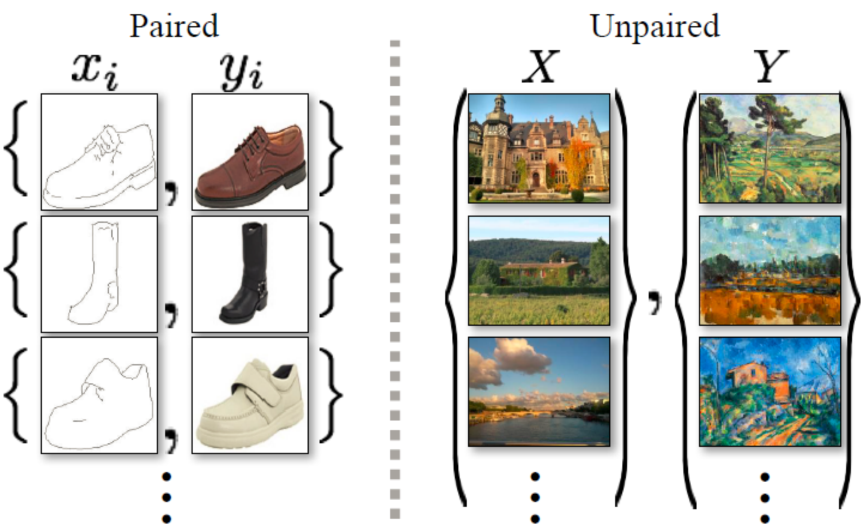
Cycle GAN

- Translating without pix-to-pix correspondences [Zhu-2017]



Cycle GAN

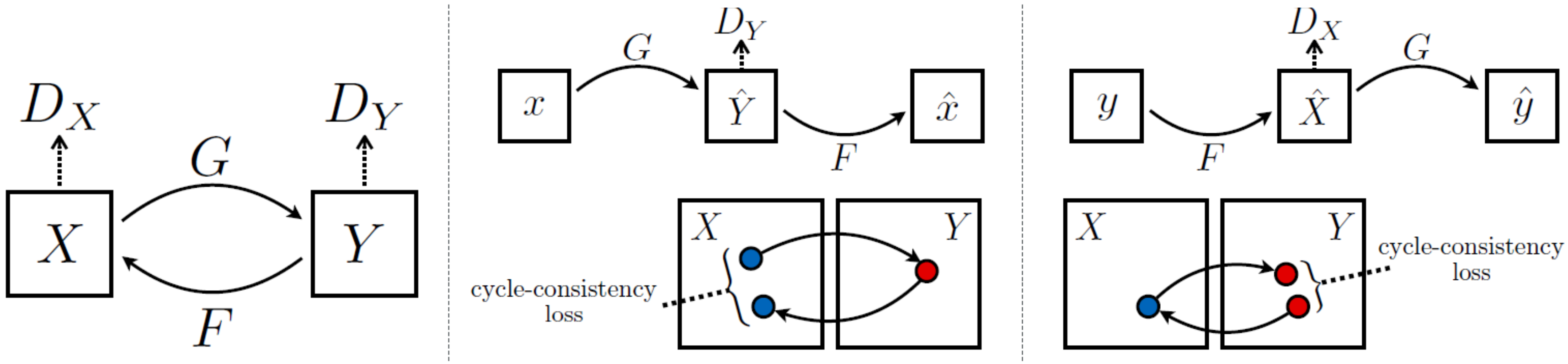
- Unpaired set of images to train the translation



$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda\mathcal{L}_{\text{cyc}}(G, F),$$

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y\sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$$

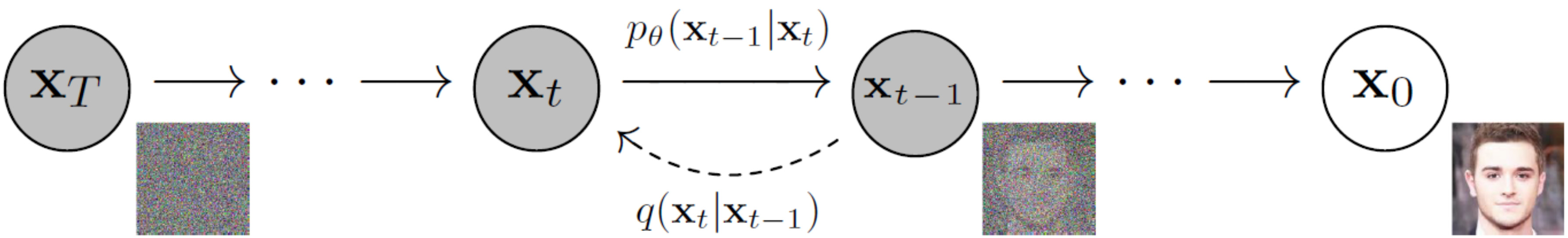
- Cycle consistency



Diffusion Models

Diffusion Models

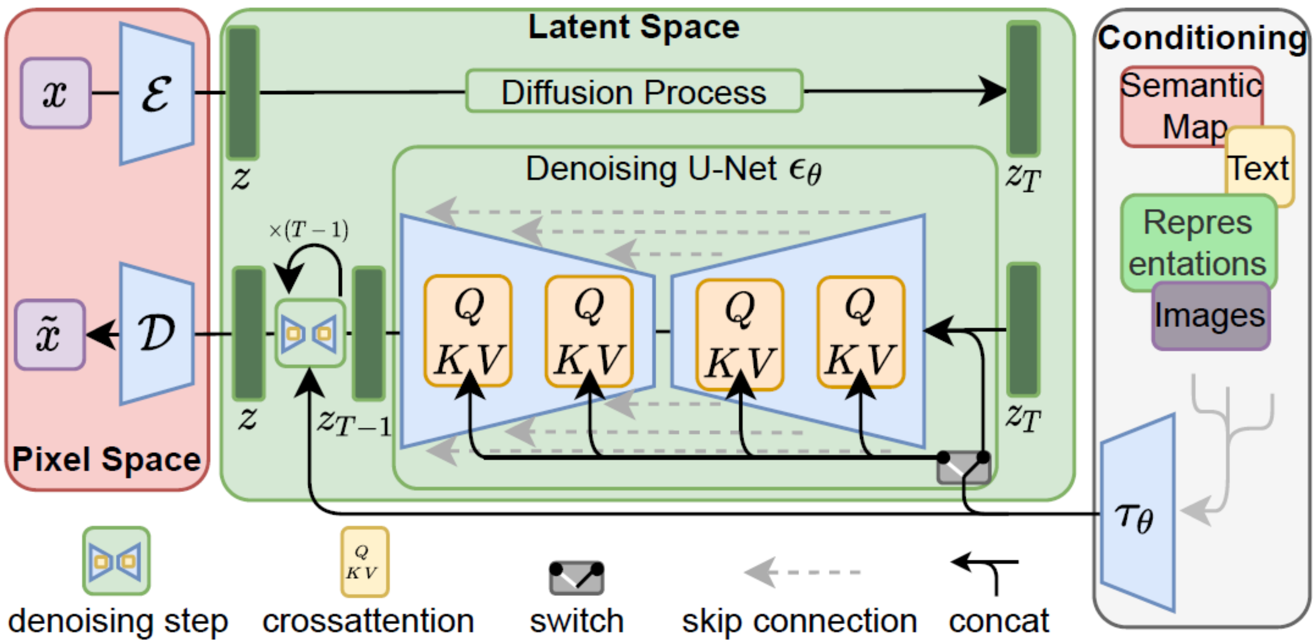
- Inspiration from thermodynamics [Sohl-Dickstein-2015]
- Denoising diffusion probabilistic models [Ho-2020]
- Main principle:
  - Forward diffusion: progressively destroy the data by injecting noise
    - Gaussian Noise
  - Reverse diffusion: learn to reverse the process to sample generation
    - denoising U-NET



$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

Stable Diffusion

- Stable Diffusion [Rombach-2022]
  - Latent diffusion model (diffusion/denoising runs in latent space)
  - Encoder/decoder pixel-latent space learnt offline
  - Conditioning by cross-attention (Text, Semantic maps, …)

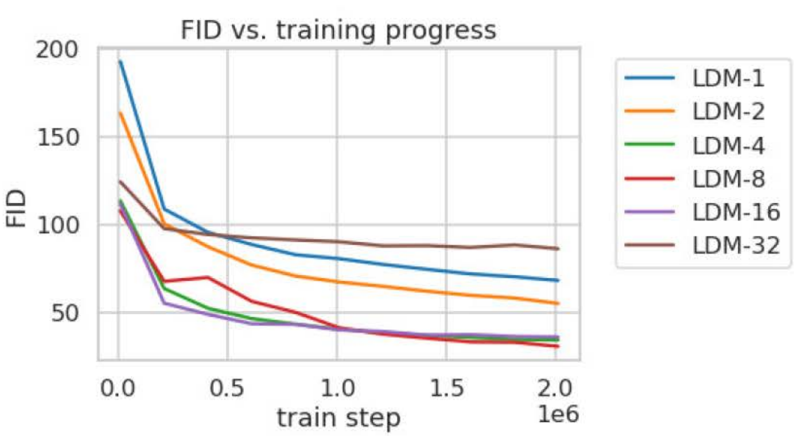


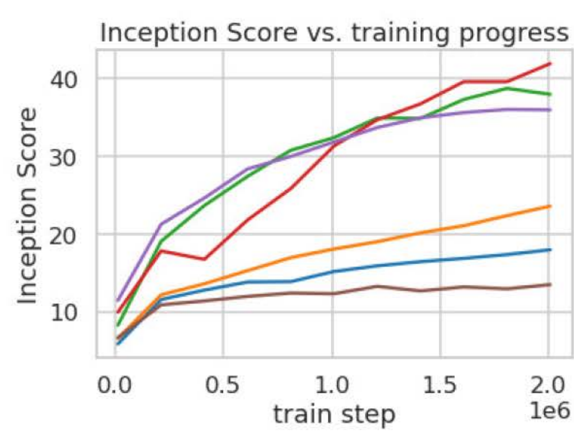
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon-\epsilon_\theta(z_t,t,\tau_\theta(y))\|_2^2\right]$$

Stable Diffusion

- Unconditioned generation (256x256)

  – Effect of the latent “compression”



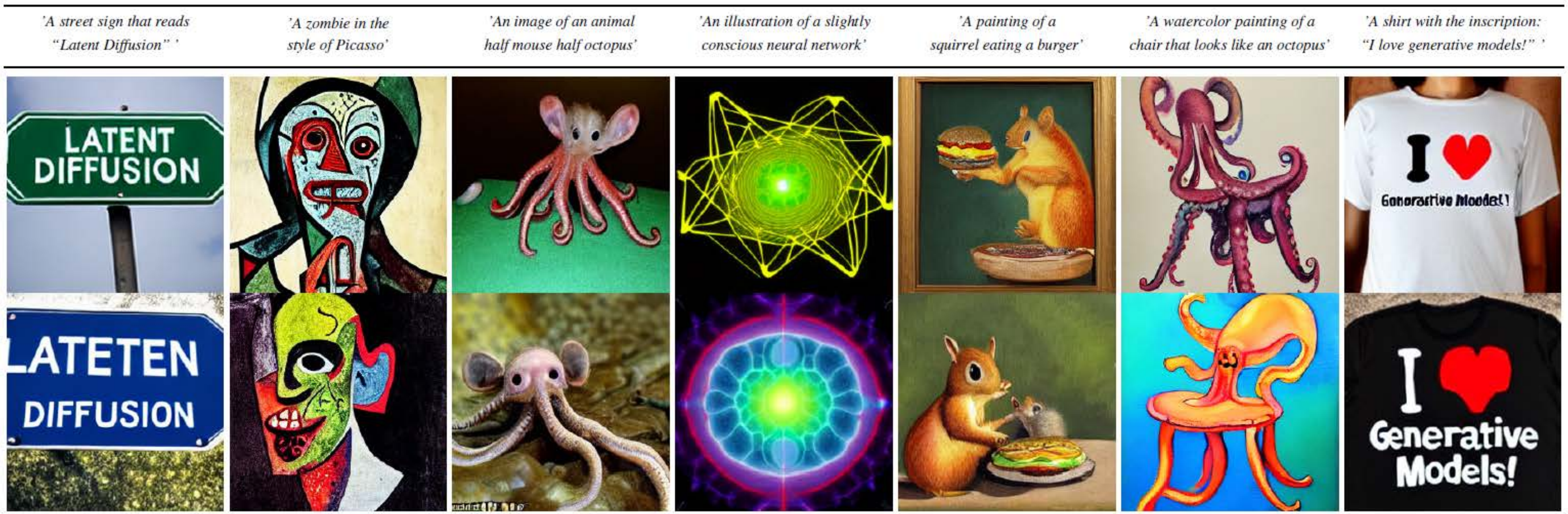


LDM-1
(pixel level)

LDM-8
(default)

Stable Diffusion

- Text-to-Image
  - Trained on LAION dataset - 1.45B
  - CLIP embedding for text



Stable Diffusion

- Semantic maps

Semantic Synthesis on Flickr-Landscapes [23]

- Inpainting

input	result

Classifier-Free Guidance (CFG)

- Classifier-free diffusion guidance [Ho-2022]
- Controls adherence with the input prompt
  - Trade-off mode coverage and sample fidelity

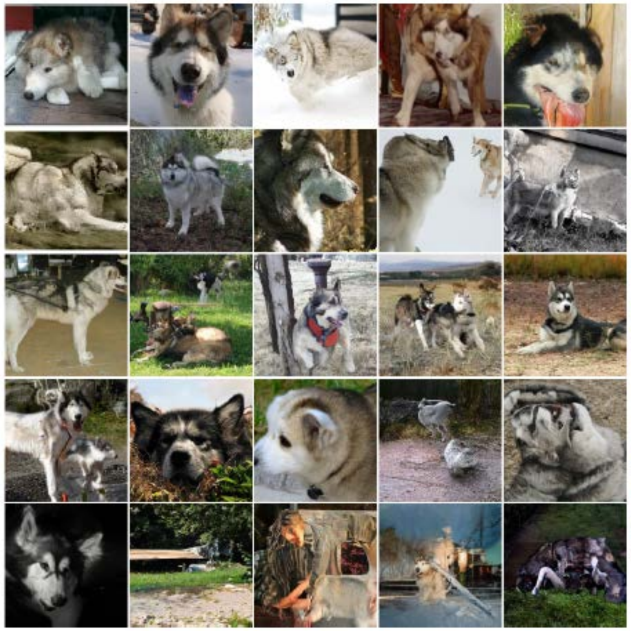
- Training: Diffusion model is trained both conditional and unconditional
  - Unconditional sample (null prompt) is given in certain probability p_{uncond} (e.g. 0.1/0.2/0.5)

- Inference:
  - Sample a weighted combination of conditioned and unconditional denoising model in each step

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda, \mathbf{c}) = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$$

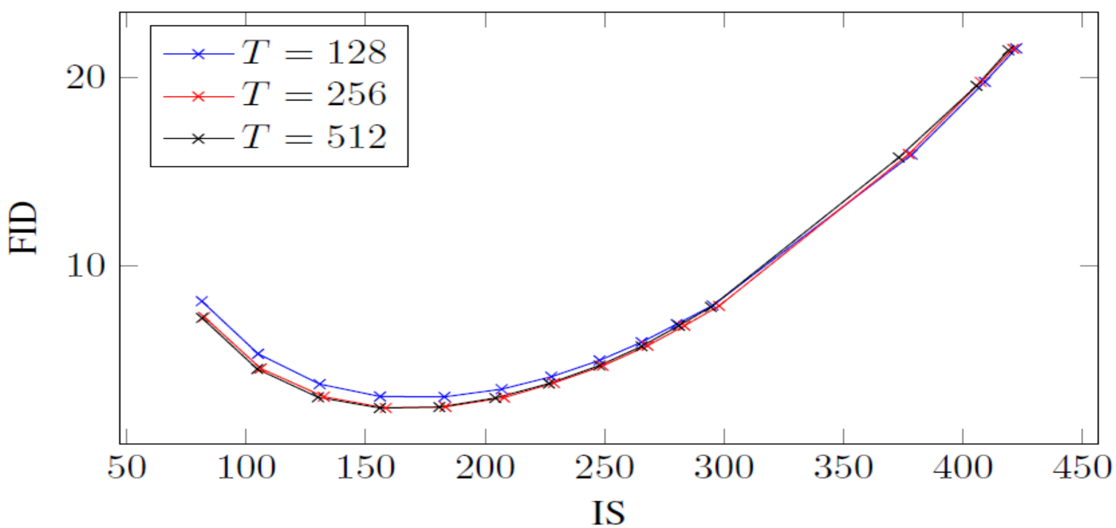
Classifier-Free Guidance (CFG)

$w = 0$



$w = 1$

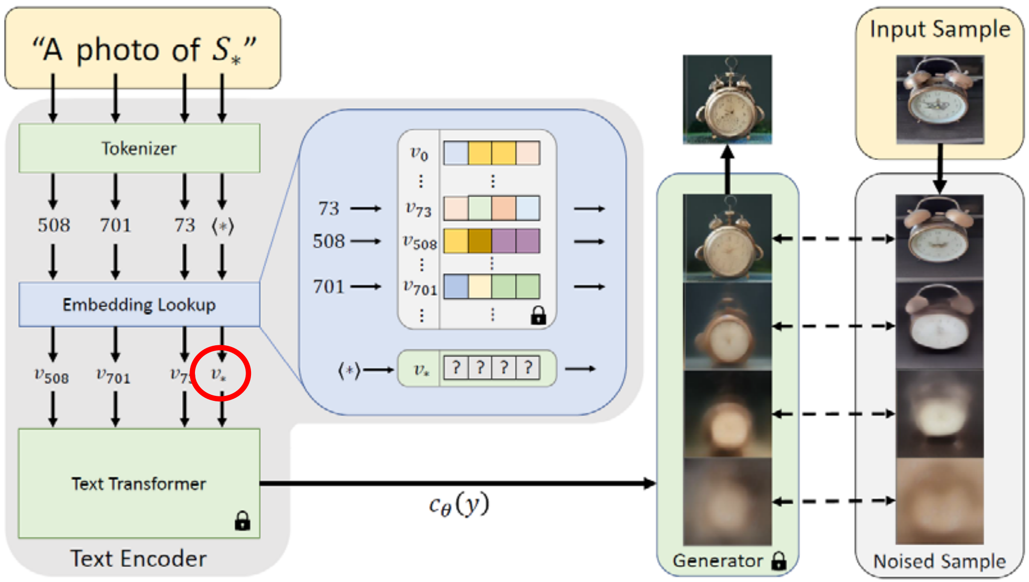
$w = 4$



Text2image using Textual Inversion

- An image is worth one word... [Gal-2022]
- Text to image with custom objects

Input samples $\xrightarrow{invert}$ "S_*" — "Painting of two S_* fishing on a boat" — "A S_* backpack" — "Banksy art of S_*" — "A S_* themed lunchbox"



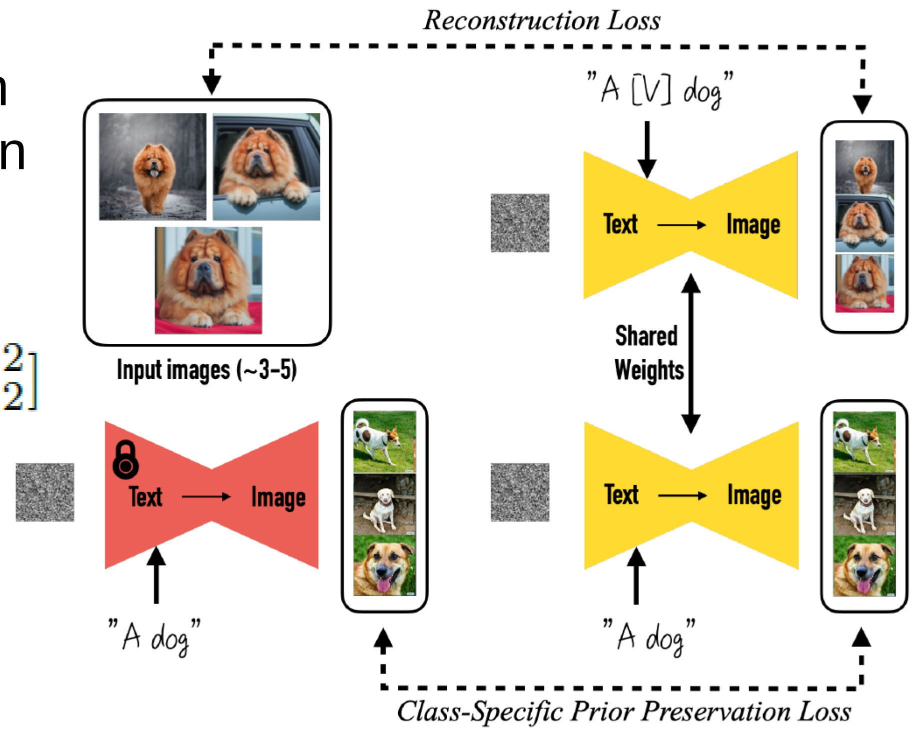
$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right]$$

DreamBooth

- DreamBooth [Ruiz-2022] (Google)
- **Fine-Tuning** Text-to-Image Diffusion Models for Subject-Driven Generation

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,\epsilon',t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\epsilon,\mathbf{c})-\mathbf{x}\|_2^2 + \lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\epsilon',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2\right]$$

  - about 5 mins on A100 GPU



Fine-tuning Diffusion Models

- LoRA (Low-Rank Adaptation) [Hu-2021] (Microsoft)
  - General light-weight adaptation of any models (including LLMs)
    - Faster training and storing
  - Many models available (e.g., CivitAI, PromptHero)

$$\mathbf{W}' = \mathbf{W} + \mathbf{BA}$$

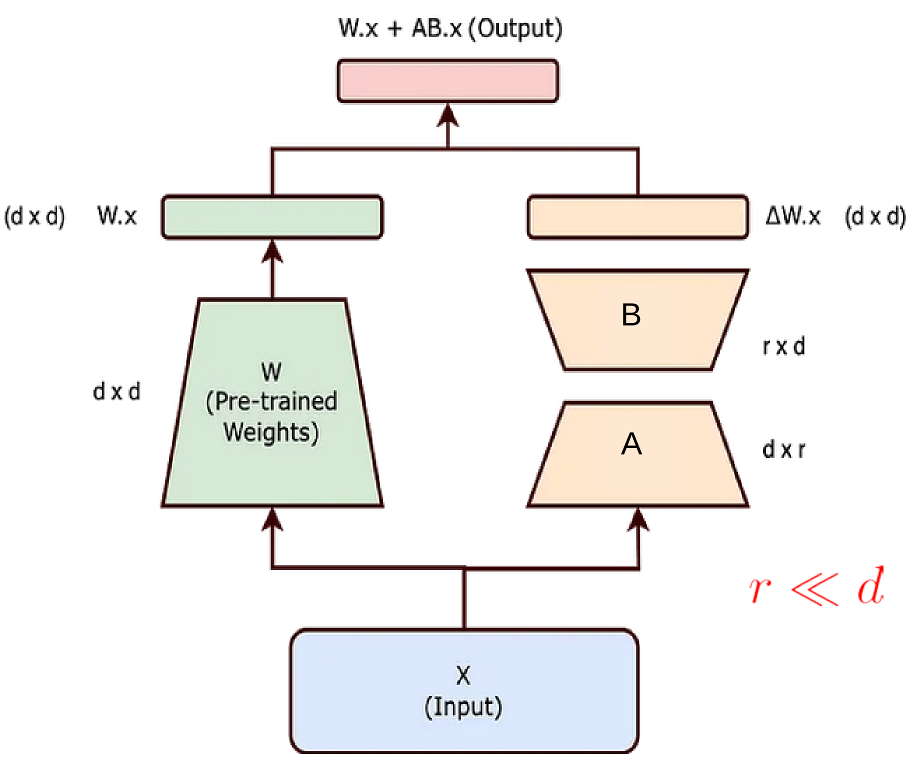
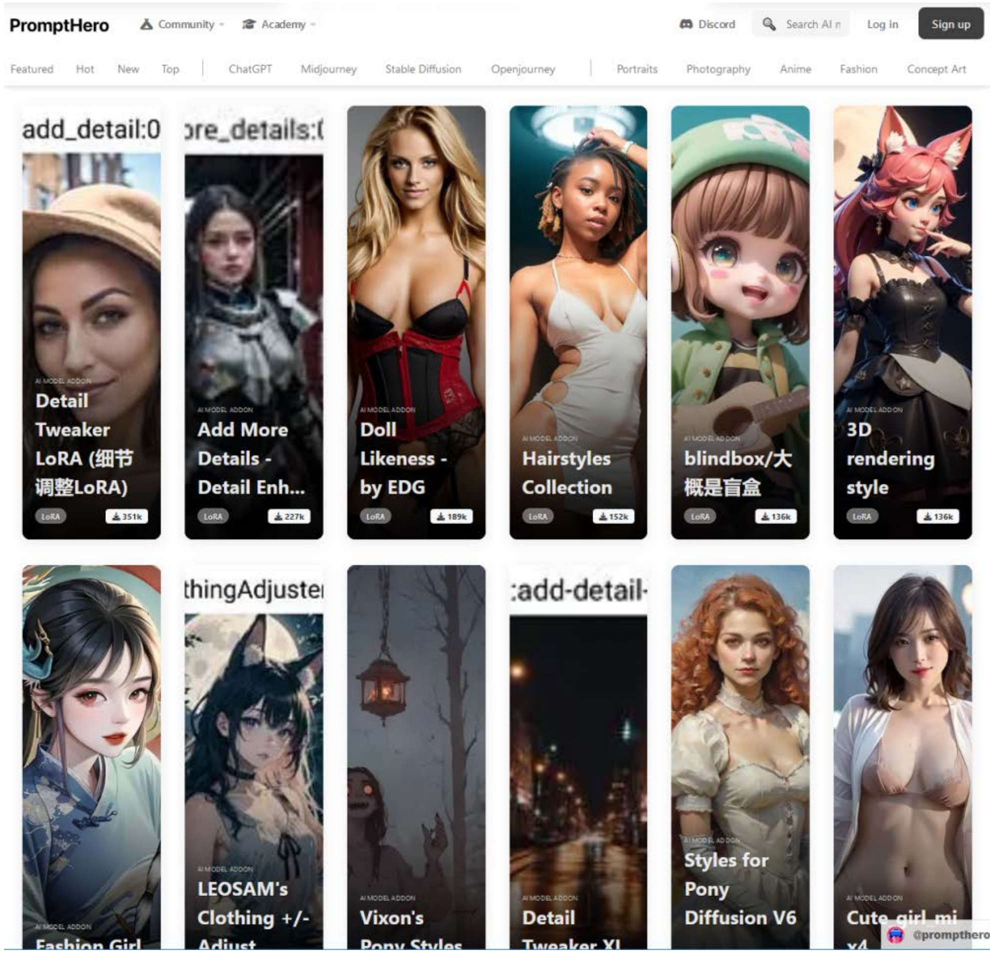
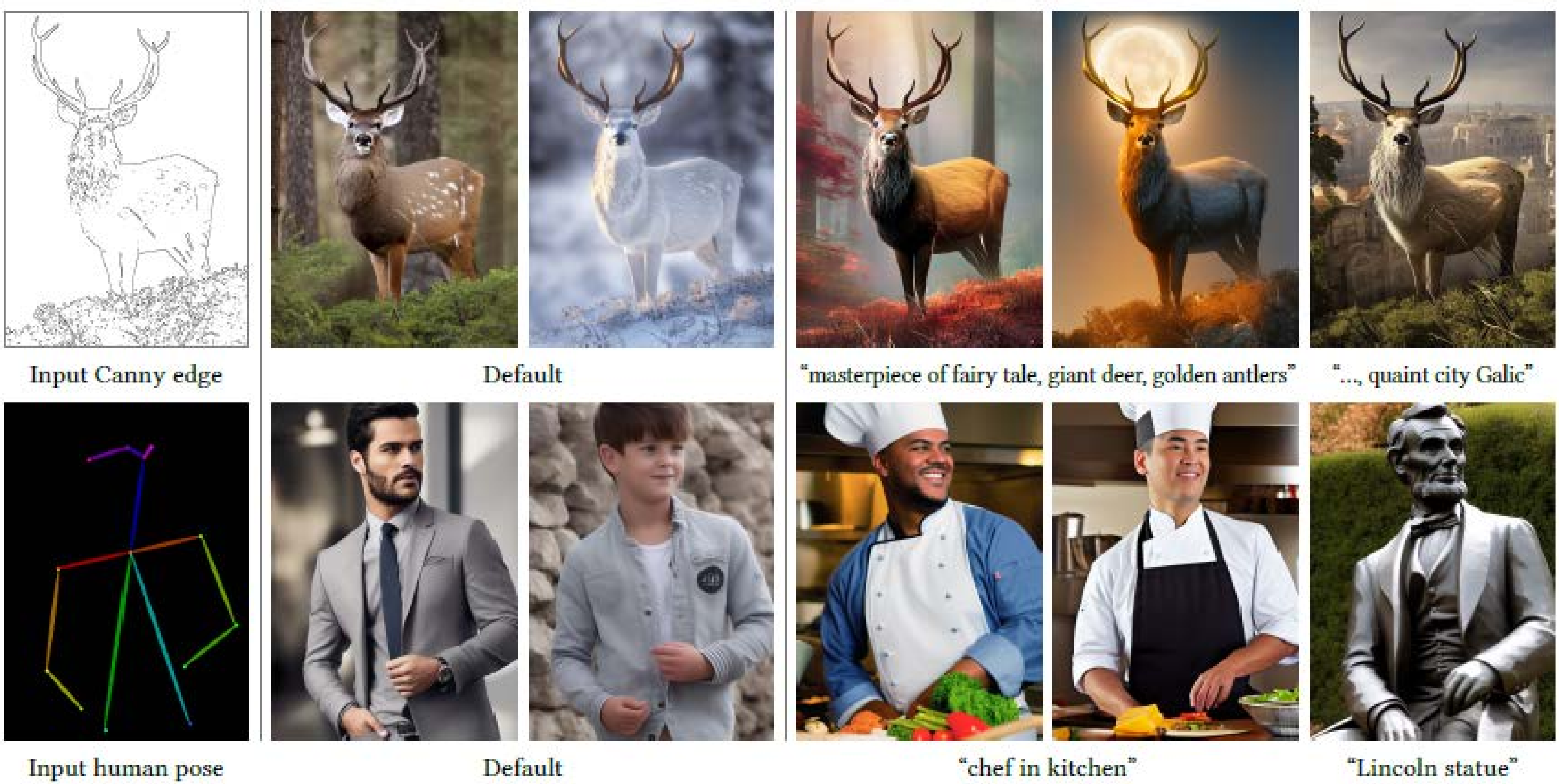


Figure credit Bhavin Jawade



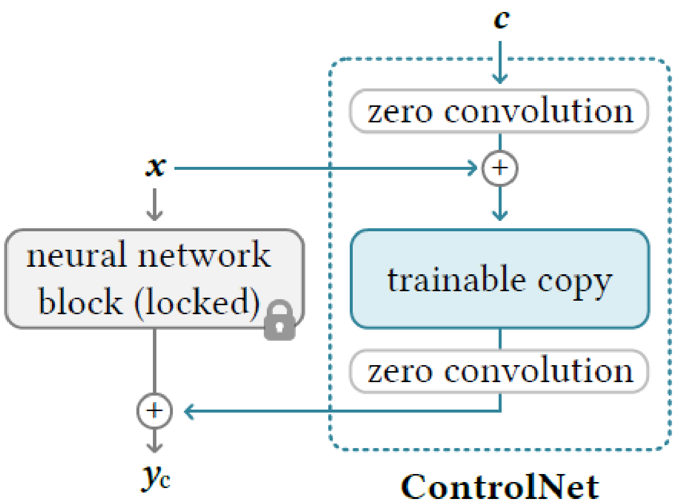
ControlNet

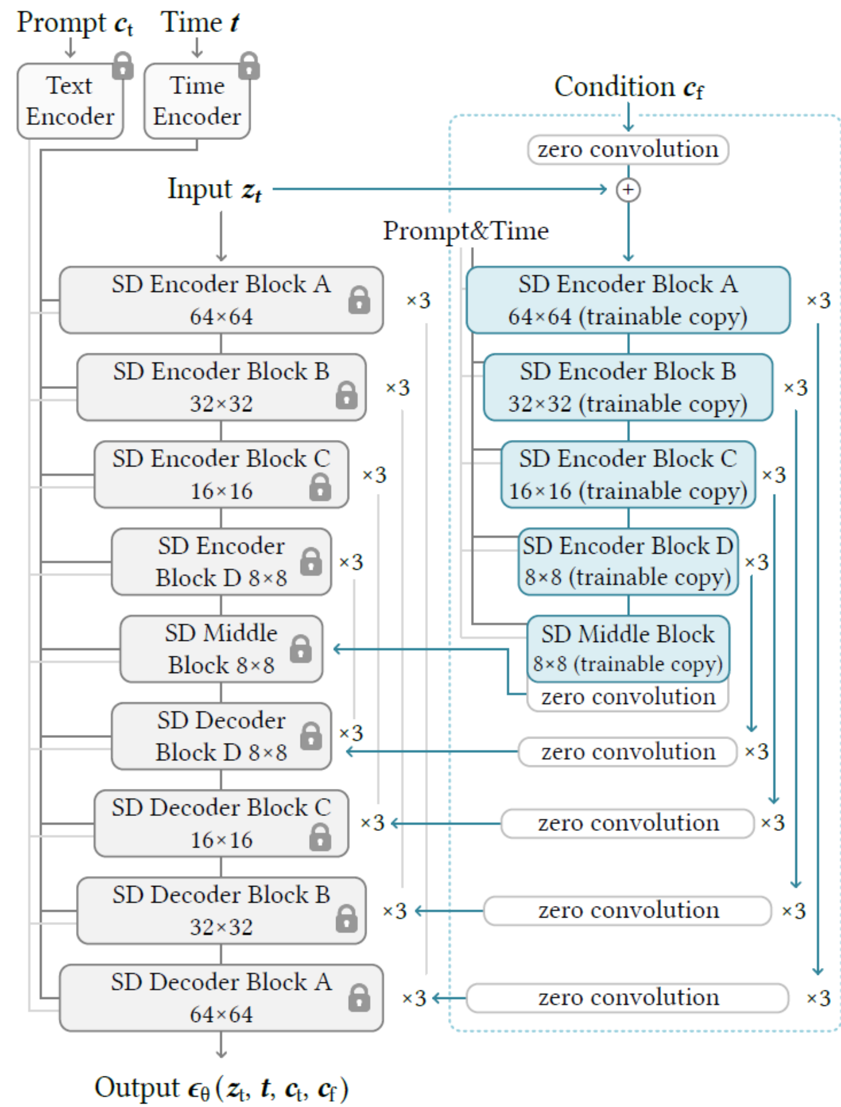
- Adding Conditional Control to Text-to-Image Diffusion Models [Zhang-2023]



ControlNet

- Fine-tuning with a trainable copy
  - Zero convolutions: 1x1 convolutions with weights initialized to zeros





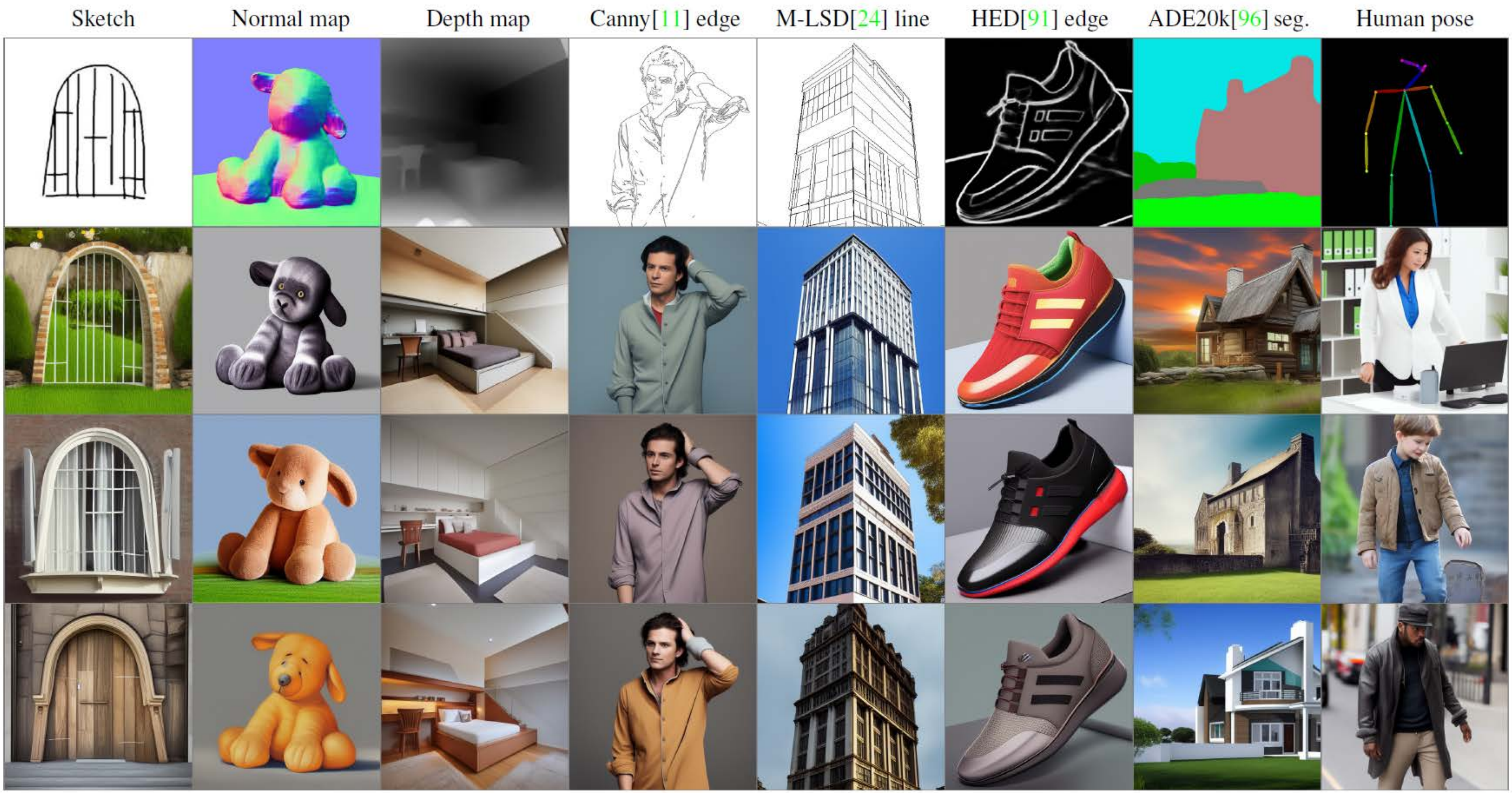
Extra cond.

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \Big[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}) \|_2^2 \Big]$$

  - 50% of text prompt c_t randomly replaced with empty string

ControlNet

- Models trained for multiple conditioning



- Demo, code and models available

Prompt2Prompt

- Prompt-to-Prompt image editing [Hertz-2022]

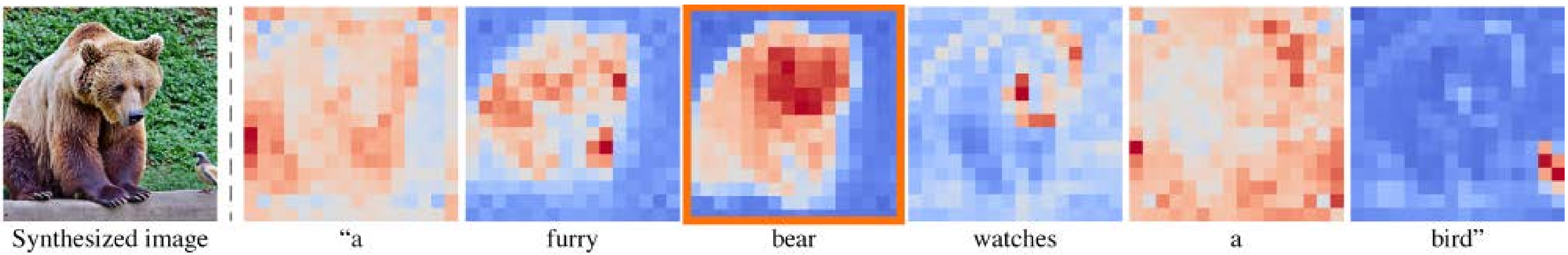
fixed random seed

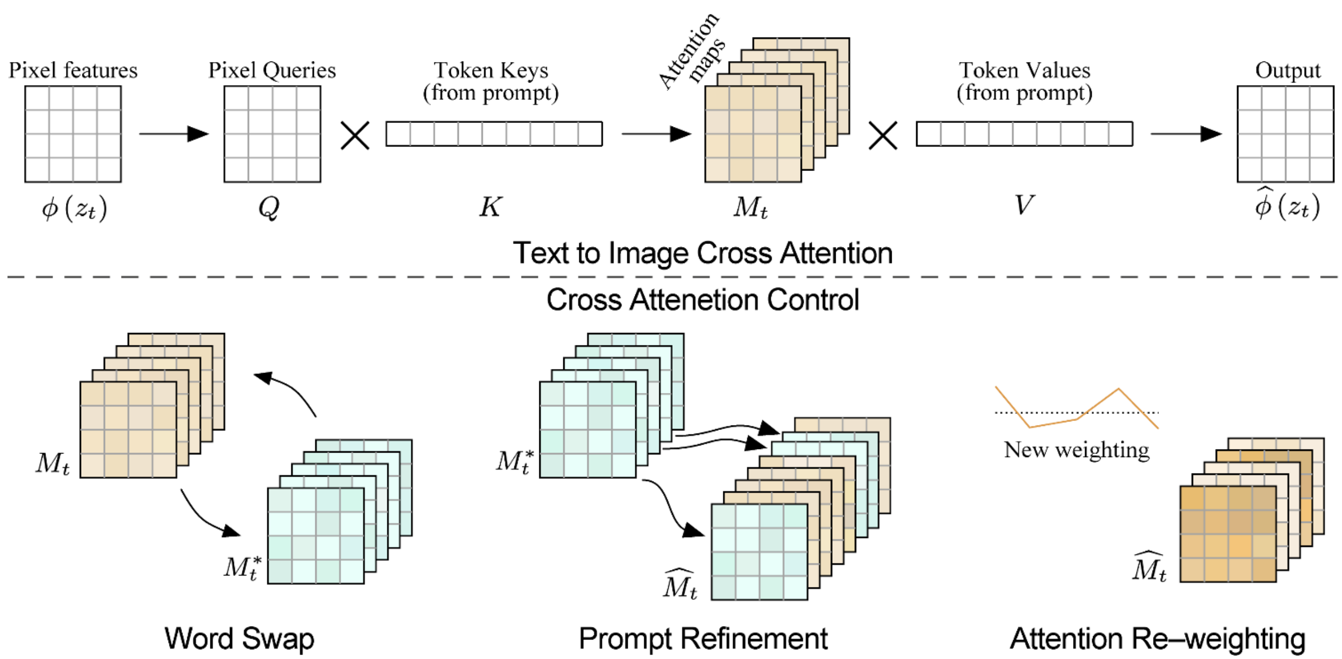
fixed random seed and attention maps

– Generated image editing
  - Stress/Weaken words, Changing words, Adding new phrases
  - Fixing random seed does not help (layout changes)

Prompt2Prompt

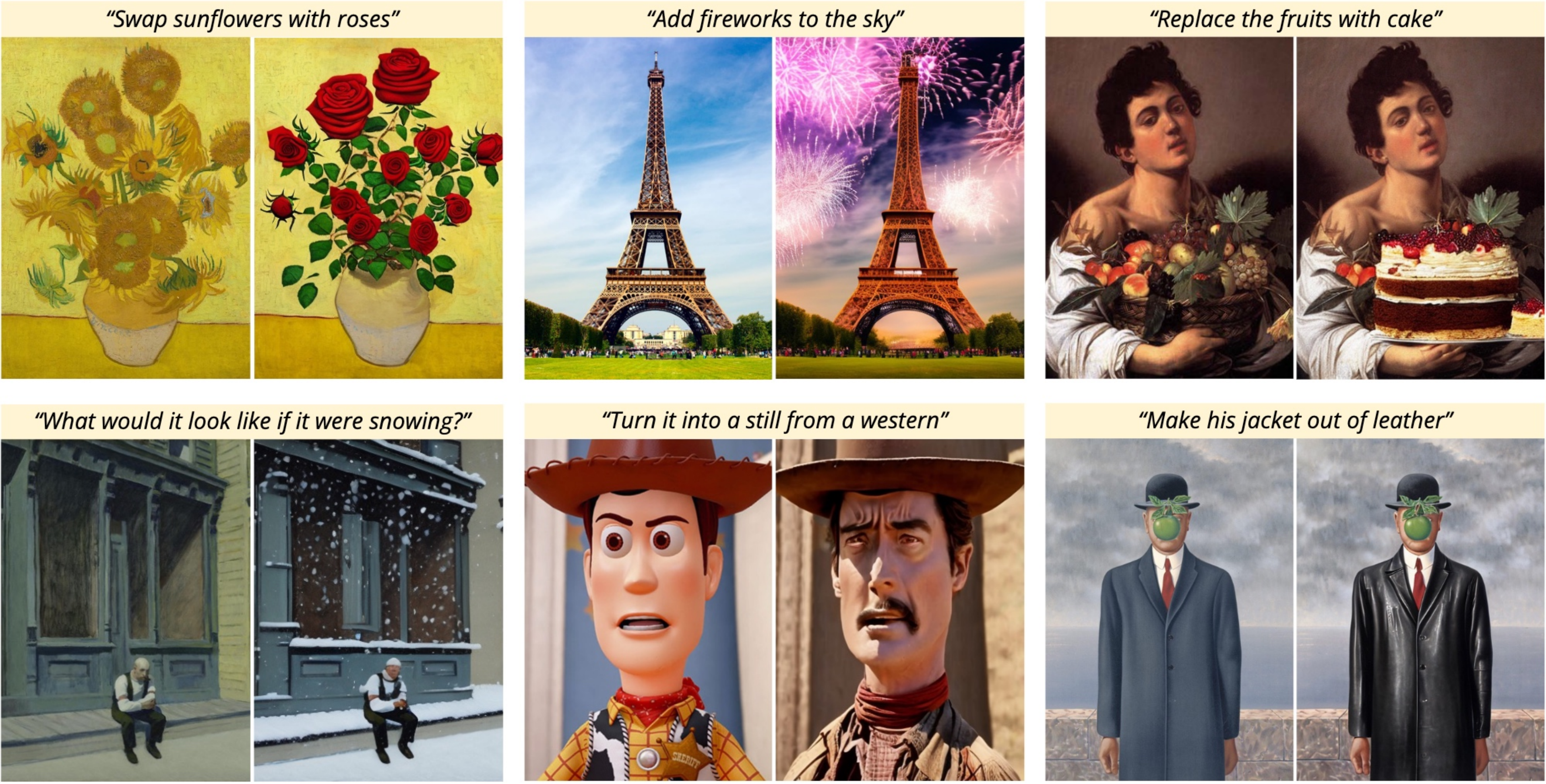
- Controlling the Cross-Attention (text x image)
  - Attention maps are responsible for spatial layout
  - Keep existing attention maps, change/add new word maps, reweight





Instruct Pix2Pix

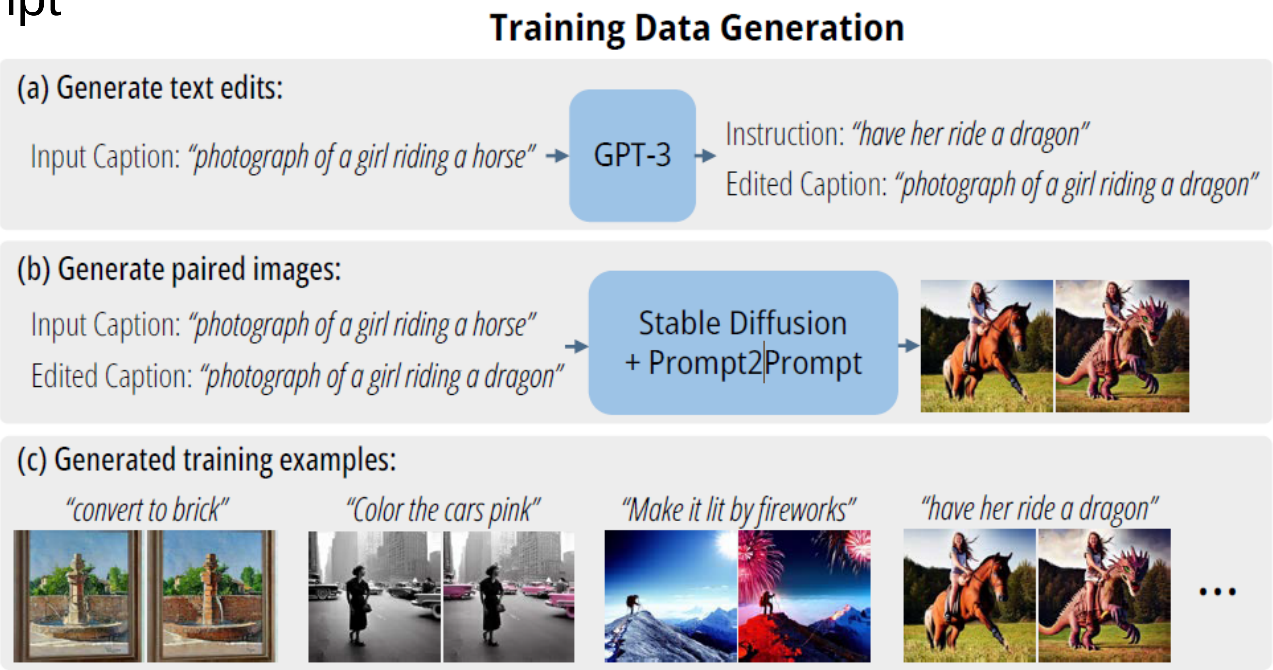
- Instruct Pix2Pix [Brooks-2023]
  - Textual editing or real images

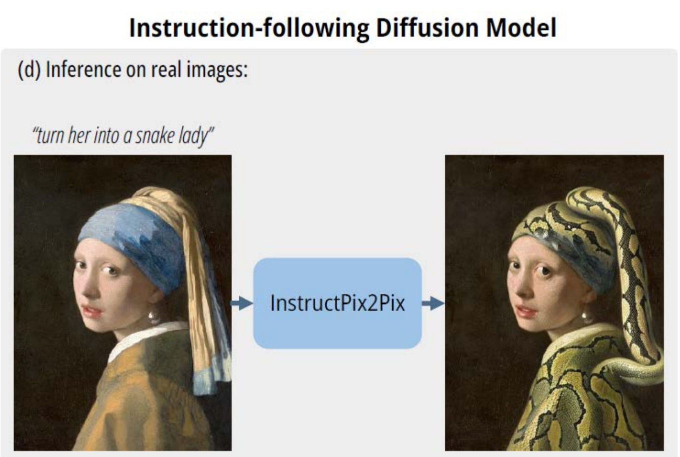


Instruct Pix2Pix

- New model trained from synthetic captions (Chat GPT) and images edited by Prompt-to-Prompt

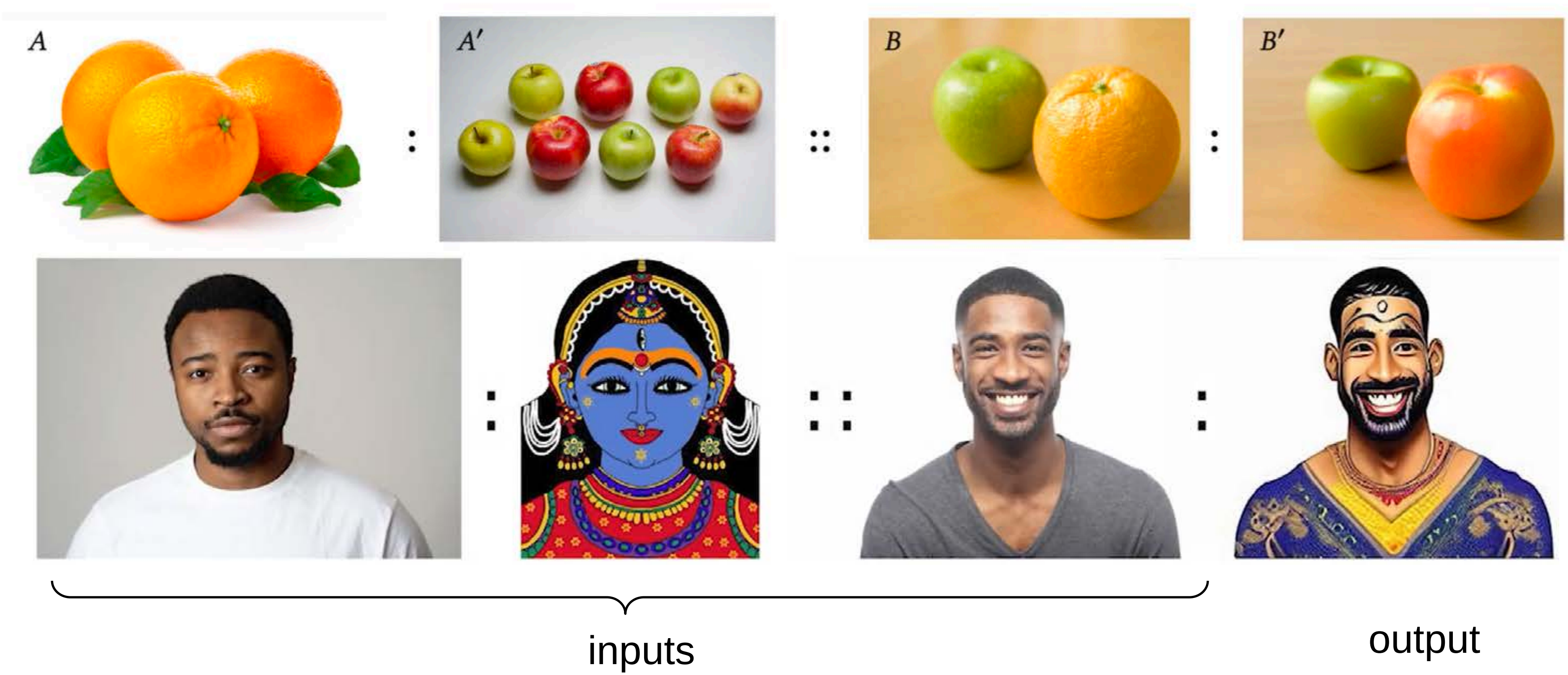
450k training examples





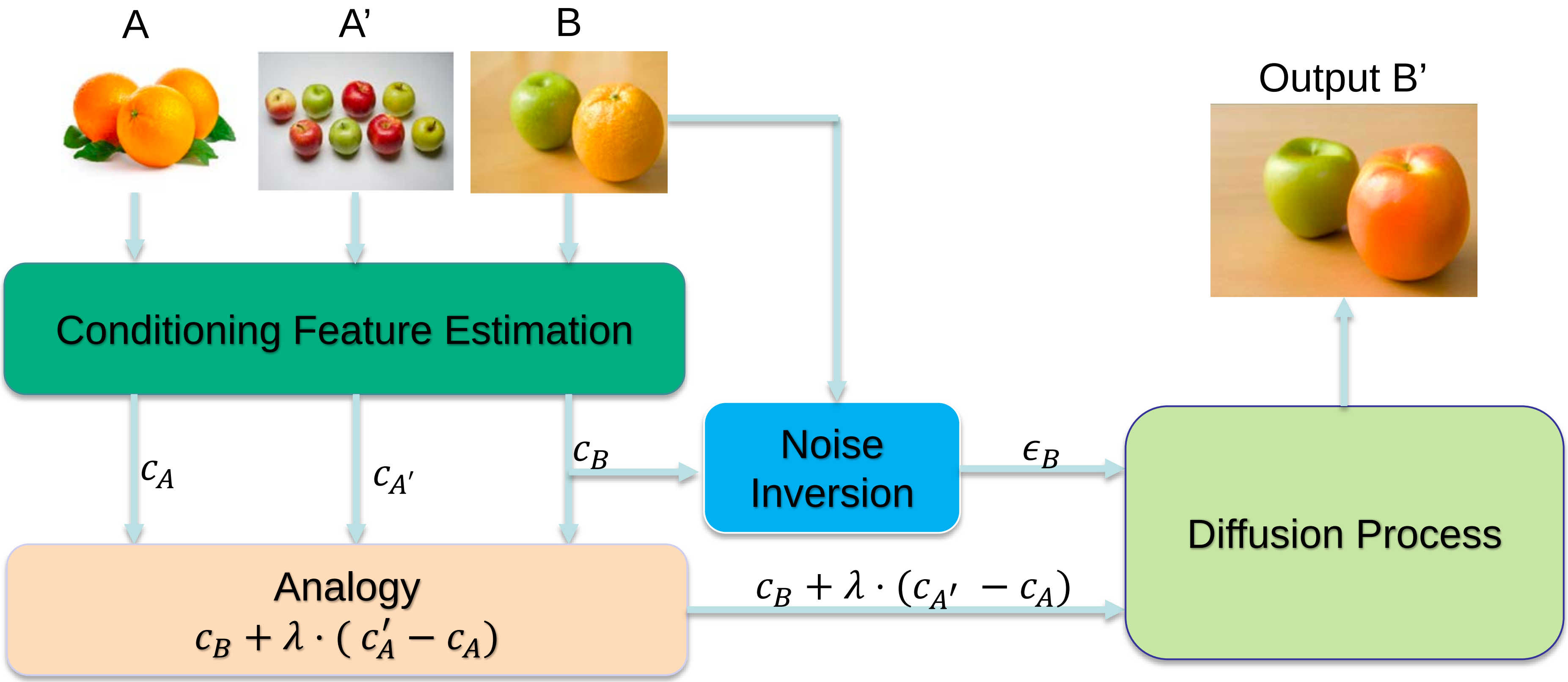
Diffusion Image Analogies

- Diffusion Image Analogies [Šubrtová-SIGGRAPH-2023]



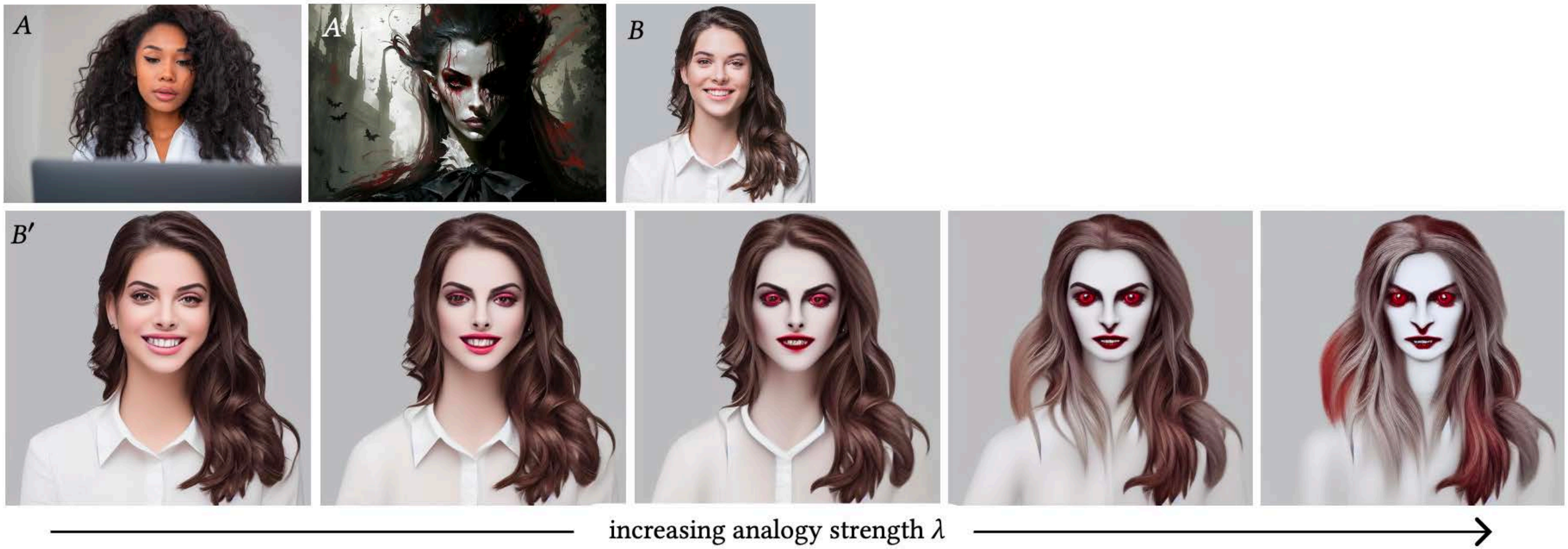
  - Real image editing by visual analogies
  - Relation between B' and B is analogous to relation between A' and A
  - Exploiting “algebra” of latent space of the Stable Diffusion

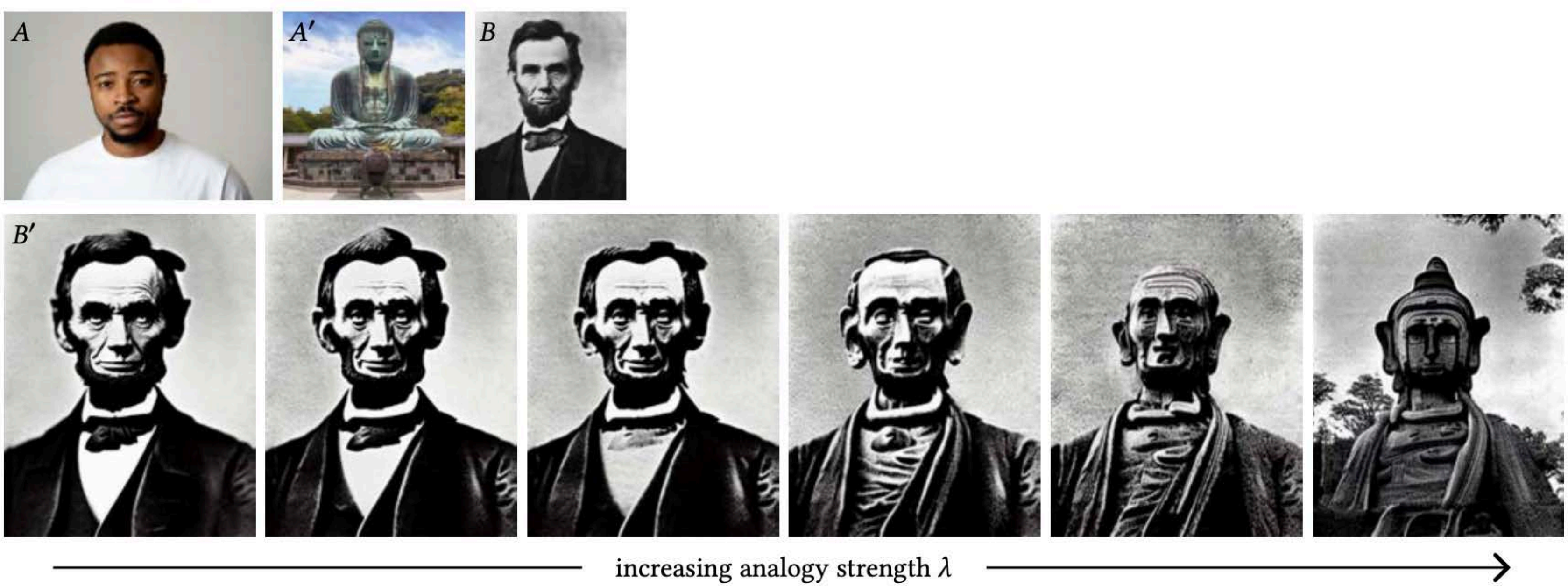
Diffusion Image Analogies



- Conditioning and Noise Inversion found by optimization
- Parameter λ controls the strength of the analogy

Diffusion Image Analogies





Conclusions

- "Photorealistic" quality has been achieved in the last couple of years
  - GANs, Diffusion models
- Deep generative models have evolved dramatically recently
- Interest / Fear of artistic community
  - Many tools greatly support content creativity
  - Certain creative artists feel endangered and exploited (some models likely trained on data without author's permission)
- Threat of high-quality deep fakes easy to perpetrate
  - Fake news, fake porn, etc.