

Interpreting Deep Generative Models for Interactive AI Content Creation

Bolei Zhou, The Chinese University of Hong Kong

Tutorial on Interpretable Machine Learning for Computer Vision at CVPR 2021

Progress for Image Generation

2014



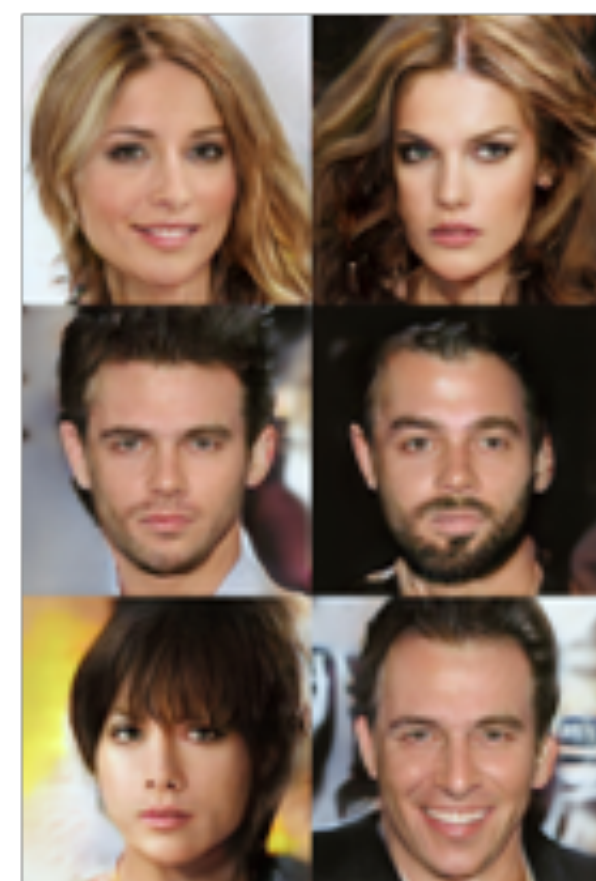
GAN

2015



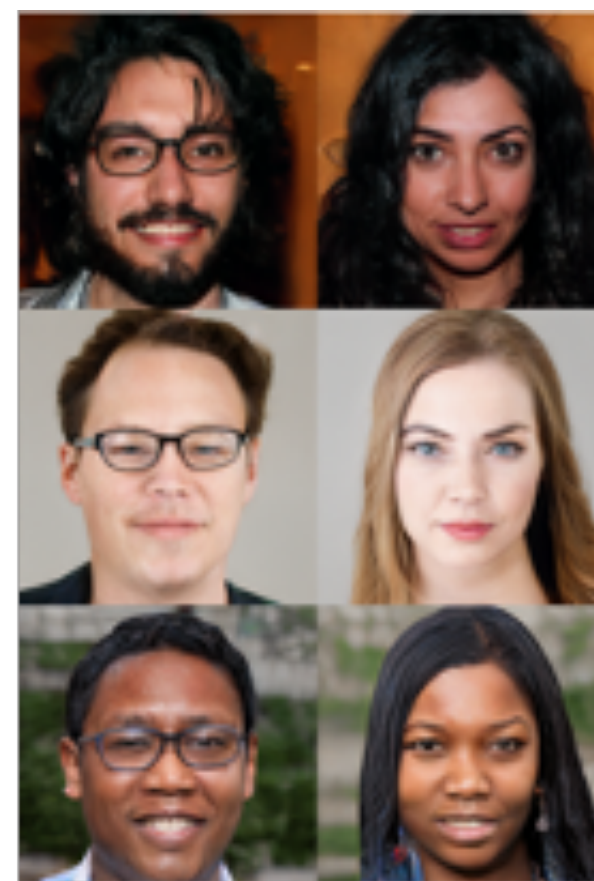
DCGAN

2017



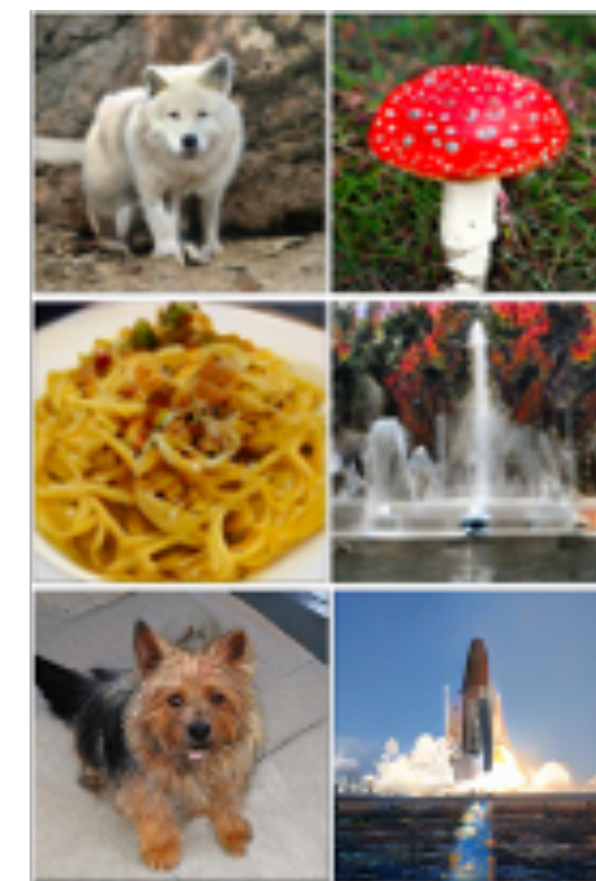
PG-GAN

2018



StyleGAN

2018



BigGAN

2019



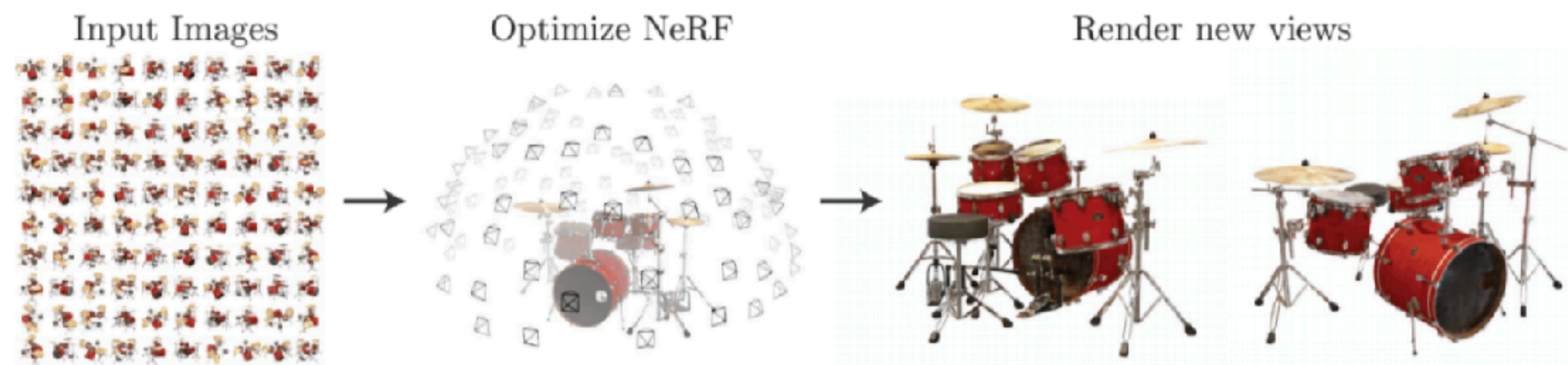
StyleGANv2

2020



StyleGAN-ADA

2020: NeRF (Neural Radiance Fields)



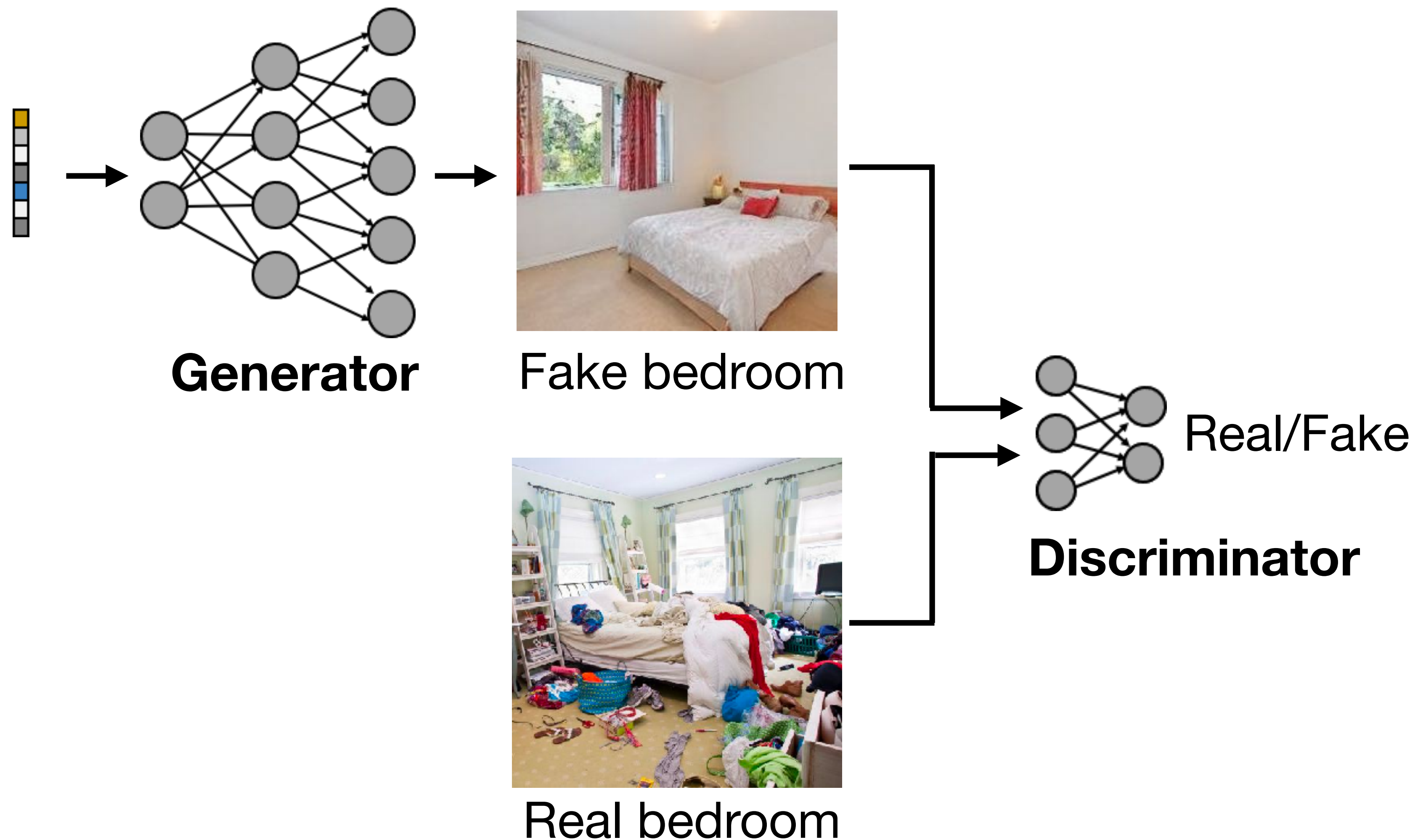
2021: OpenAI DALLE (VQ-VAE)

Arm chair in shape of avocado

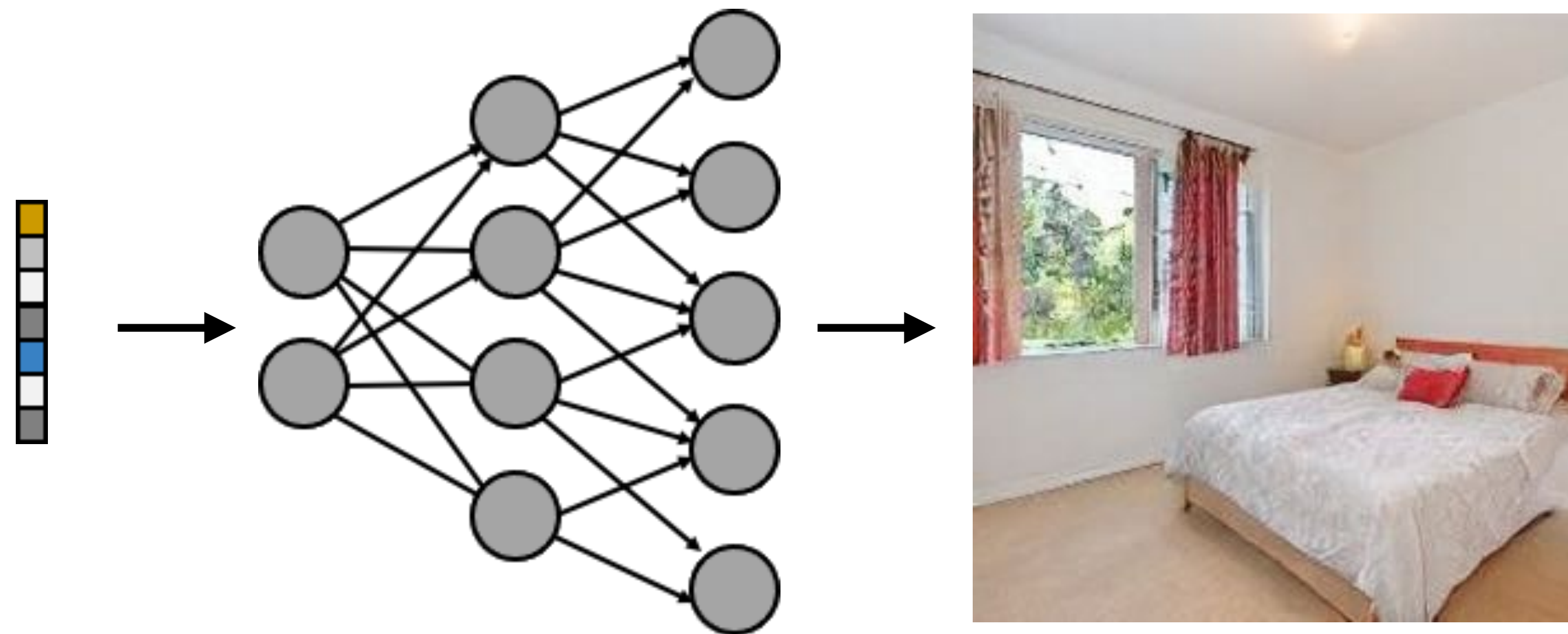


Generative Adversarial Networks (GANs)

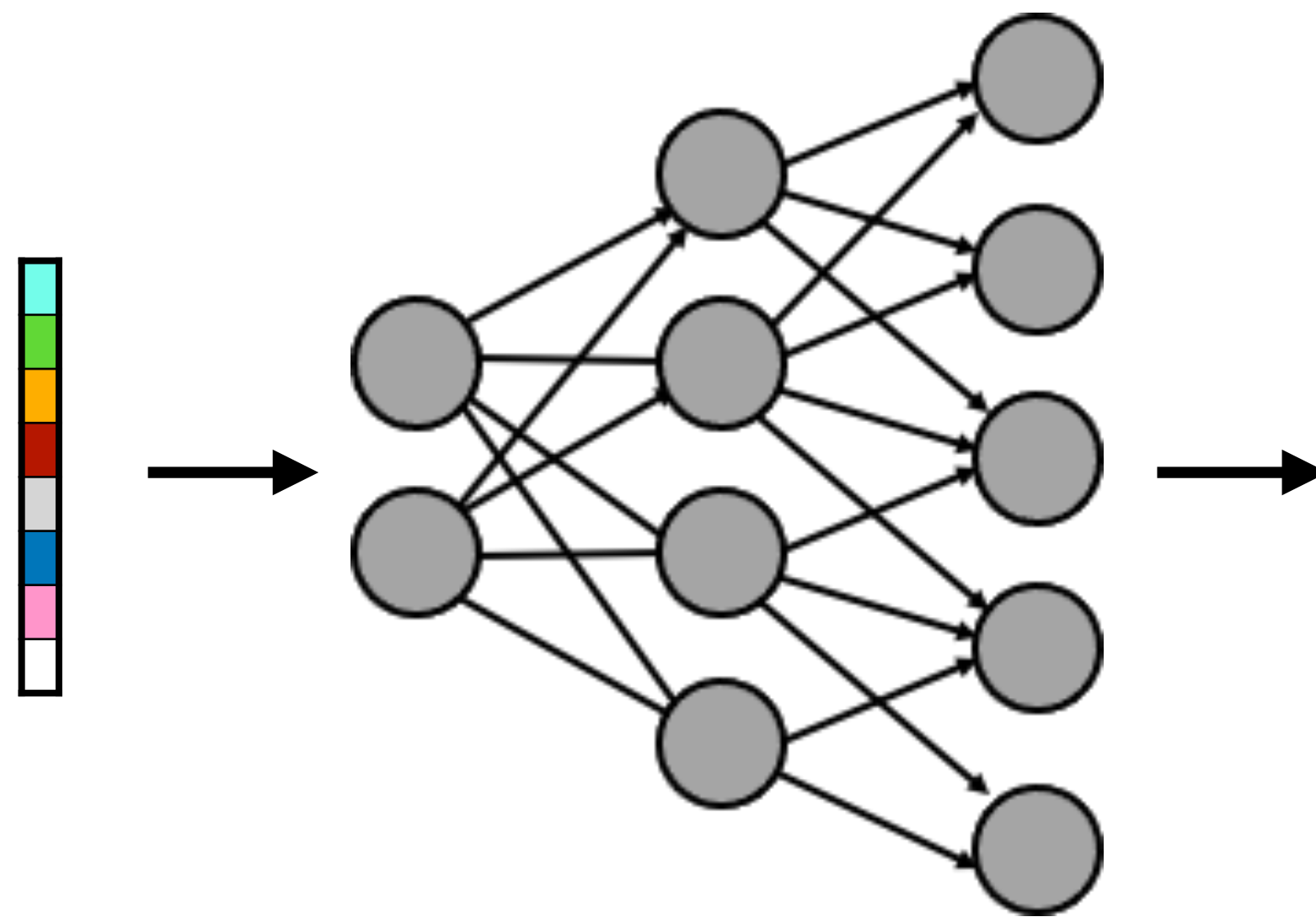
Adversarial Training



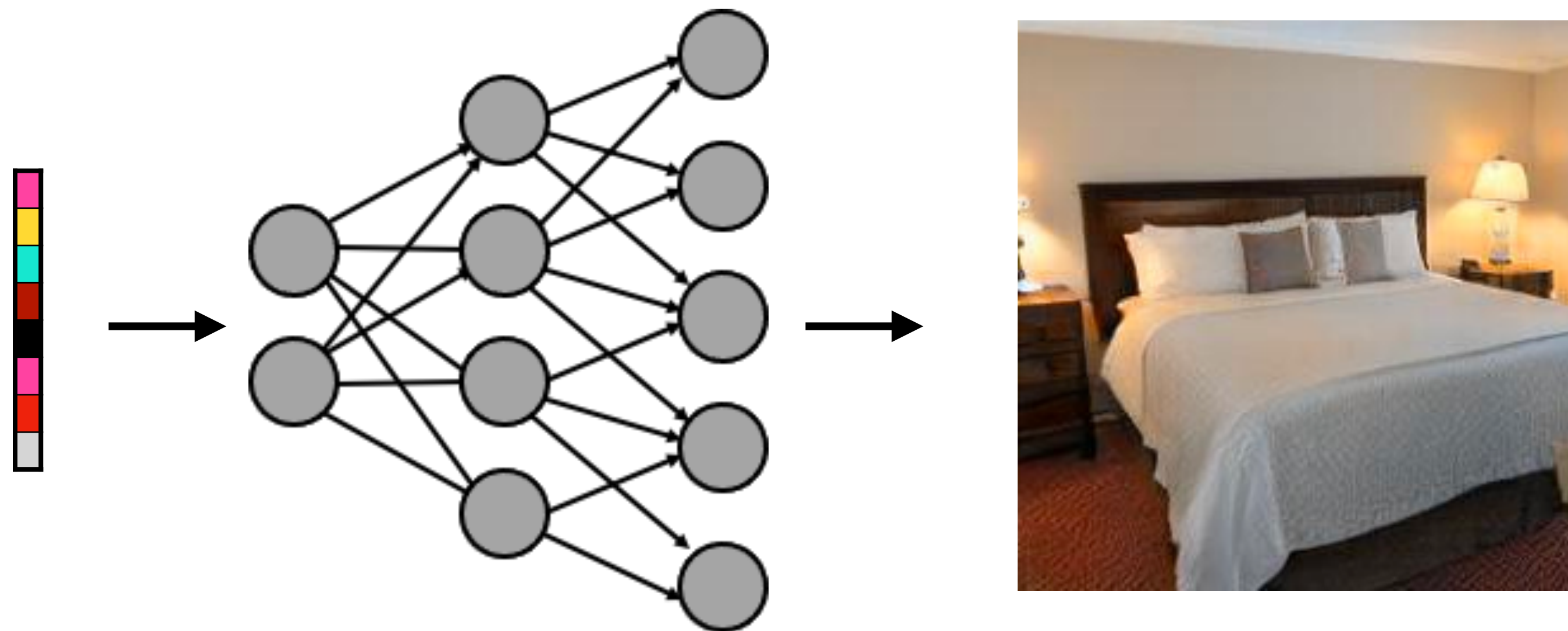
Neural Image Generation



Neural Image Generation

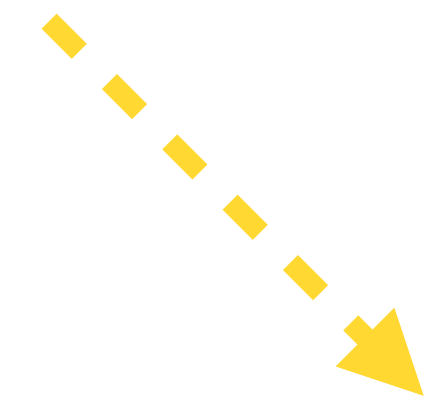
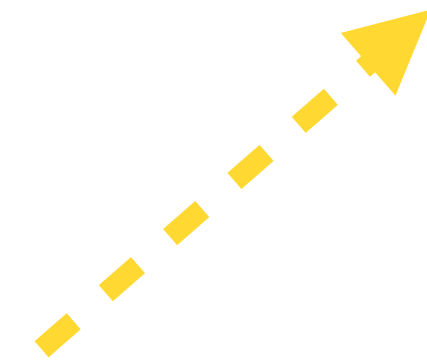
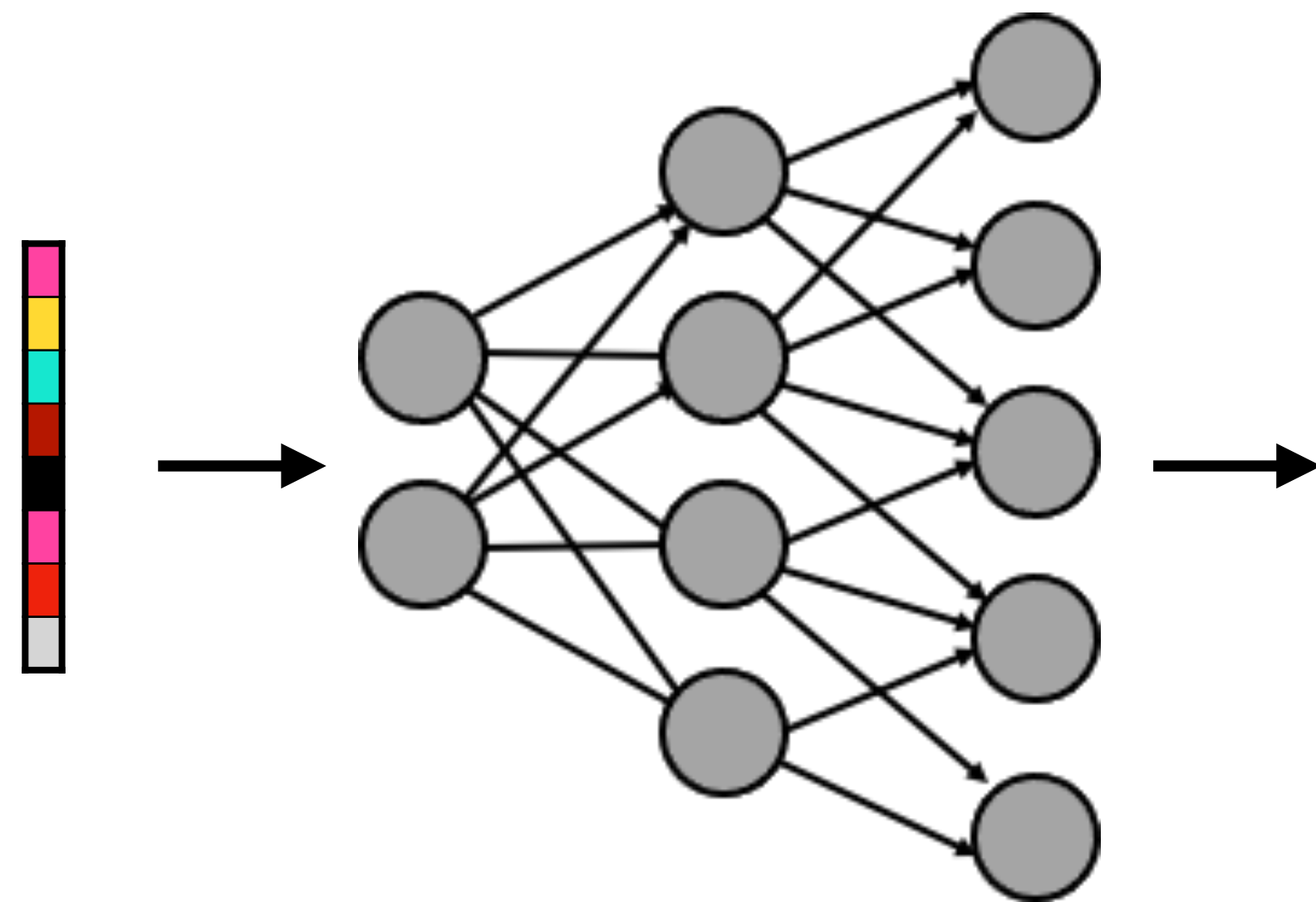


Neural Image Generation

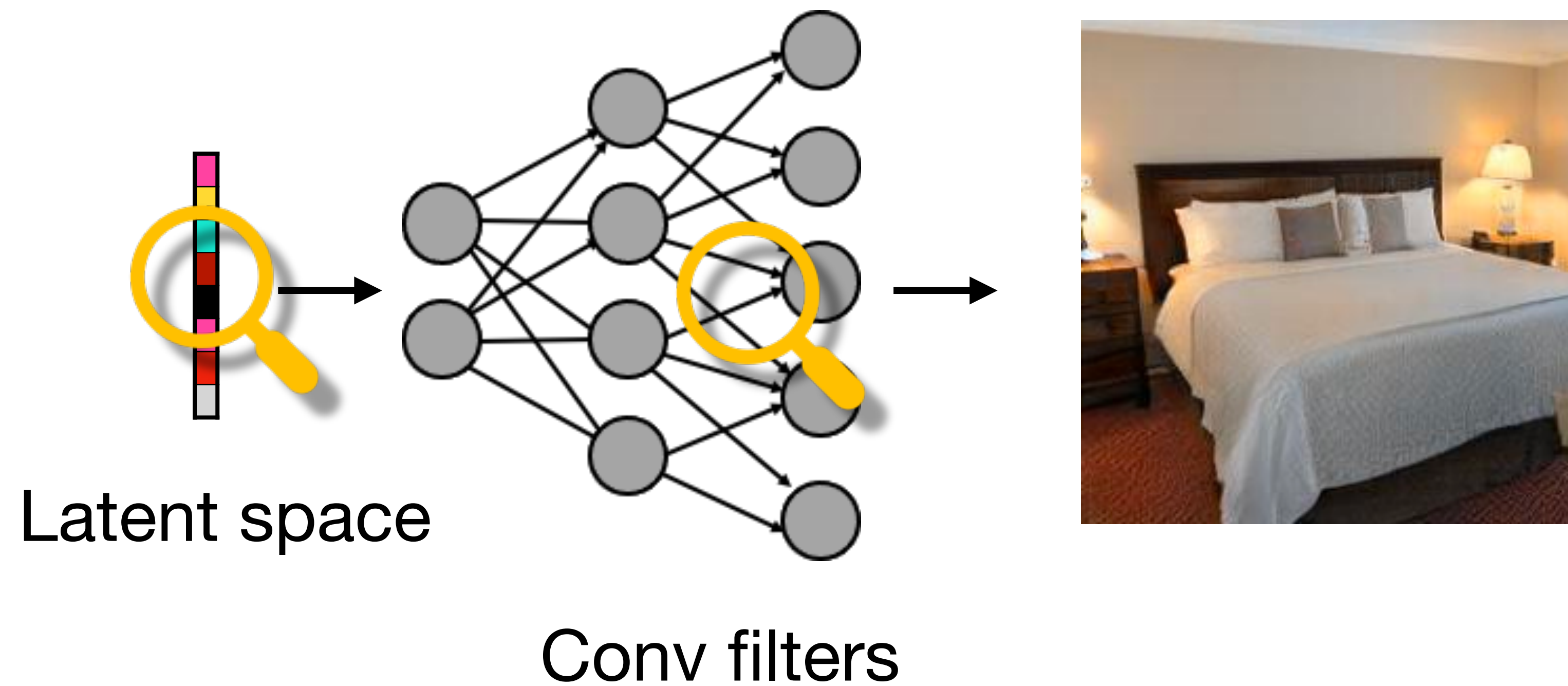


How to Steer Neural Image Generation?

- Interpret the generative representations with human understandable concepts
- Put human in the loop of AI content creation



Deep Generative Representations



Interpretation Approaches

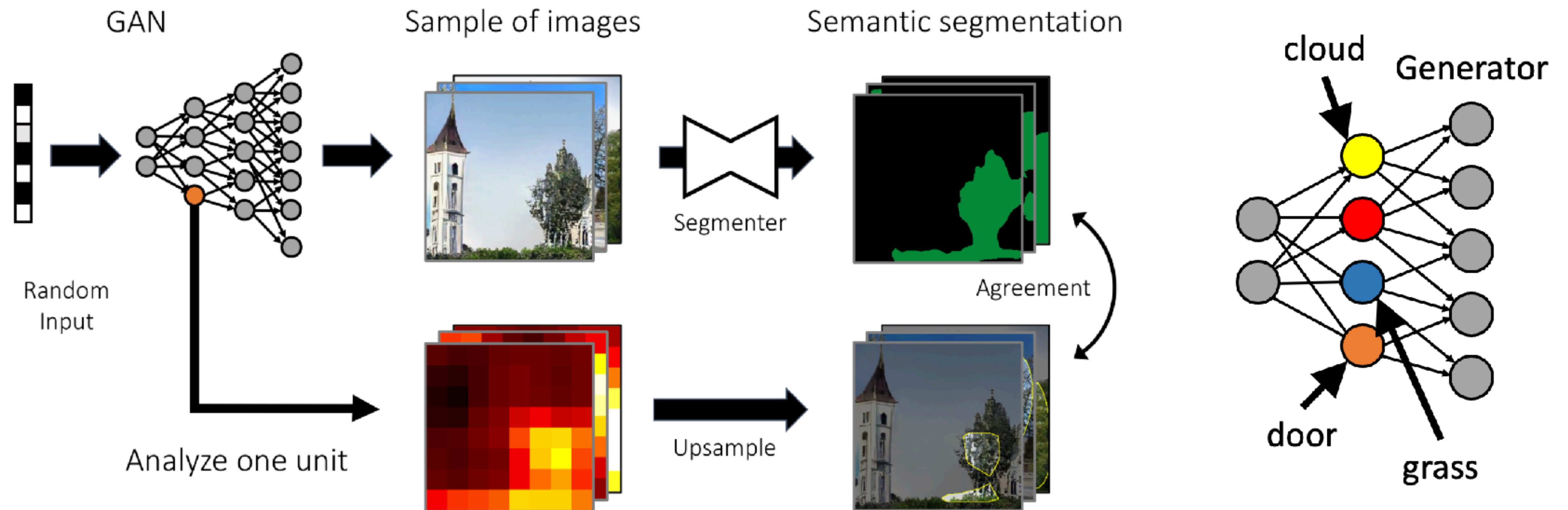
- **Supervised approach:** use labels or trained classifiers to probe the representation of the generator
- **Unsupervised approach:** identify the controllable dimensions of generator without labels/classifiers
- **Zero-shot approach:** align language embedding with generative representations

Interpretation Approaches

- **Supervised approach:** use labels or trained classifiers to probe the representation of the generator
- **Unsupervised approach:** identify the controllable dimensions of generator without labels/classifiers
- **Zero-shot approach:** align language embedding with generative representations

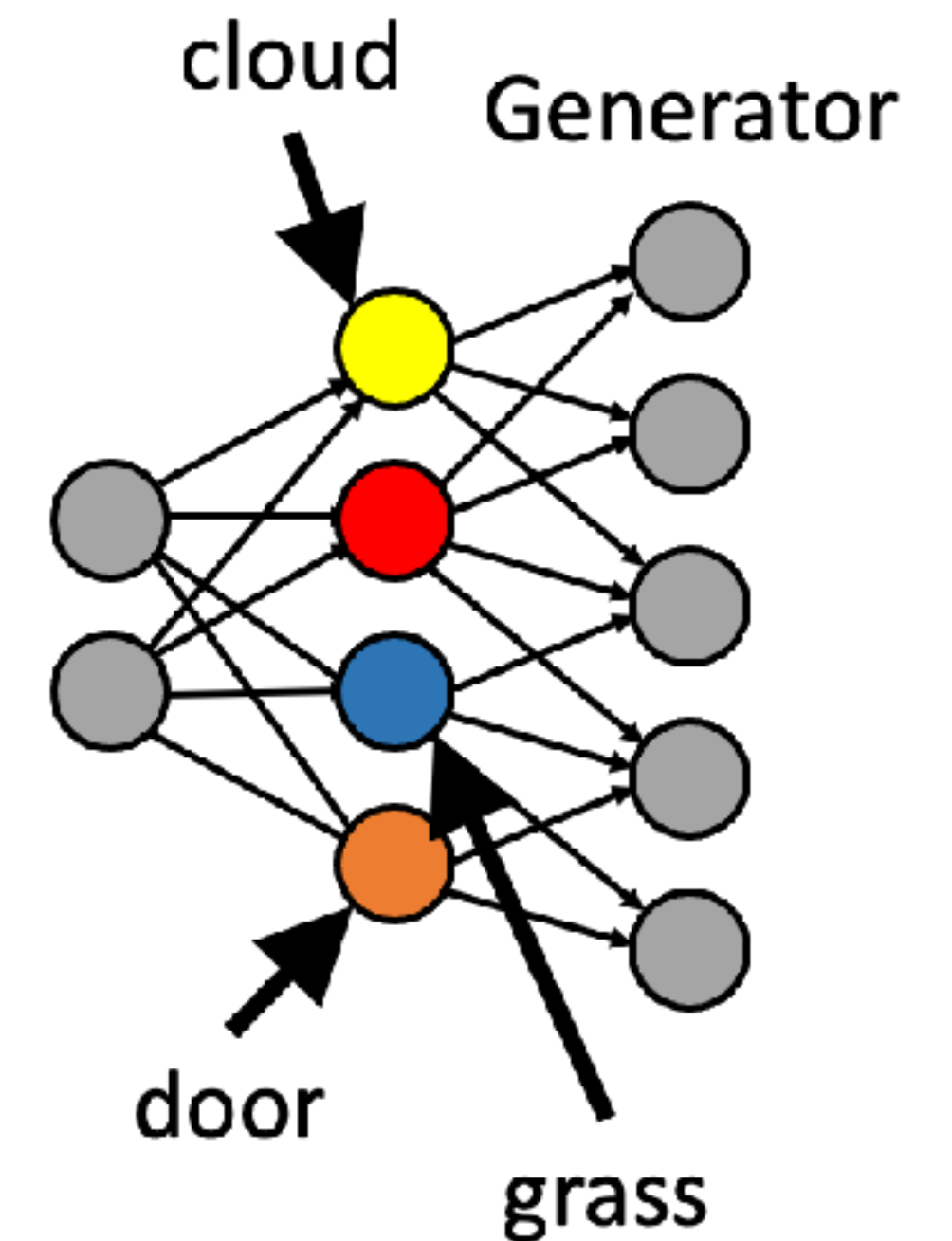
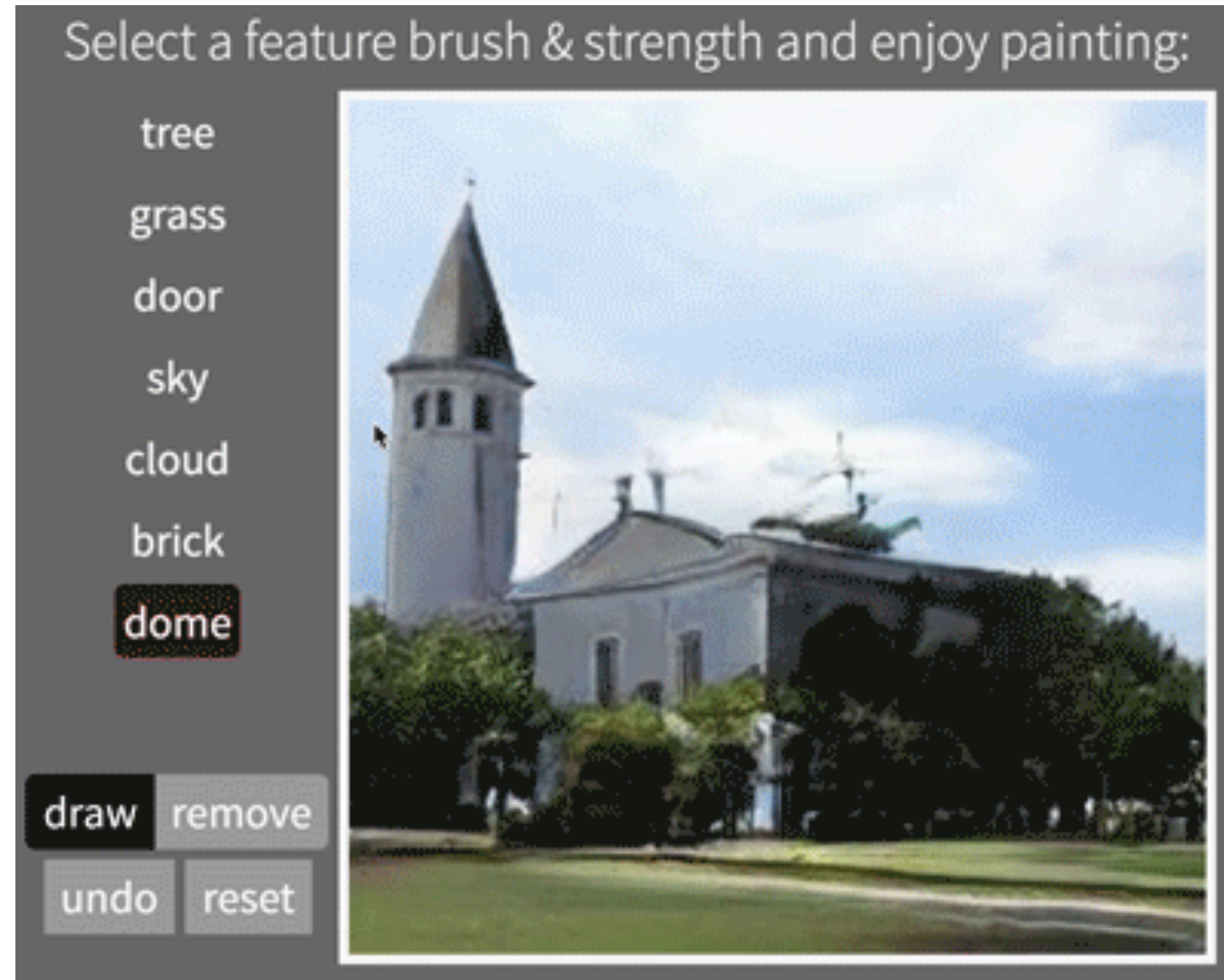
Supervised Approach

GAN Dissection: Aligning semantic segmentation with GAN feature map

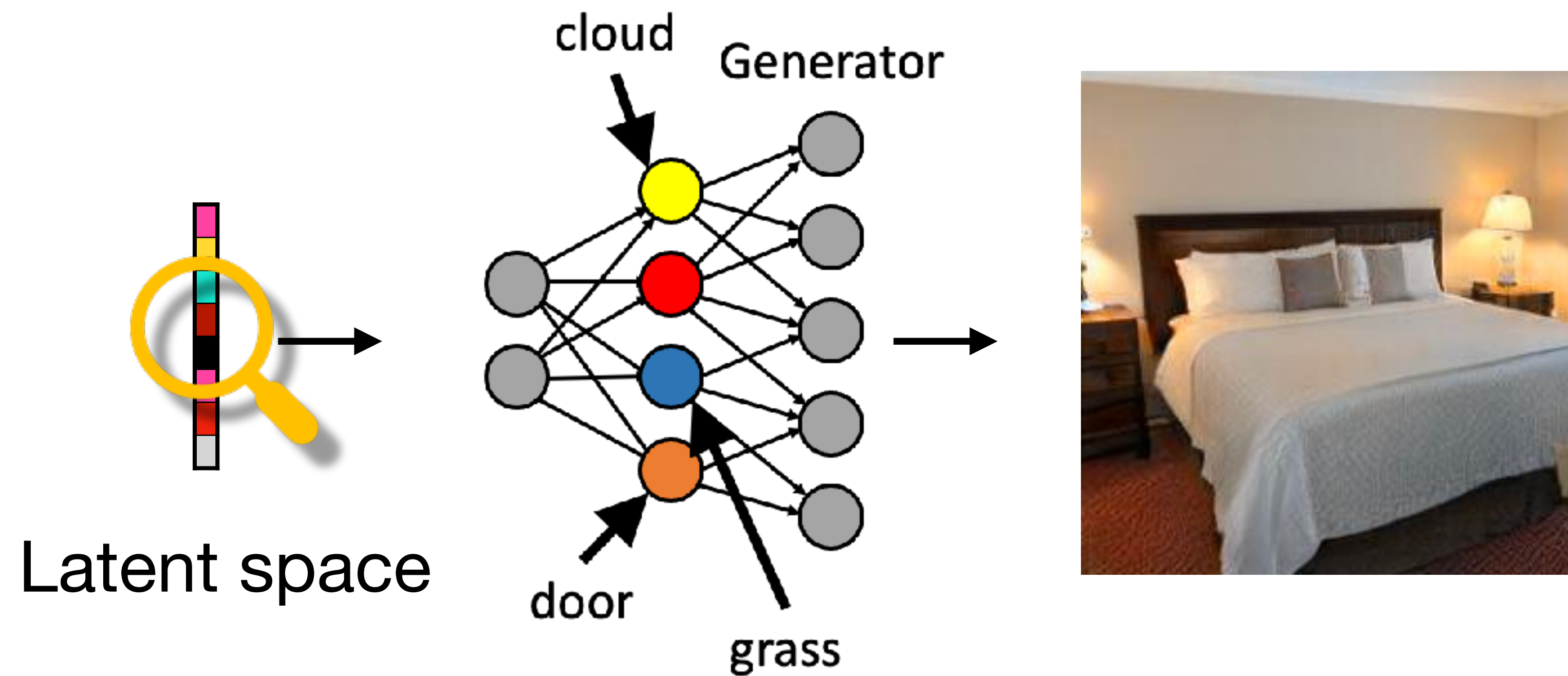


Supervised Approach

GAN Dissection: Aligning semantic segmentation with GAN feature map

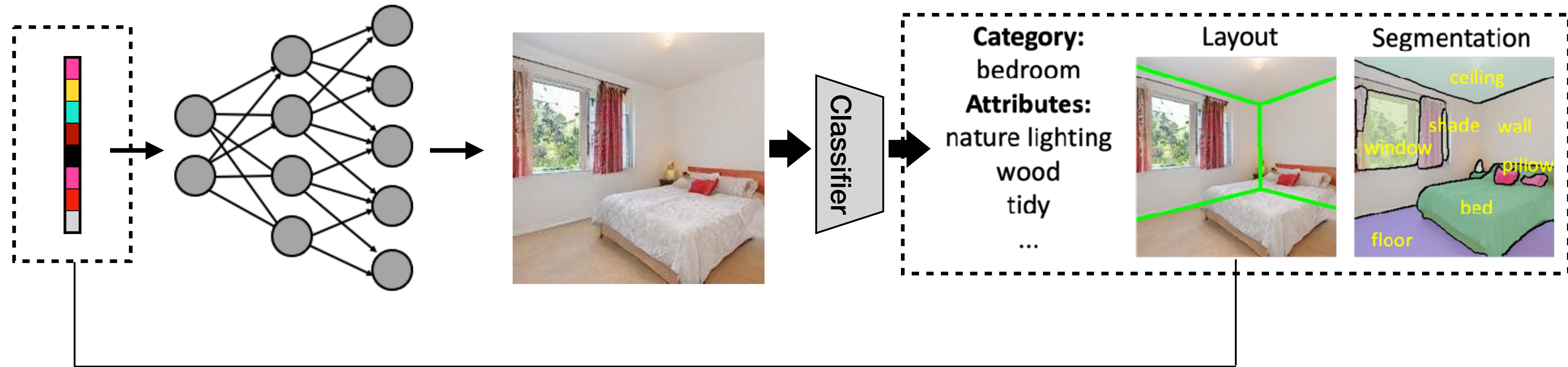


Representations of GAN's Generator



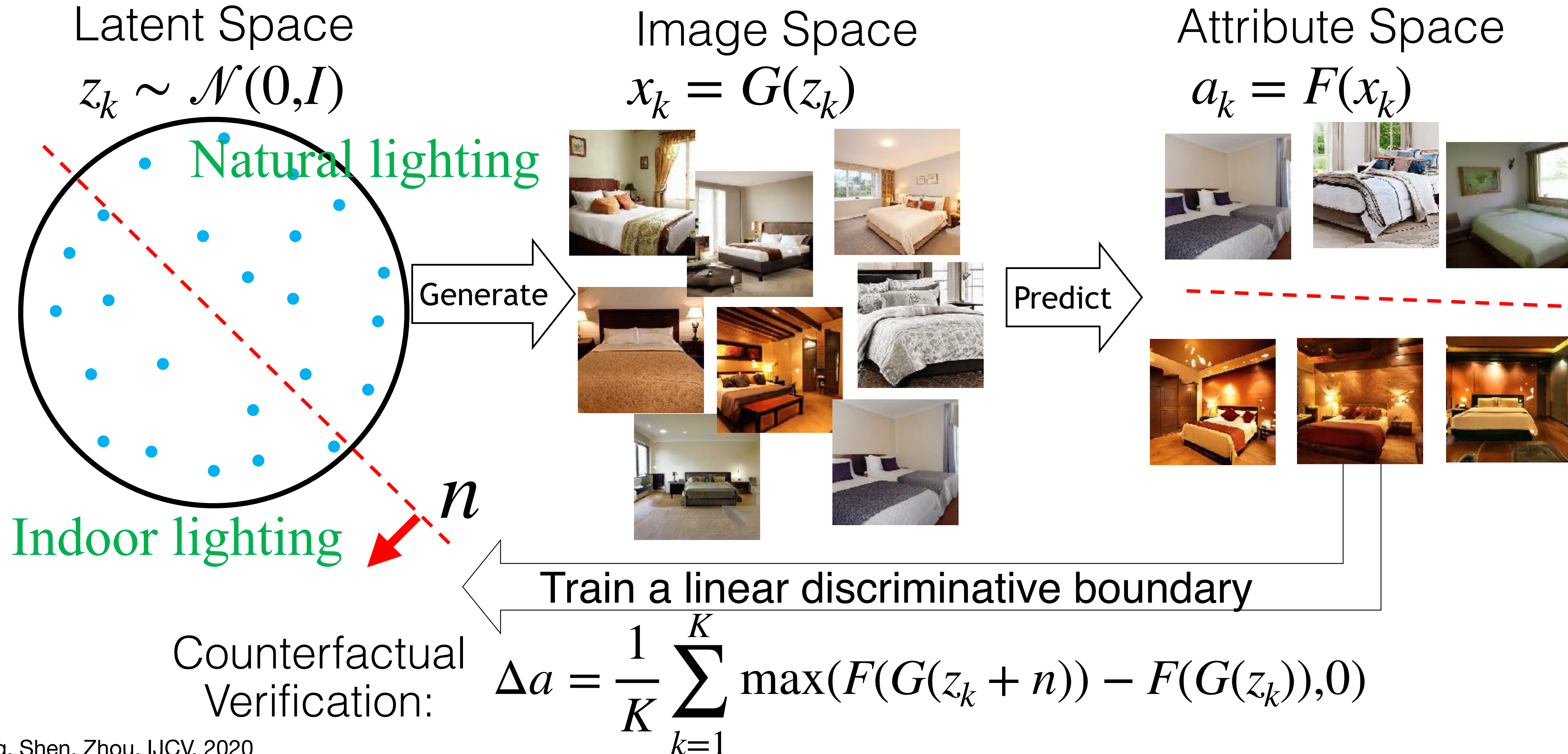
Supervised Approach

Probing latent space with linear classifier

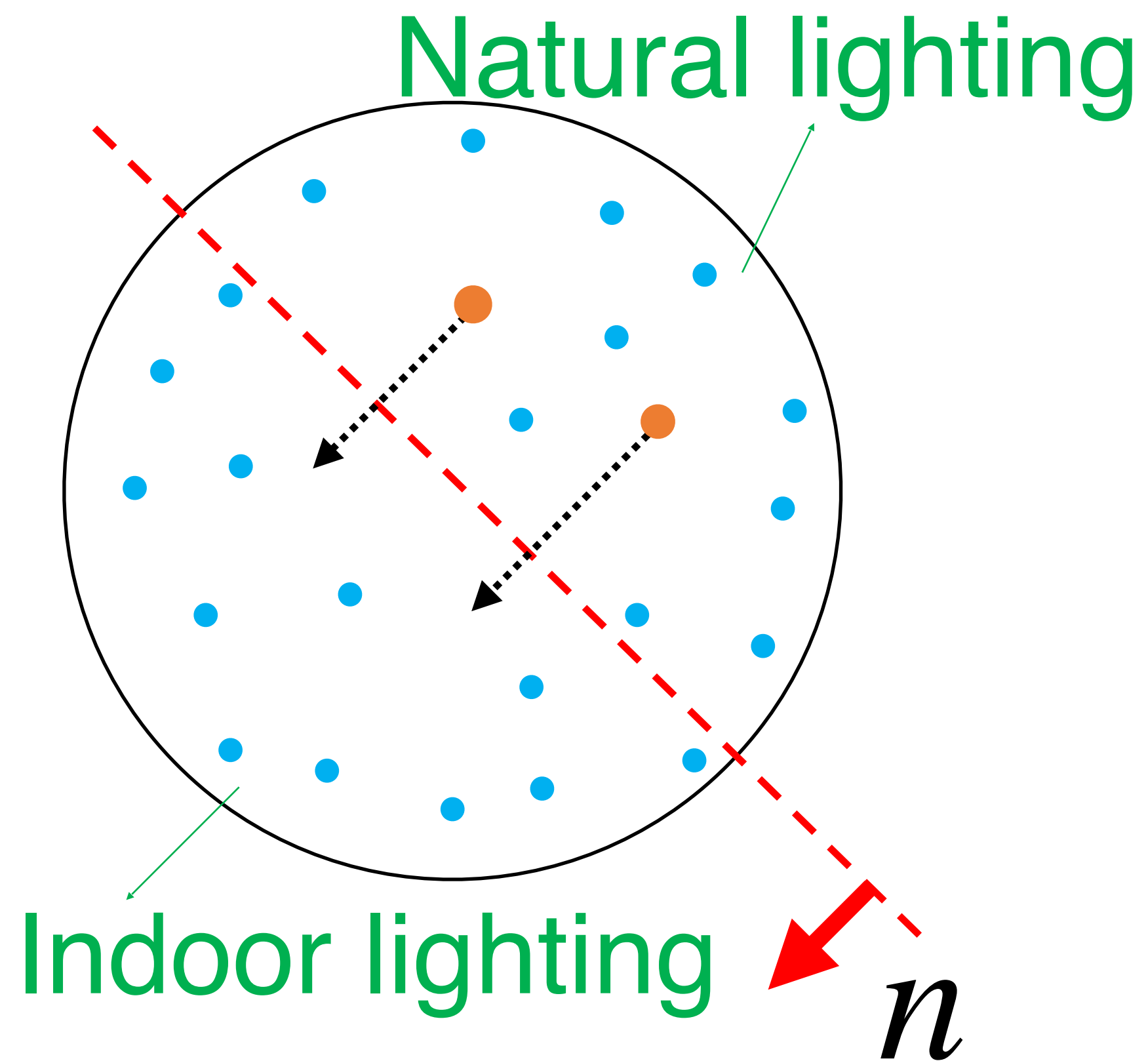


Identifying Causality Relations

Identifying Causality Relations

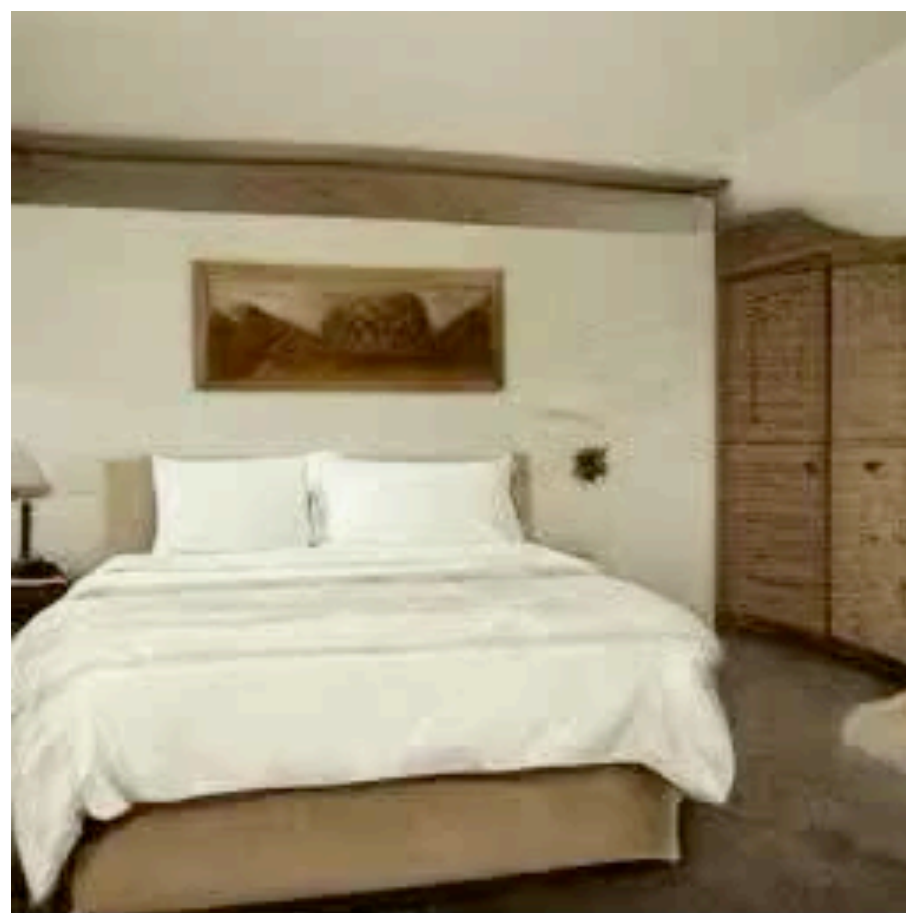
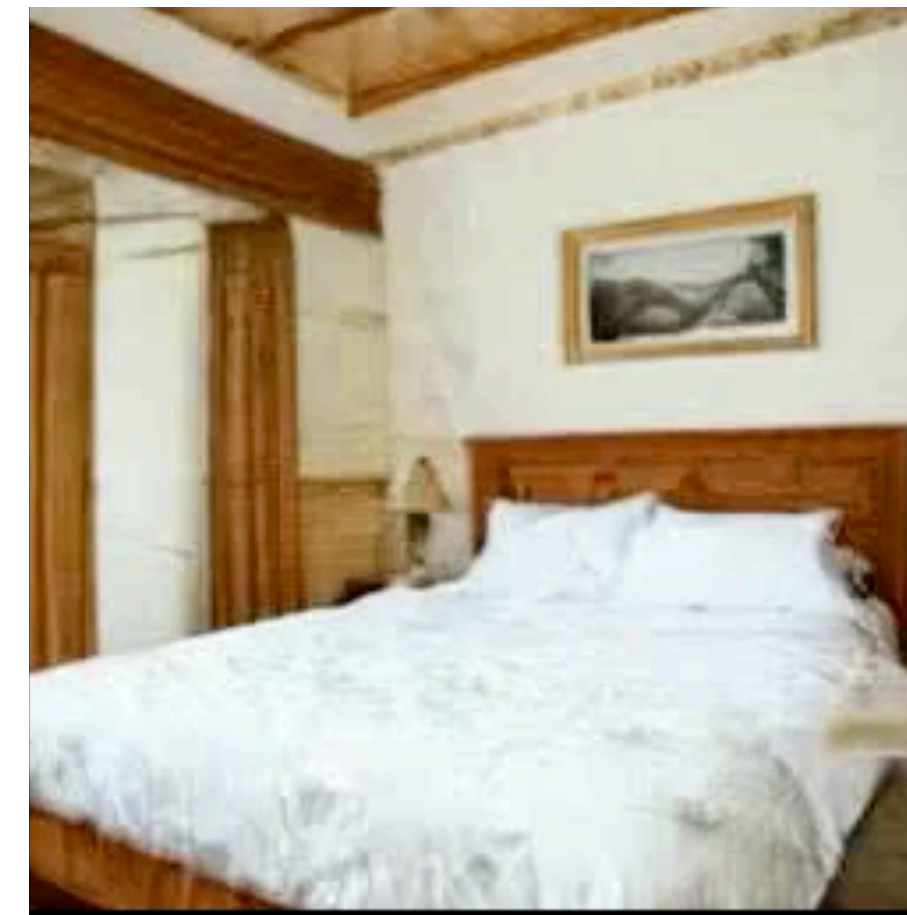


Linear Manipulation on Latent Code



Steering Generative Model

Changing Indoor lighting



Steering Generative Model

Adding clouds



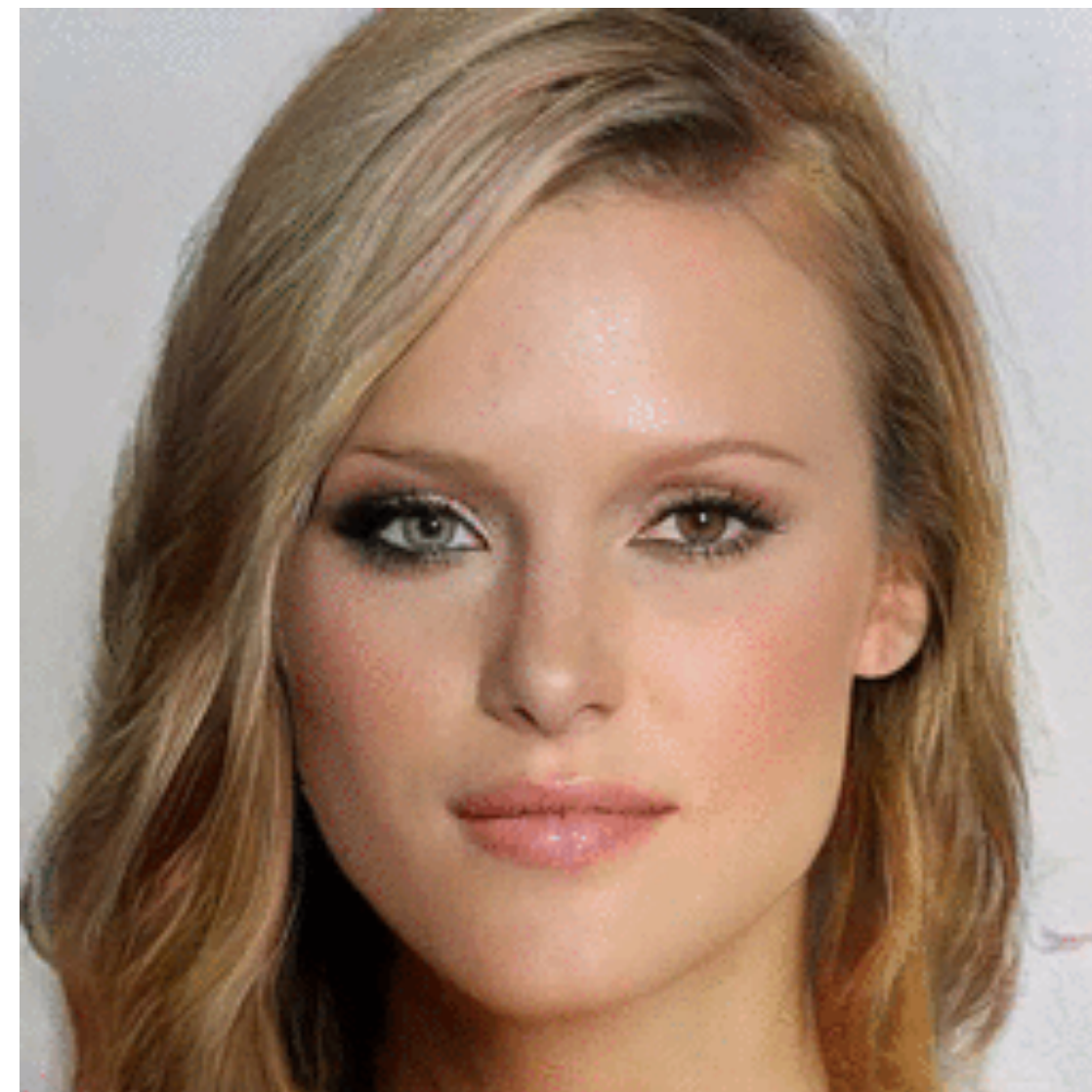
Supervised Approach

InterFaceGAN: Probing latent space of face GAN with linear classifier

Age



Gender



Pose



Artifact



Supervised Approach

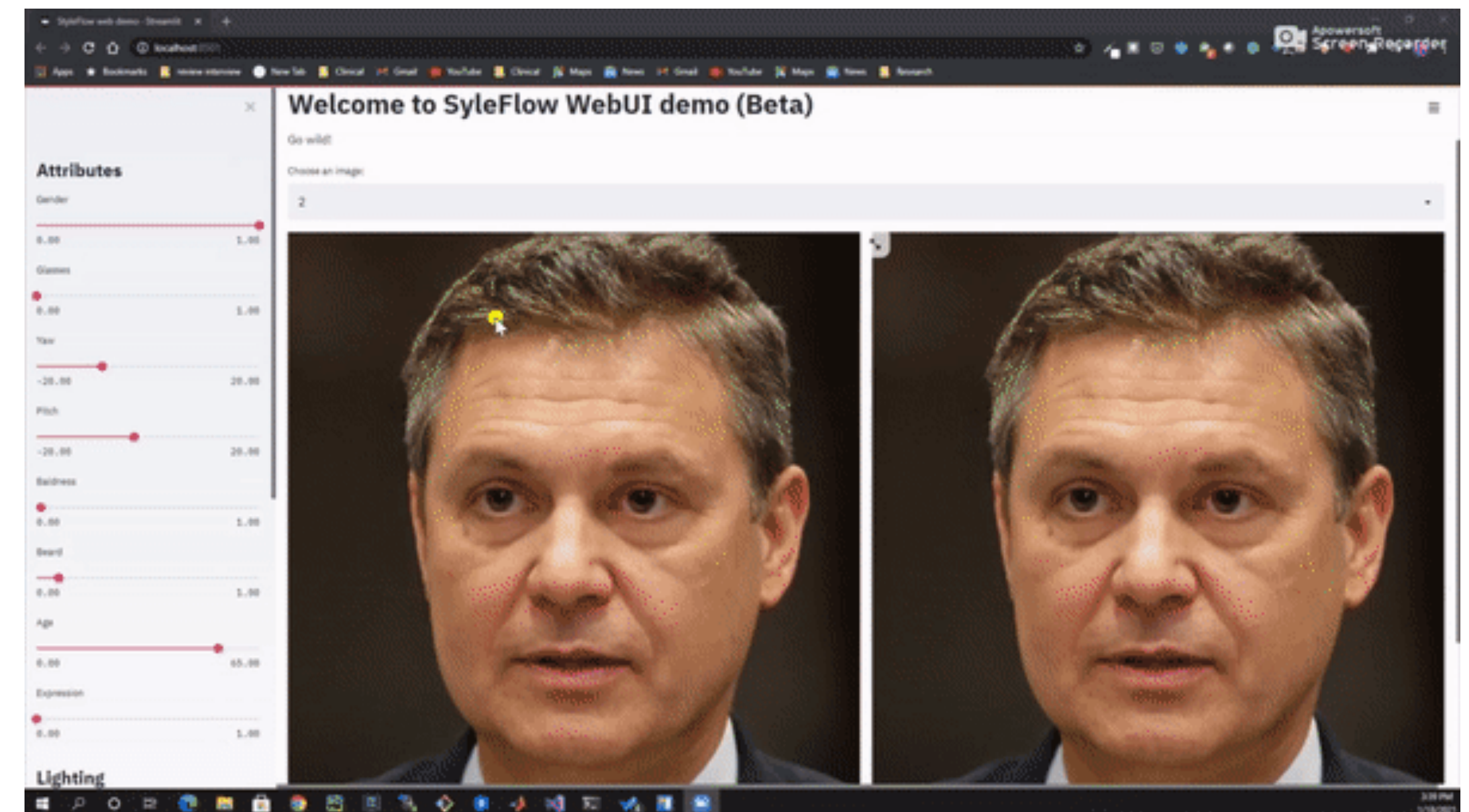
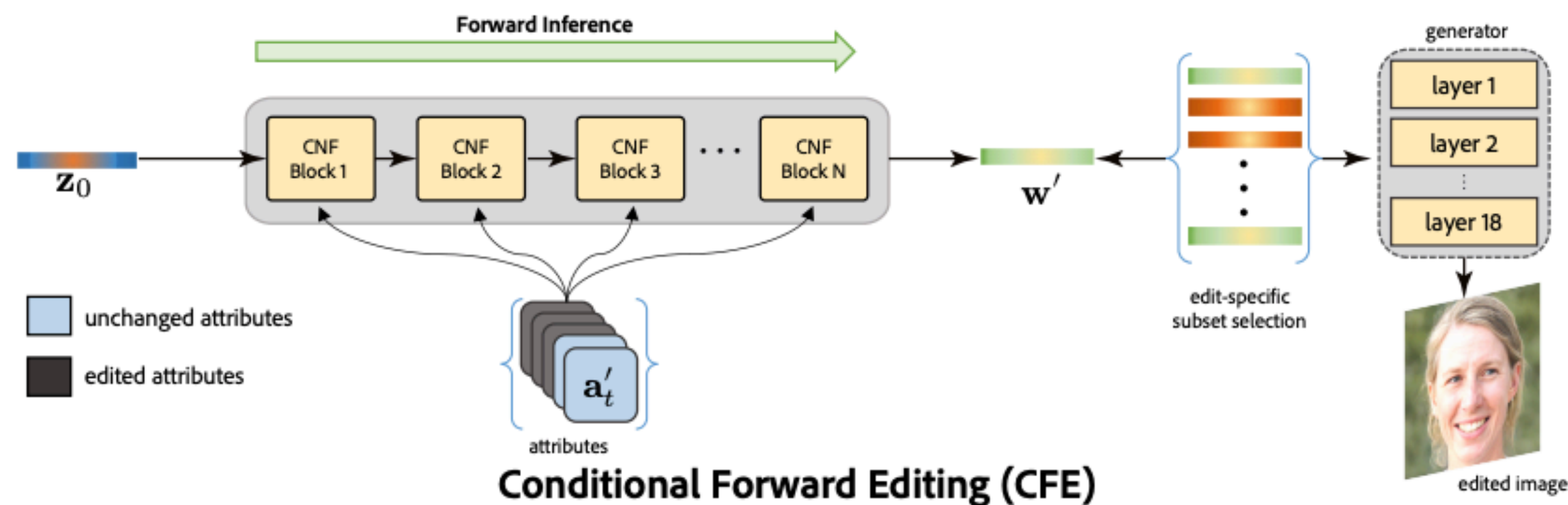
StyleFlow: StyleGAN + flow-based conditional model

- Previous work assumes the linear manipulation model:

$$I' = G(w + \lambda n_a), w = F(z), z \sim \mathcal{N}(0,1)$$

- StyleFlow: Replace the MLP with an invertible flow model conditioned on attributes

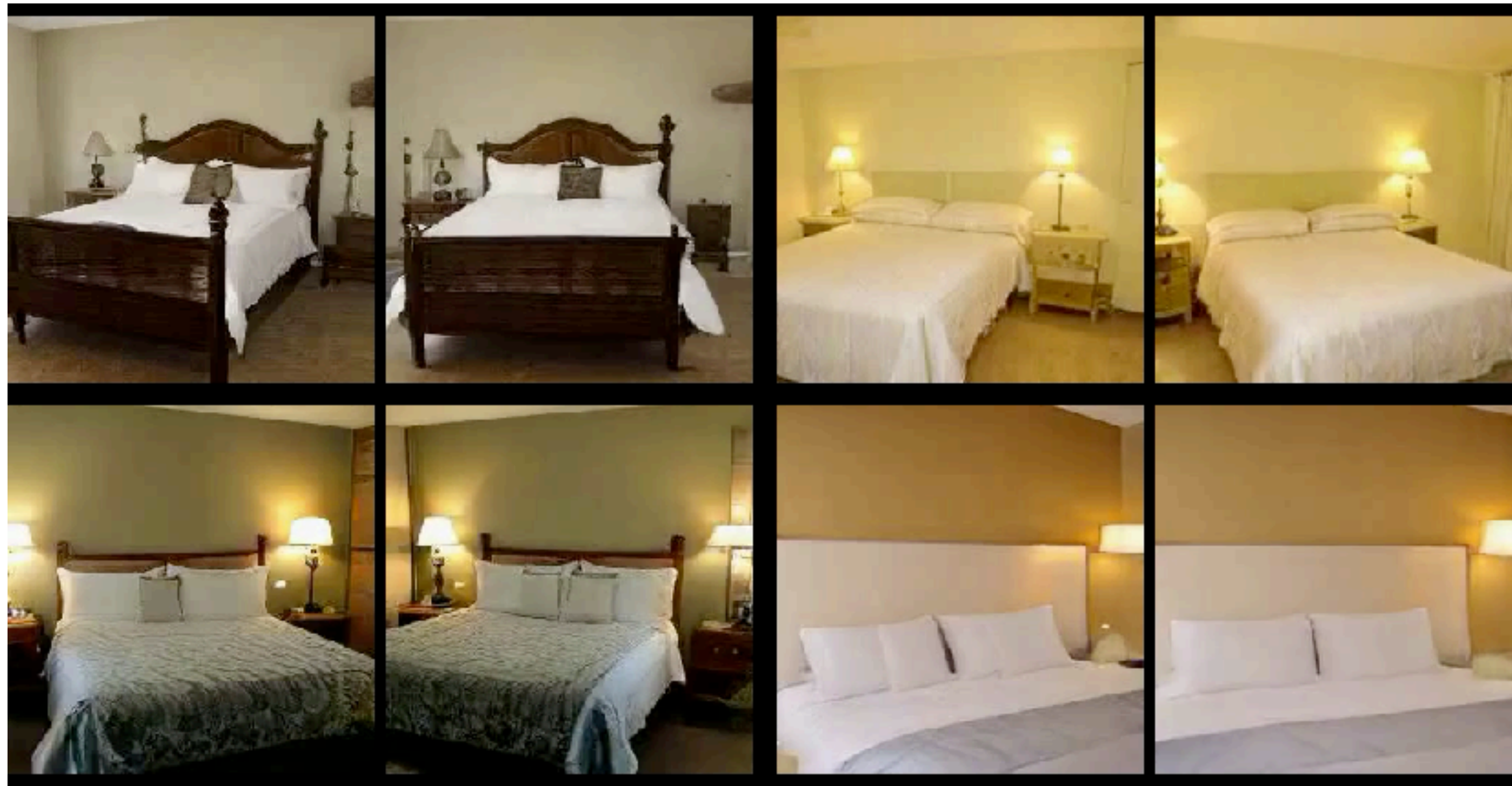
$$w = \Phi(z, a), z \sim \mathcal{N}(0,1)$$



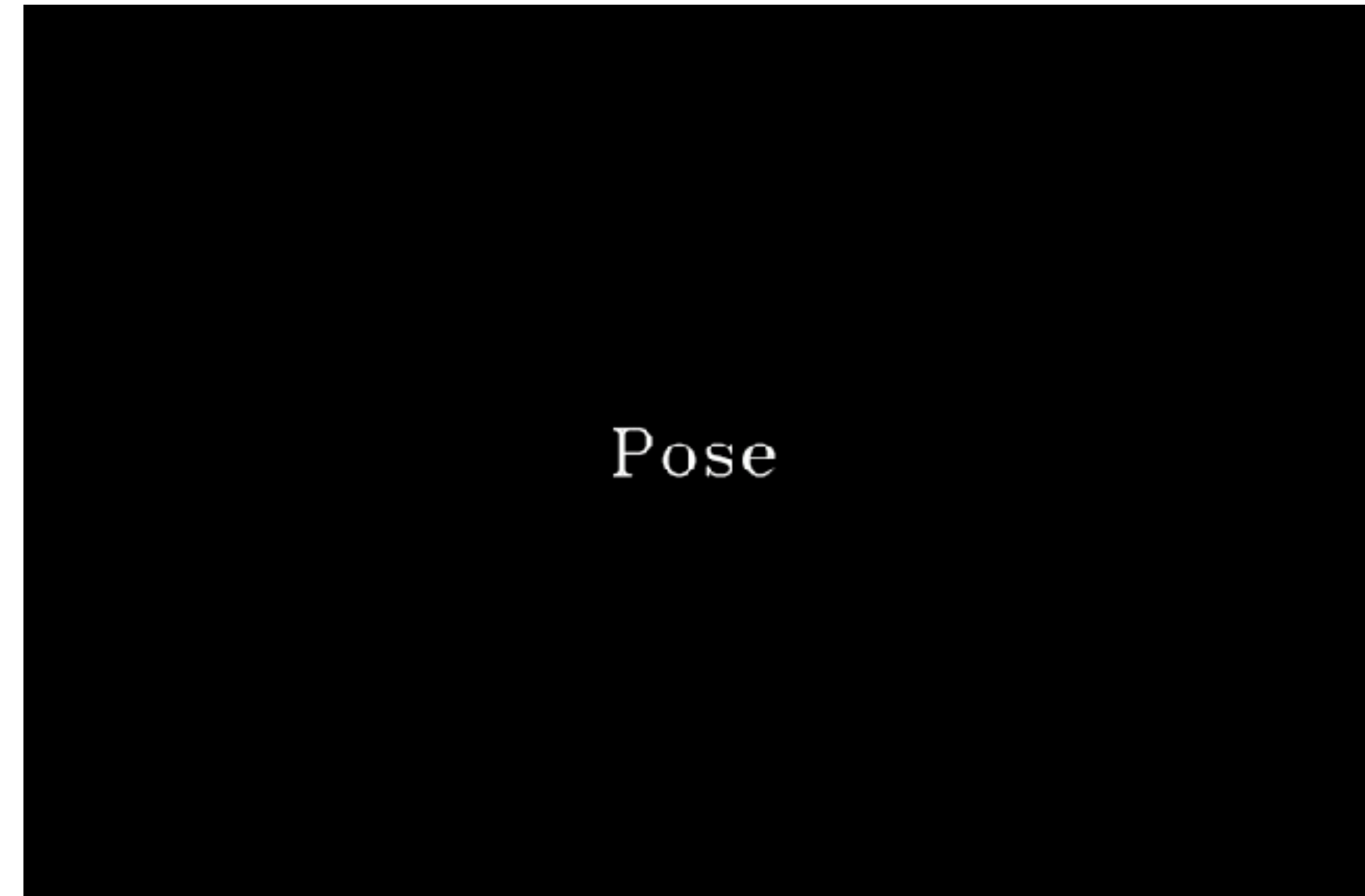
Supervised Approach

Does 3D structure emerge from 2D image generation?

Changing scene view (Yang et al, IJCV)



Changing face pose (Shen et al, CVPR'20)

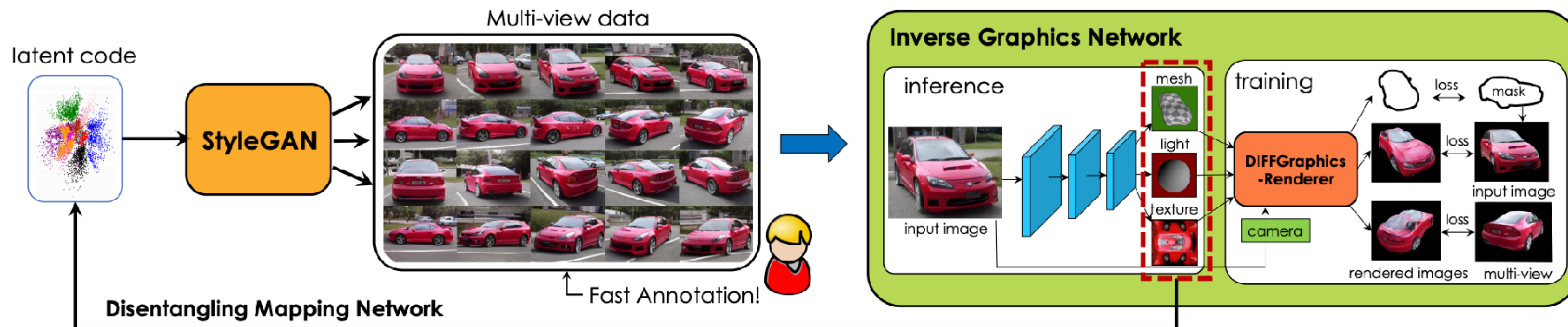


Supervised Approach

Parsing 3D Information from 2D Image Generator

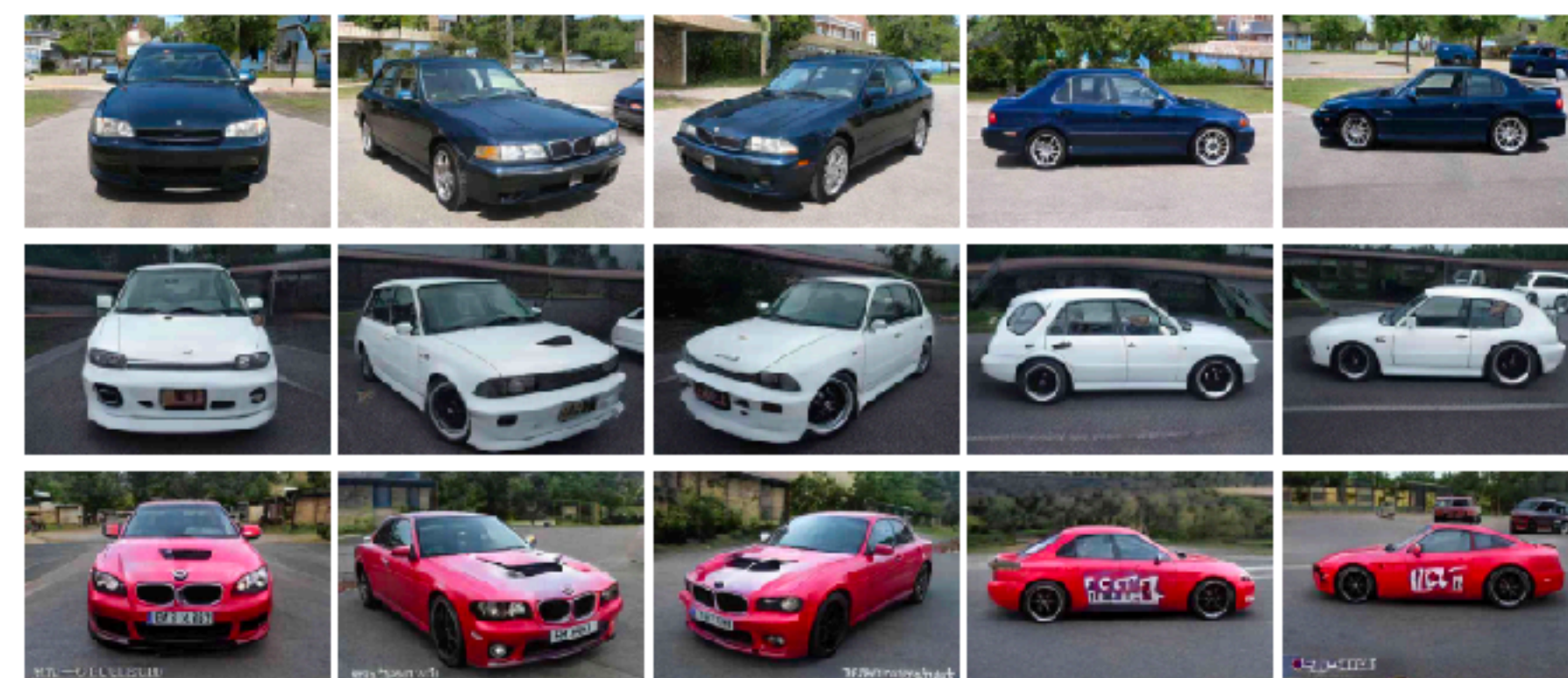
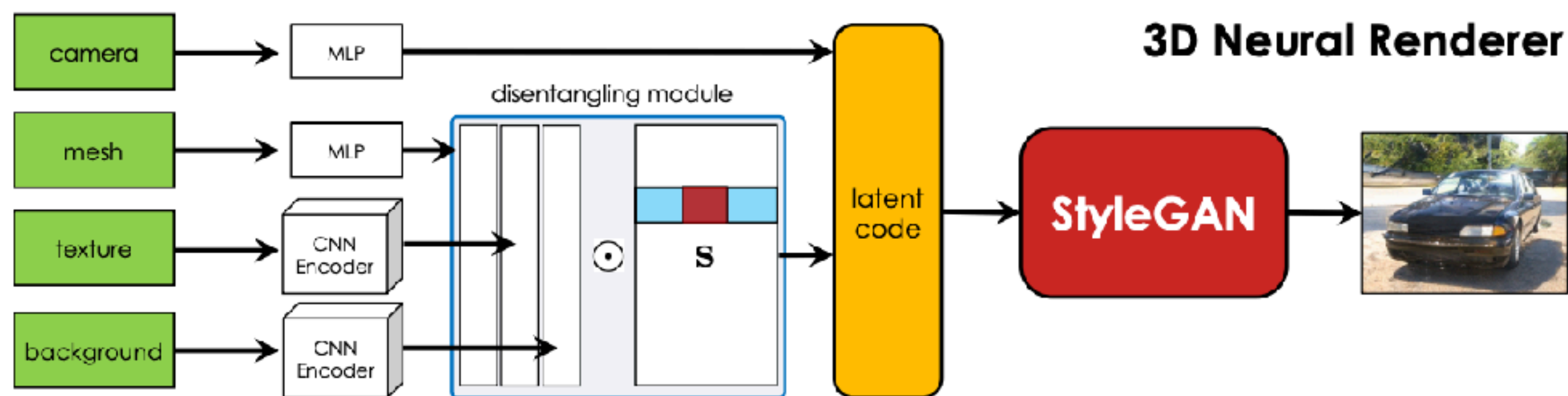
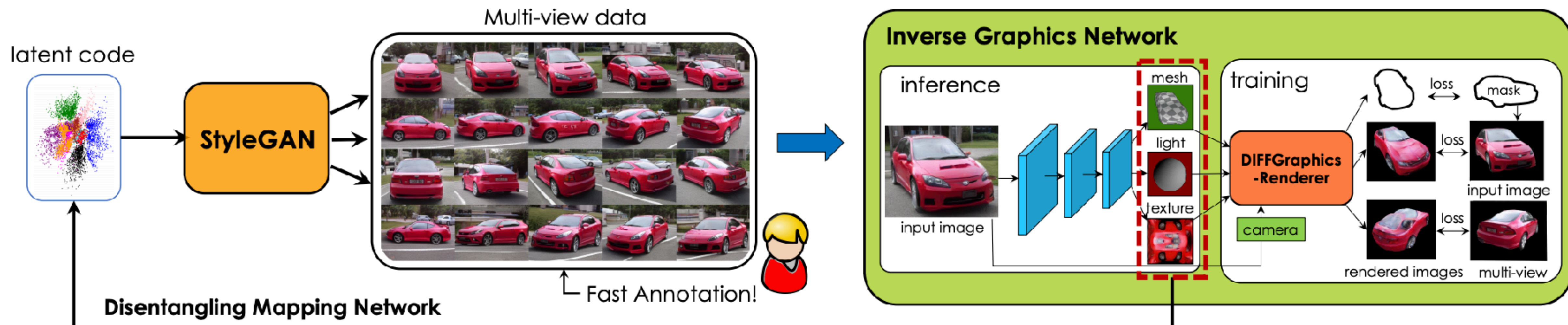
Differentiable rendering for inverse graphics and interpretable 3D rendering

1. Use multi-view synthetic data to train inverse graphics network
2. Use trained inverse graphics net to train mapping network



Supervised Approach

Parsing 3D Information from 2D Image Generator



Controllable output

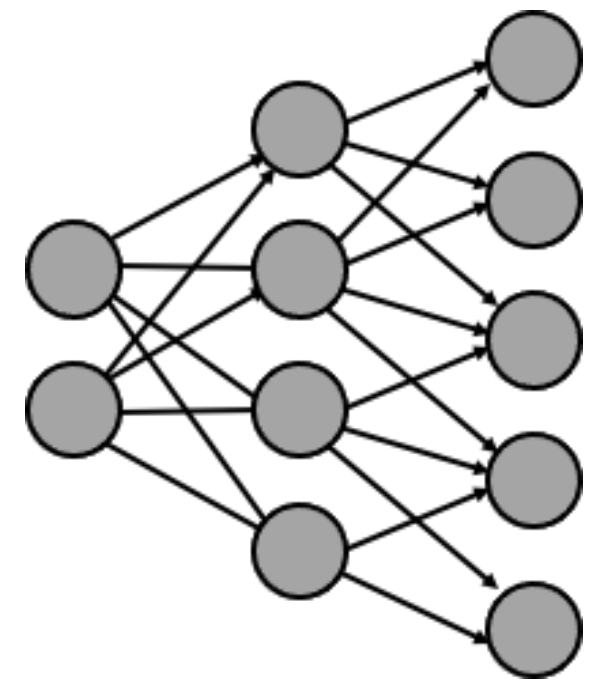
Challenges for Supervised Approach

- How to expand the annotated dictionary size?
- How to further disentangle the relevant attributes?
- How to align latent space with image region attributes?

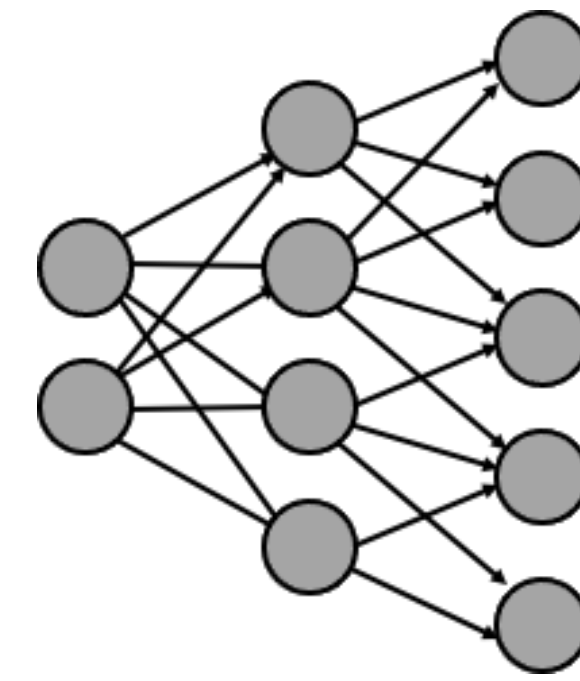
Interpretation Approaches

- **Supervised approach:** use labels or trained classifiers to probe the representation of the generator
- **Unsupervised approach:** identify the controllable dimensions of generator without labels/classifiers
- **Zero-shot approach:** align language embedding with generative representations

Unsupervised Approach



Generative model
for cats

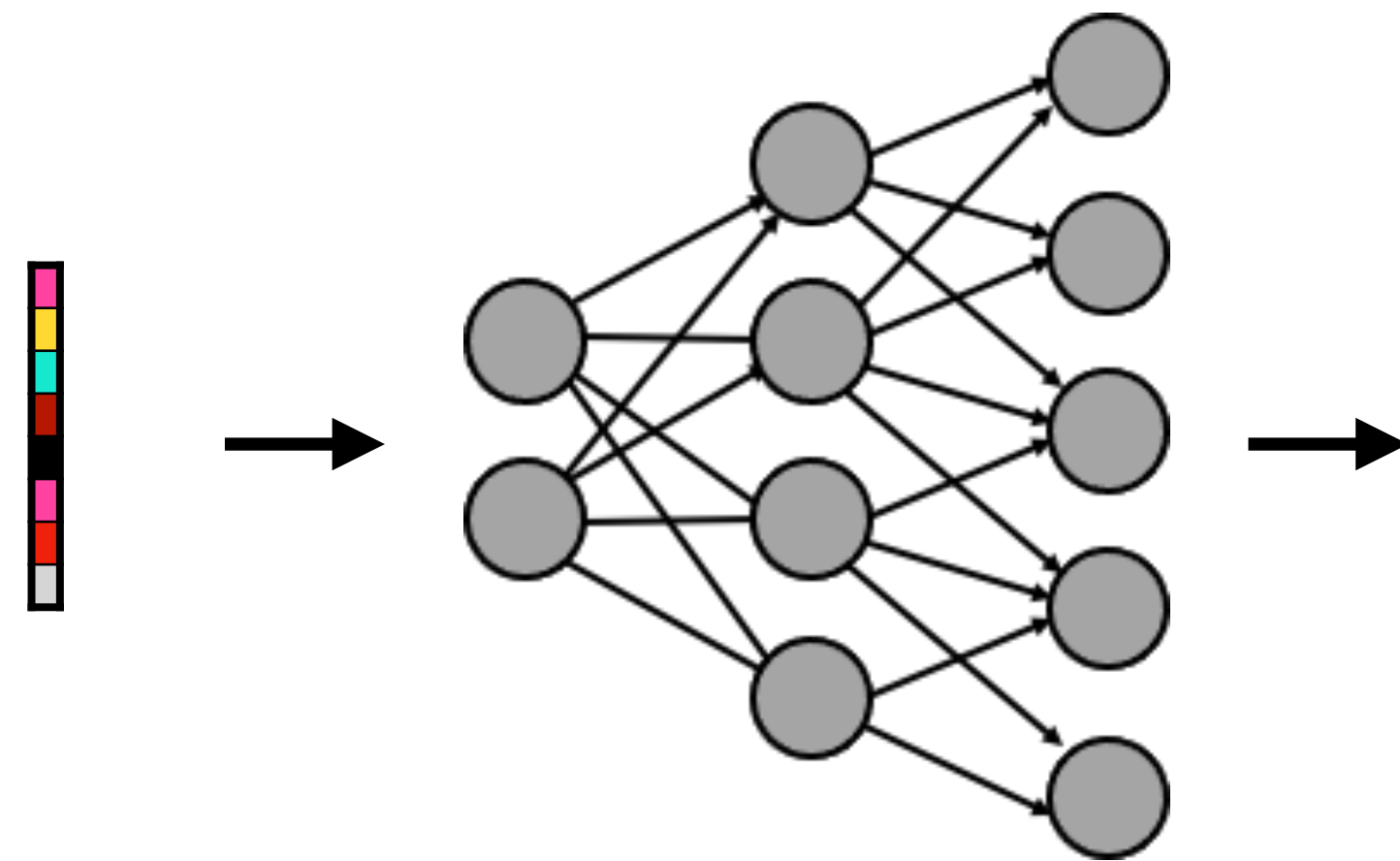


Generative model
for cartoons



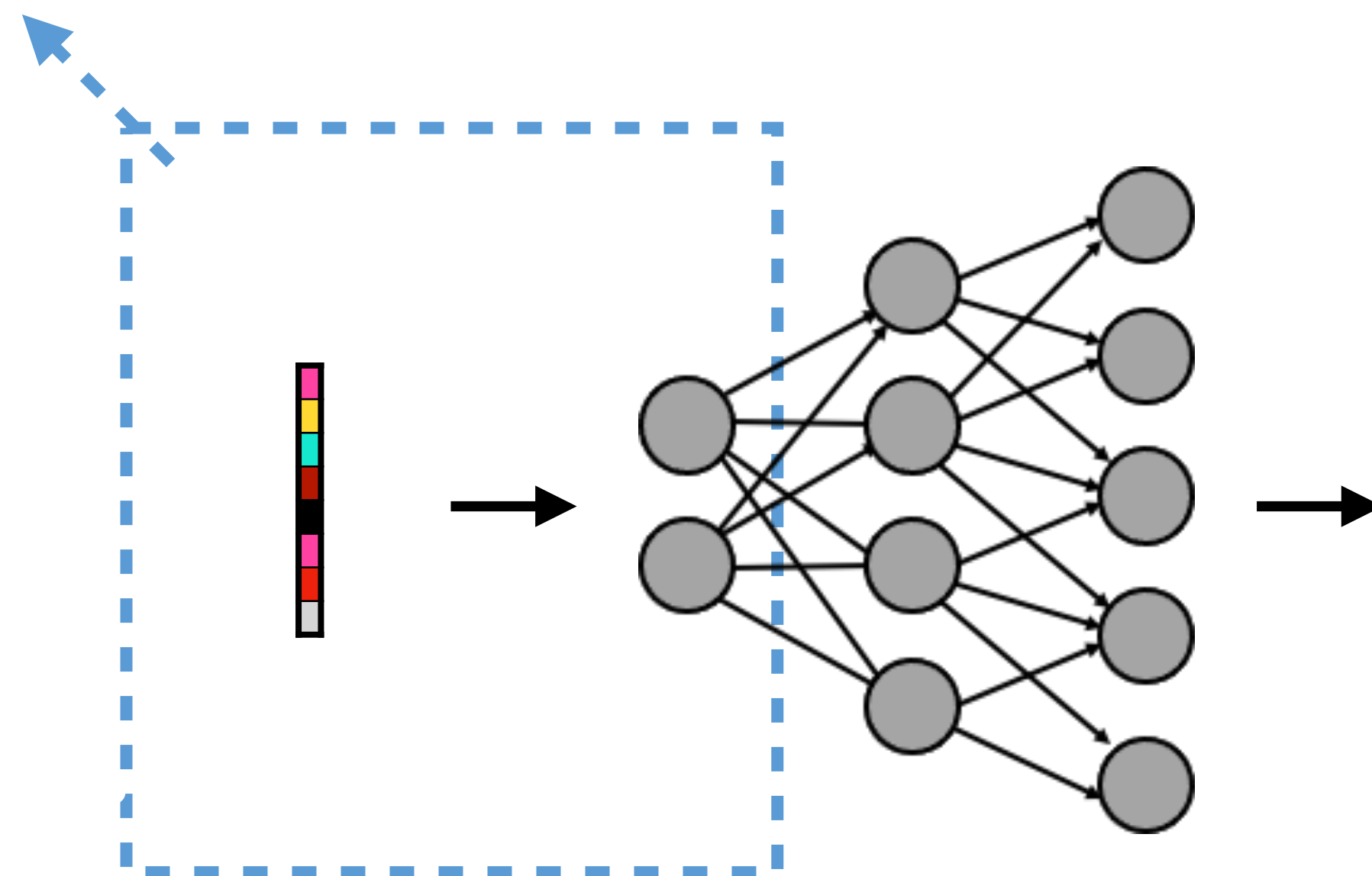
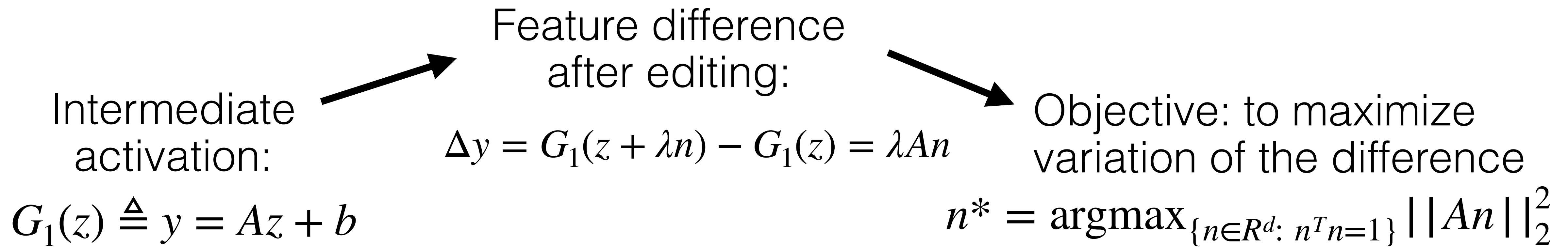
Unsupervised Approach

SeFa: Closed-form factorization of latent space in GANs



Unsupervised Approach

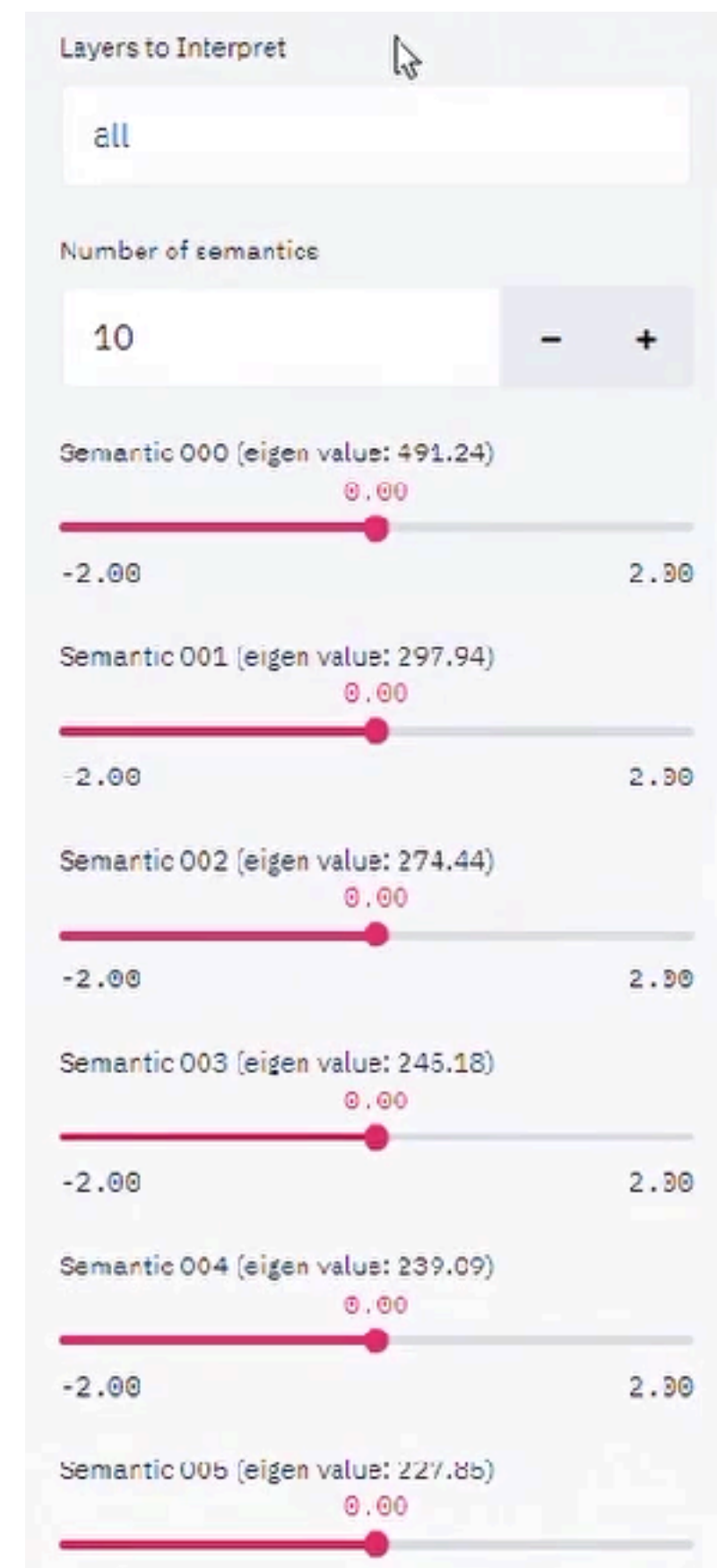
SeFa: Closed-form factorization of latent space in GANs



Unsupervised Approach

SeFa: Closed-form factorization of latent space in GANs

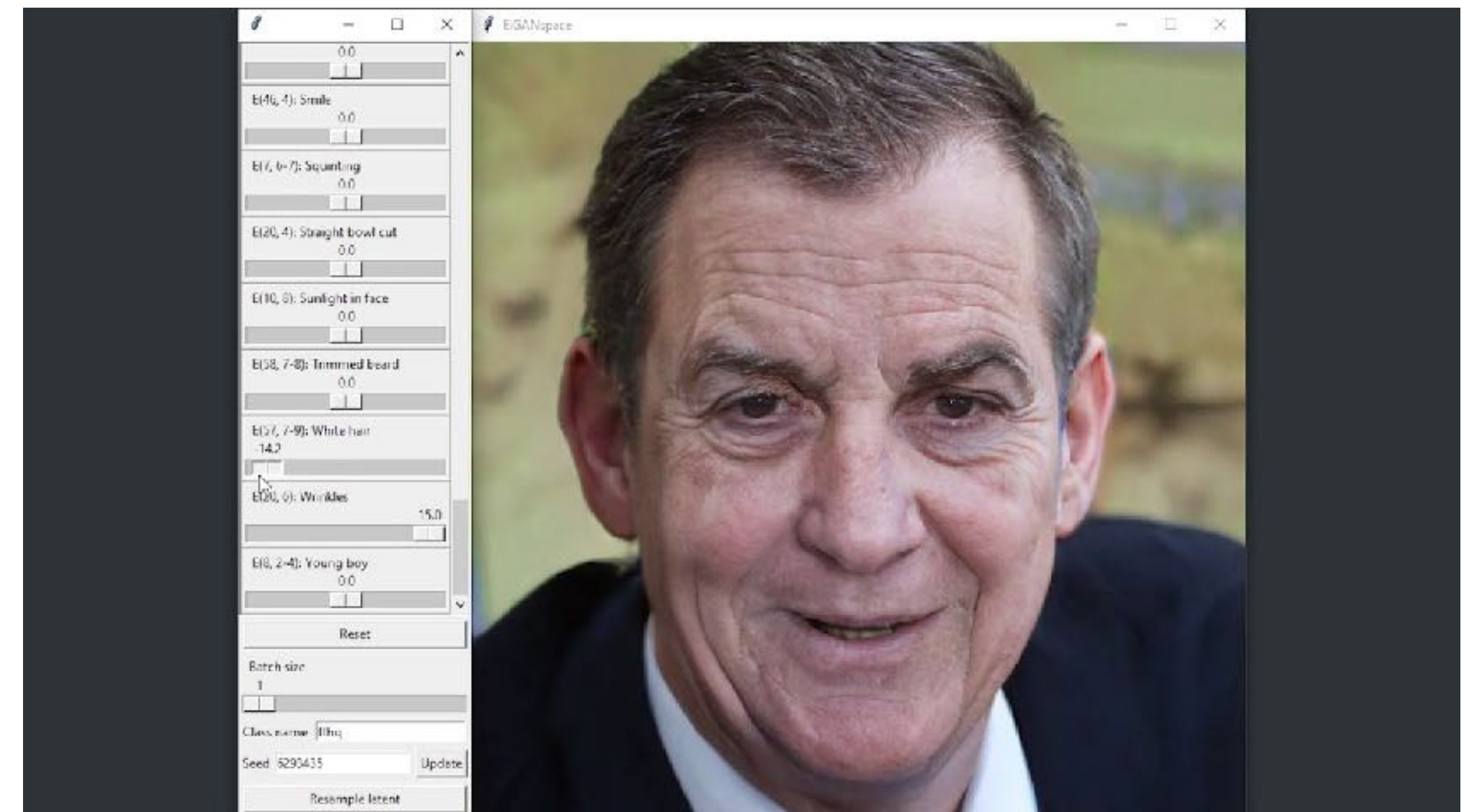
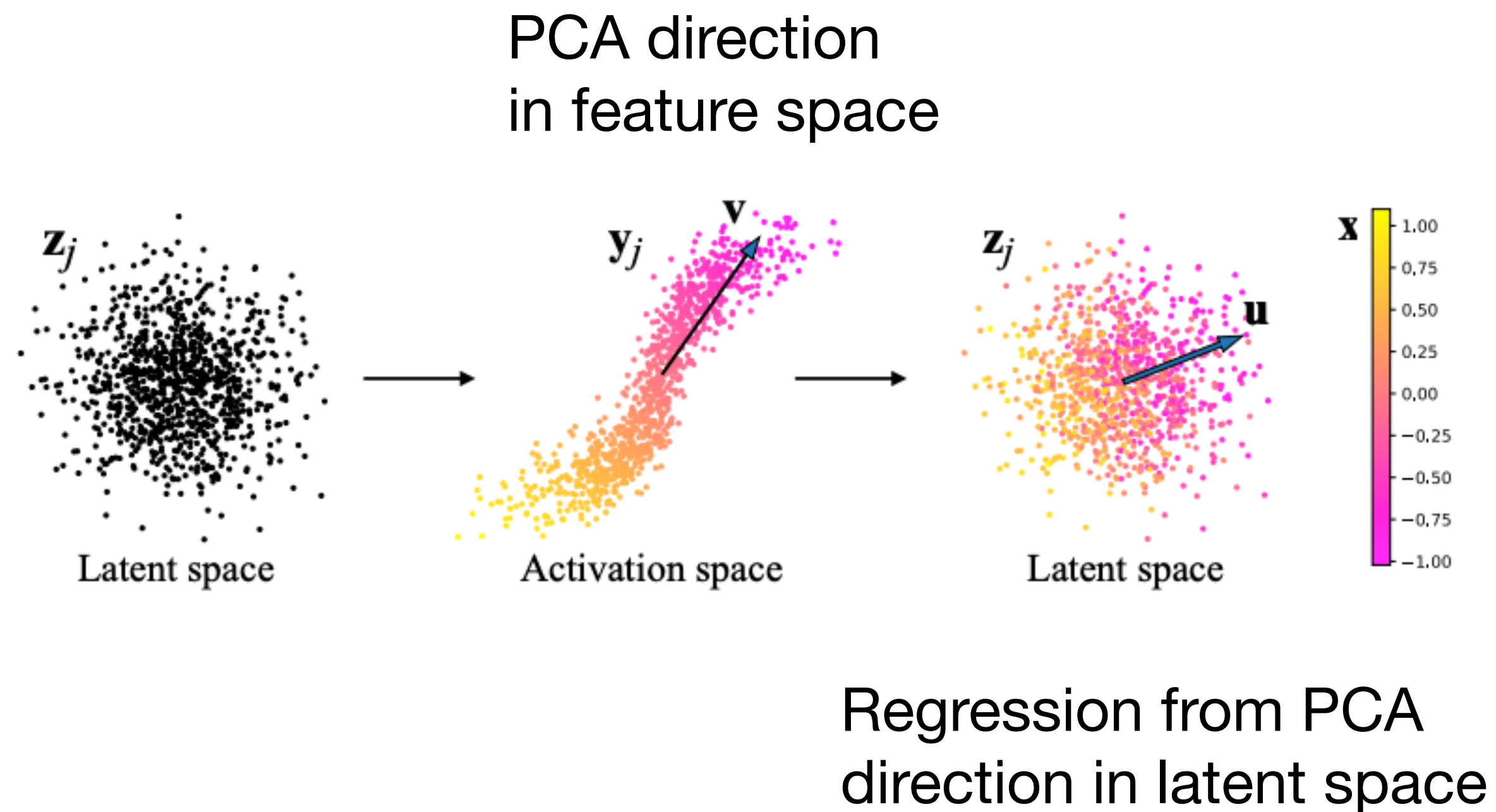
Human-in-the-loop
AI content creation



<https://genforce.github.io/sefa>

Unsupervised Approach

GANspace: PCA applied to the latent space of StyleGAN



Unsupervised Approach

Hessian Penalty: A weak prior for unsupervised disentanglement.

$$I = G(z)$$

Hessian Matrix:

$$H_{ij} = \frac{\partial^2 G}{\partial z_i \partial z_j} = \frac{\partial}{\partial z_j} \left(\frac{\partial G}{\partial z_i} \right) = 0.$$

Hessian penalty in training:

$$\mathcal{L}_H(G) = \sum_{i=1}^{|z|} \sum_{j \neq i}^{|z|} H_{ij}^2.$$

The Hessian Penalty:
A Weak Prior for Unsupervised Disentanglement

William Peebles¹, John Peebles², Jun-Yan Zhu³, Alexei Efros¹, Antonio Torralba⁴

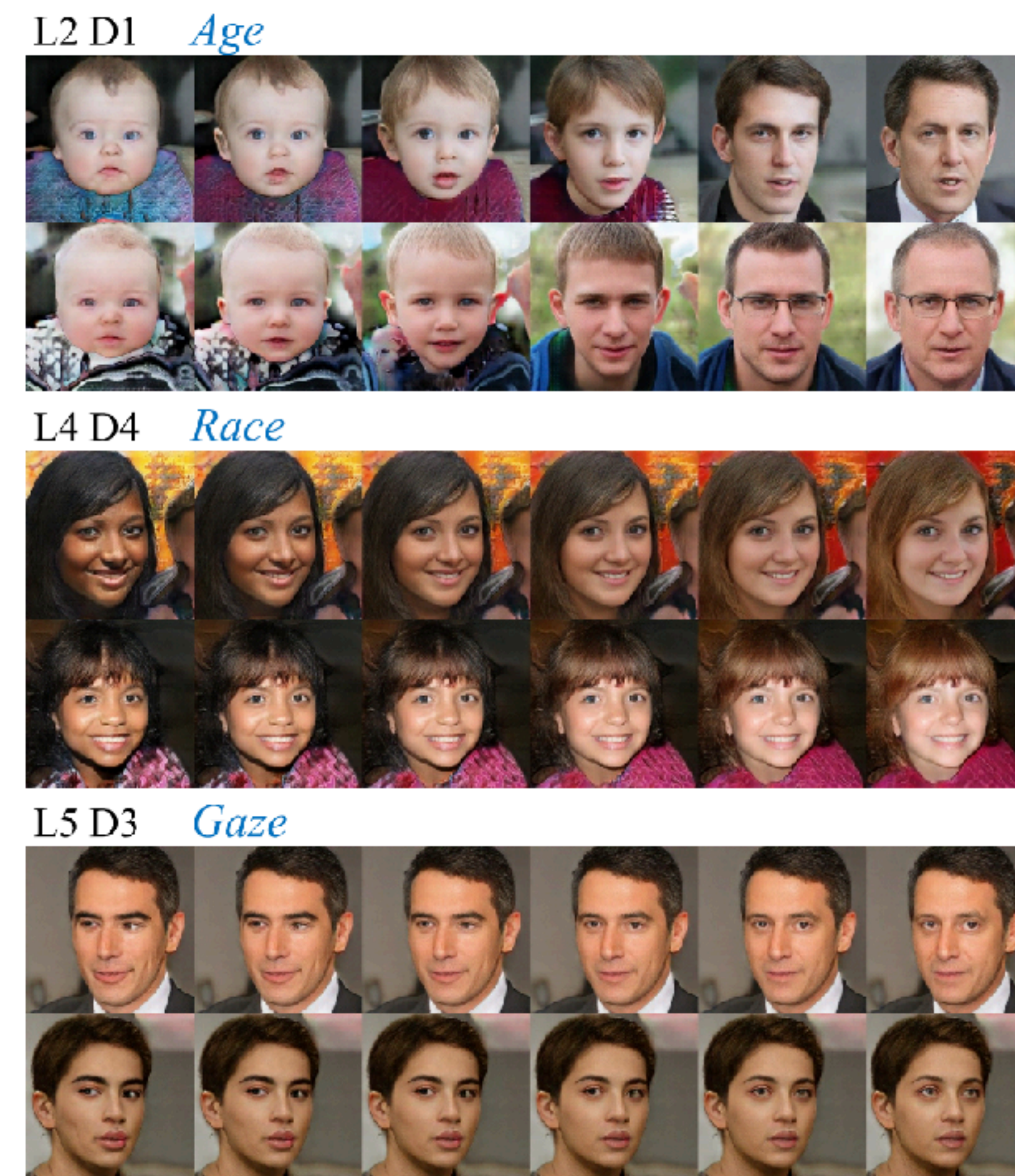
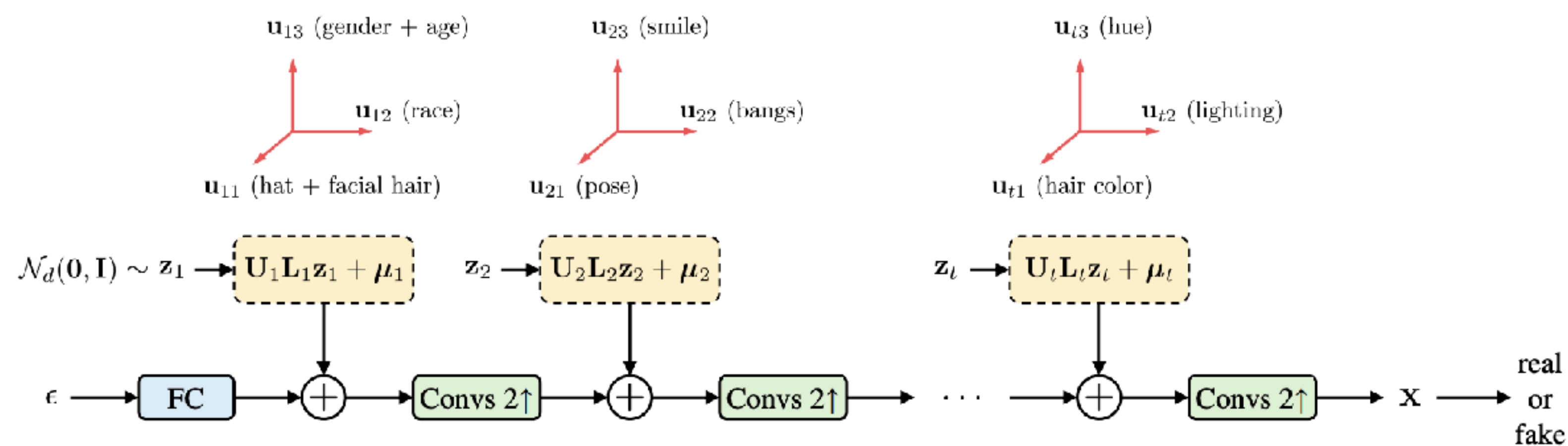
¹UC Berkeley, ²Yale University, ³Adobe Research, ⁴MIT CSAIL

Unsupervised Approach

EigenGAN: Layer-Wise Eigen-Learning for GANs

Design inductive bias of disentanglement in the generator:



Challenges for Unsupervised Approach

- How to evaluate the results?
- How to annotate each disentangled dimensions?
- How to improve the disentanglement in GAN training?

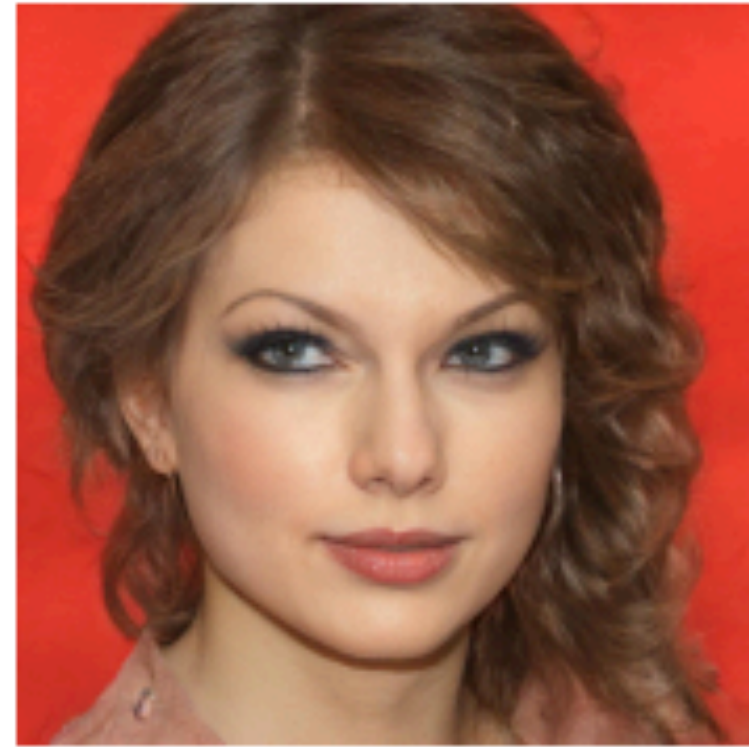
Interpretation Approaches

- **Supervised approach:** use labels or trained classifiers to probe the representation of the generator
- **Unsupervised approach:** identify the controllable dimensions of generator without labels/classifiers
- **Zero-shot approach:** align language embedding with generative representations

Zero-Shot Approach

StyleCLIP: CLIP + StyleGAN

Source:



Text input:

“Mohawk hairstyle”

“Without makeup”

“Cute cat”

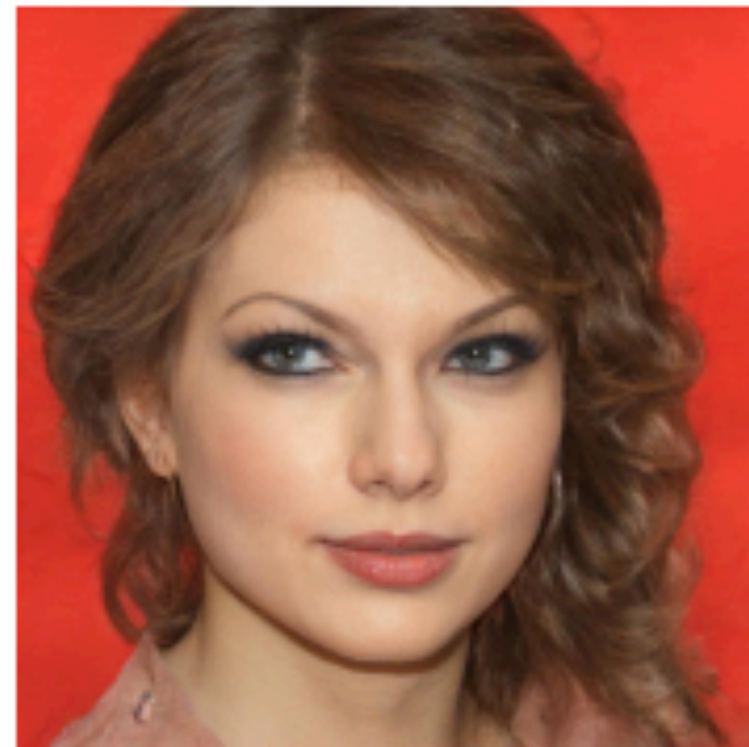
“Lion”

“Gothic church”

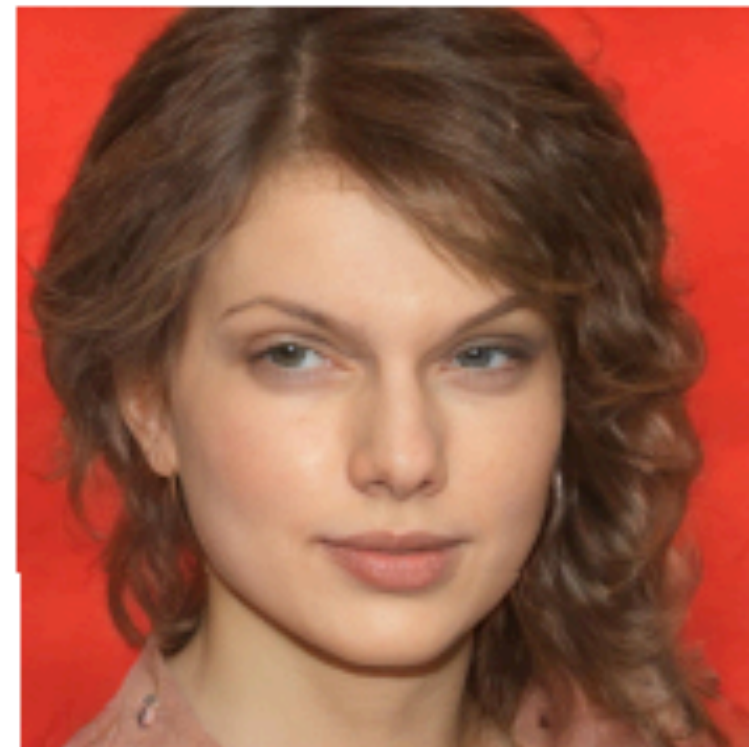
Zero-Shot Approach

StyleCLIP: CLIP + StyleGAN

Source:



Output:



Text input:

“Mohawk hairstyle”

“Without makeup”

“Cute cat”

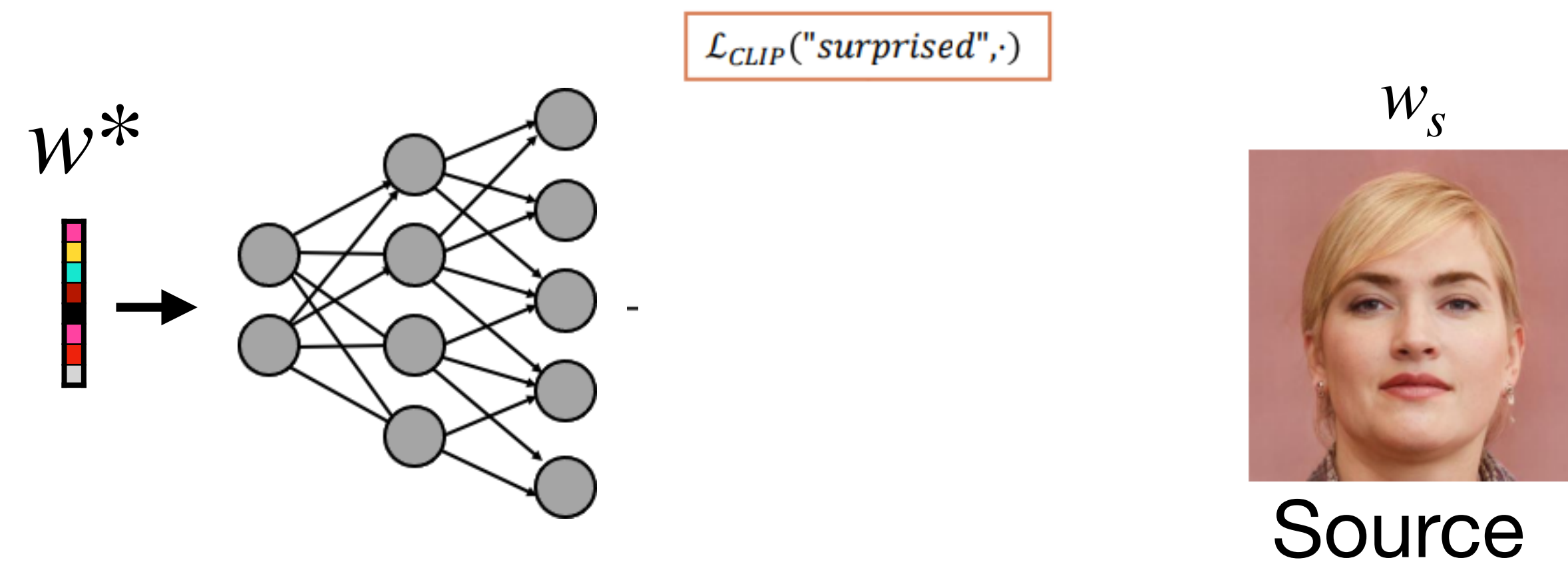
“Lion”

“Gothic church”

Zero-Shot Approach

StyleCLIP: CLIP + StyleGAN

- Contrastive Language-Image Pre-training (CLIP): pretrained model from 400 million image-text pairs: <https://github.com/openai/CLIP>



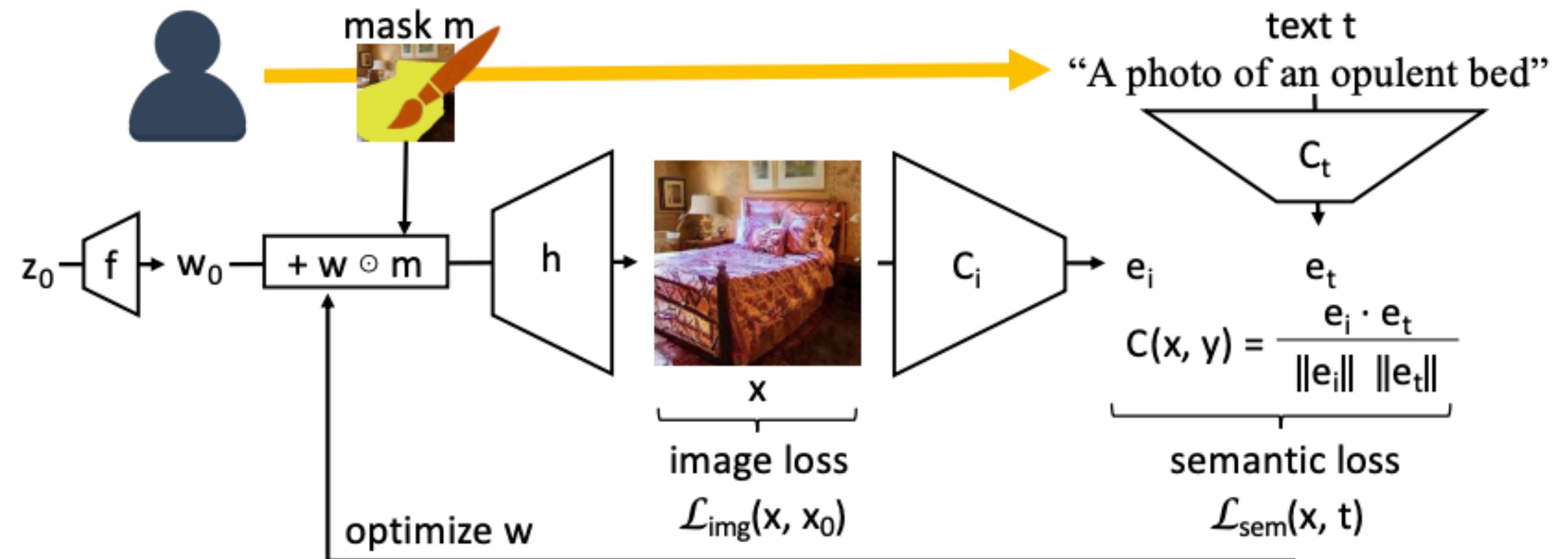
$$w^* = \arg \min_{w \in \mathcal{W}+} D_{CLIP}(G(w), t) + \lambda_{L2} \|w - w_s\|_2 + \lambda_{ID} \mathcal{L}_{ID}(w),$$

CLIP's joint embedding space

Regularizers on similarity and identity

$$\mathcal{L}_{ID}(w) = 1 - \langle R(G(w_s)), R(G(w)) \rangle$$

Paint by Word: CLIP + Region-based StyleGAN inversion



Zero-Shot Approach

Massive data-driven OpenAI DALL.E

12-billion parameter model trained on **250 million text-images pairs** from the internet

1. Train a discrete variational autoencoder (dVAE)
2. Train an autoregressive transformer to model the joint distribution of text and image tokens



(a) a tapir made of accordion. (b) an illustration of a baby hedgehog in a christmas sweater walking a dog (c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

Interpretation Approaches

- **Supervised approach:** use labels or trained classifiers to probe the representation of the generator
- **Unsupervised approach:** identify the controllable dimensions of generator without labels/classifiers
- **Zero-shot approach:** align language embedding with generative representations

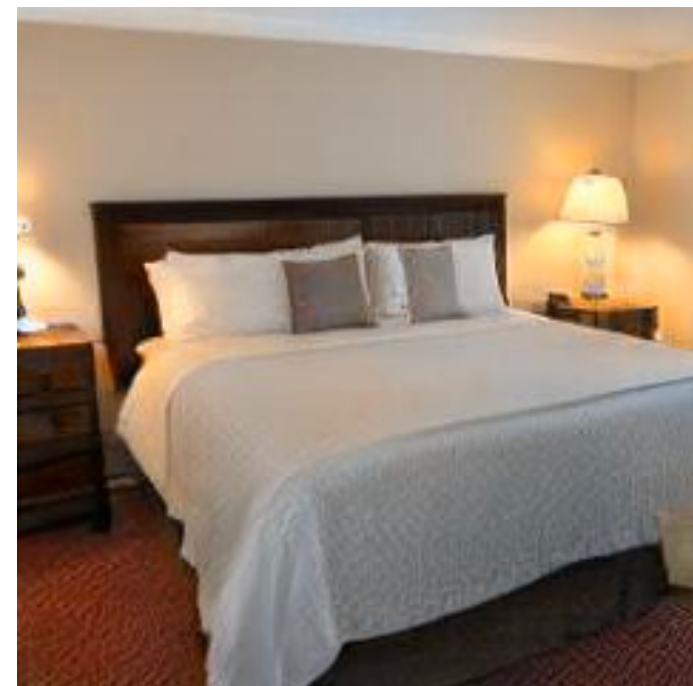
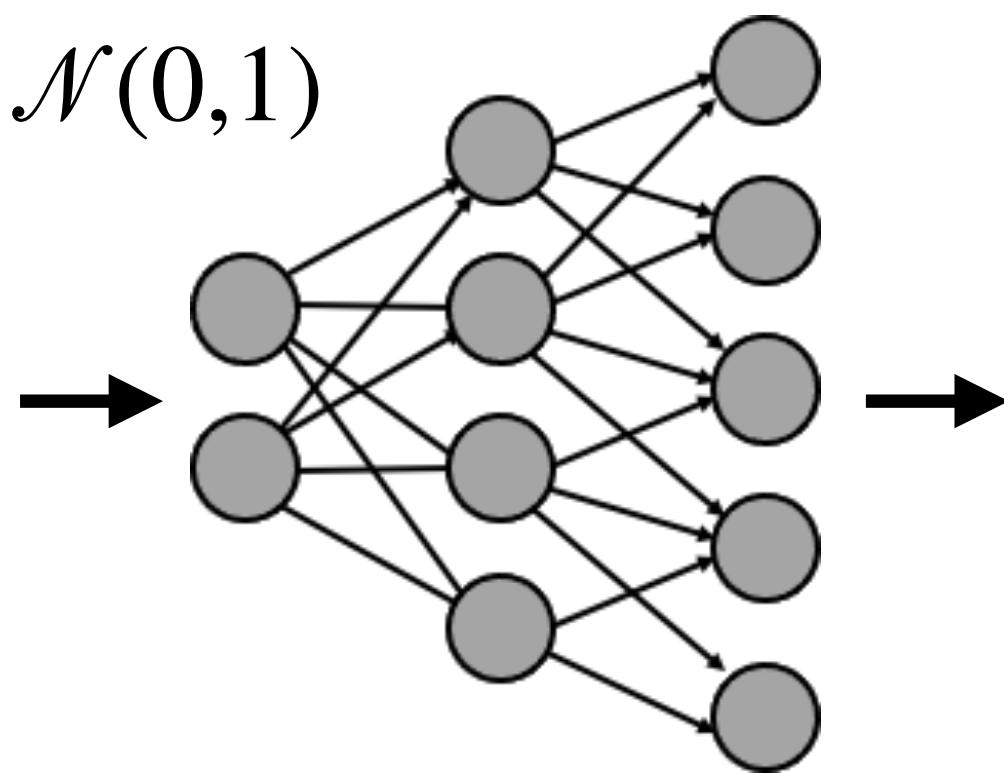
Latent Spaces of GAN's Generator

Z space, W space, StyleSpace (S space), W+ space, P/P+ space

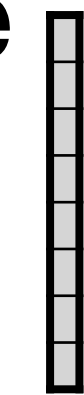
Z space $z \sim \mathcal{N}(0,1)$ W space $w = F(z)$ ← Most of the linear manipulation

Z space

$z \sim \mathcal{N}(0,1)$



S space $\{y_i = A_i(w)\}$
(Layer-wise codes)



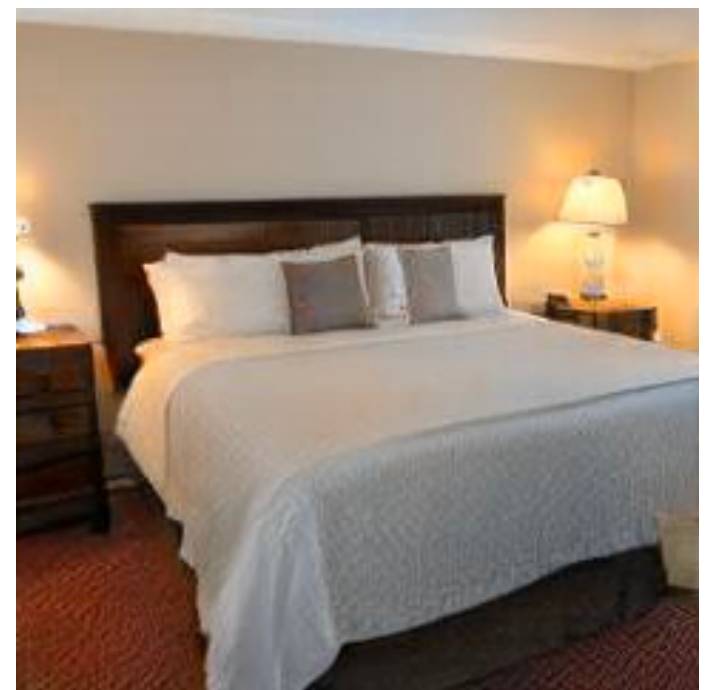
StyleGAN

AdaIN

AdaIN

AdaIN

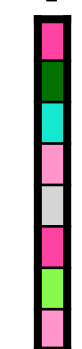
AdaIN



P space: $\text{LeakyReLU}_5(w)$

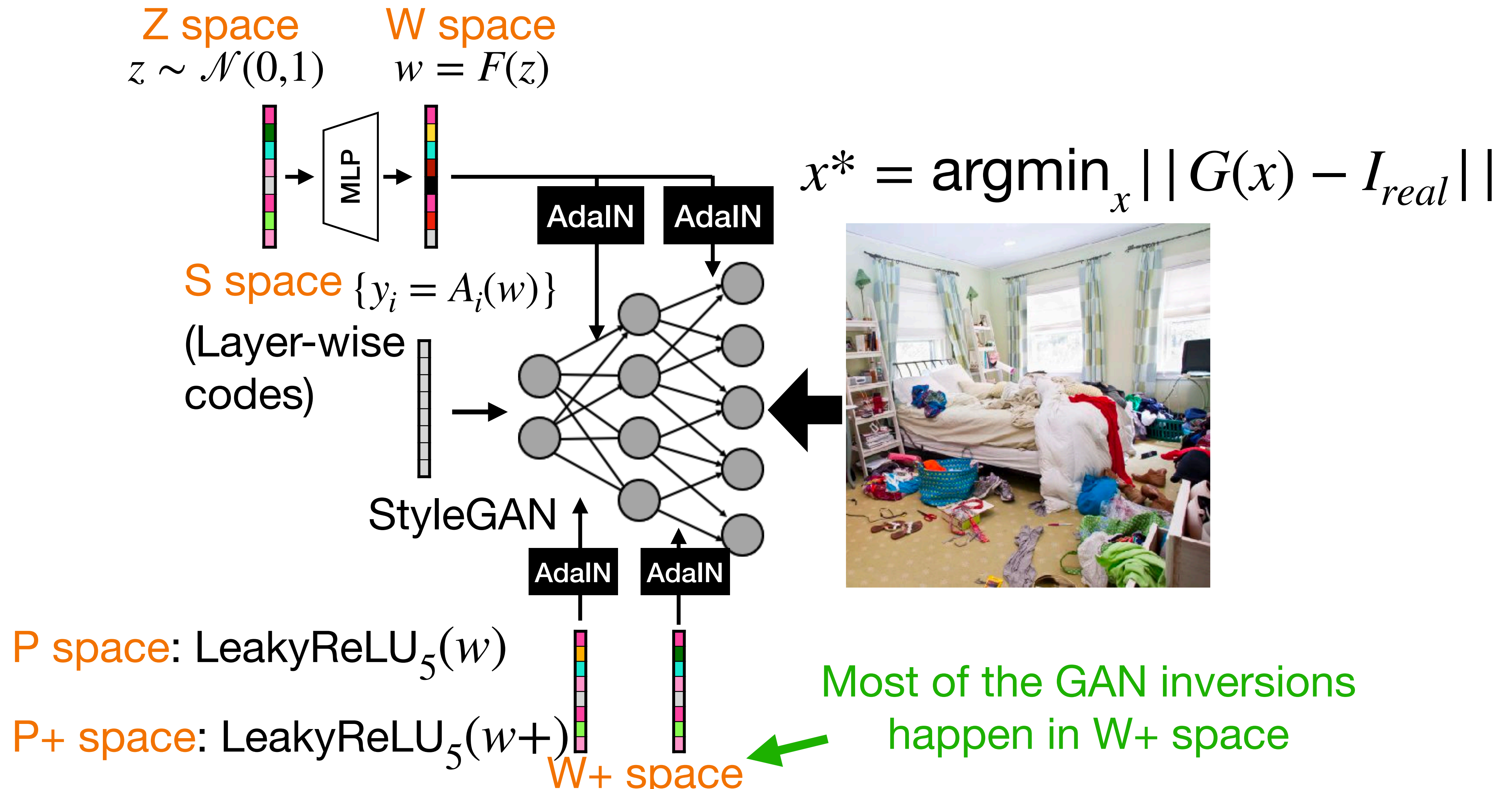
P+ space: $\text{LeakyReLU}_5(w+)$

W+ space

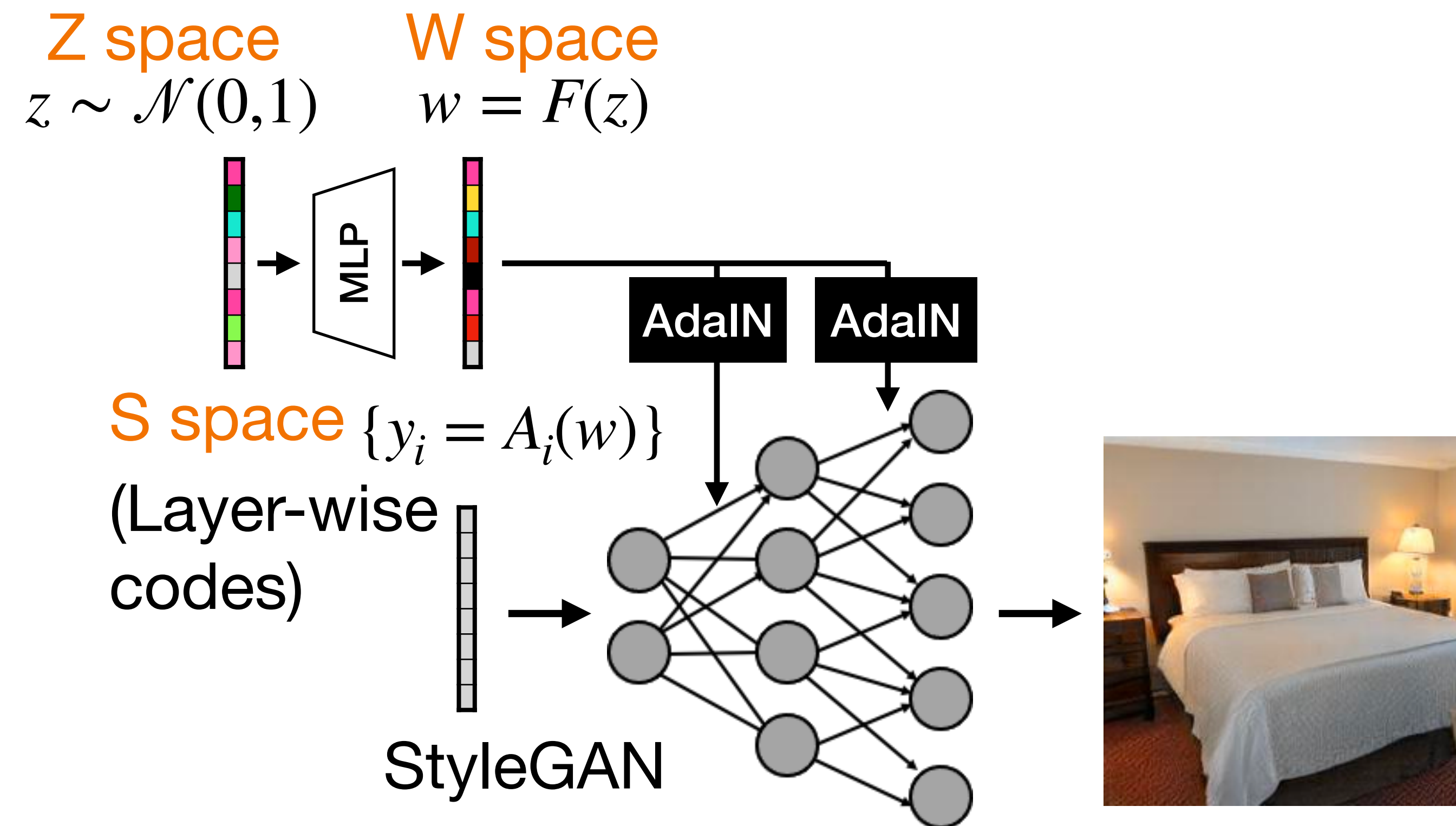


Latent Spaces of GAN's Generator

Z space, W space, StyleSpace (S space), W+ space, P/P+ space



Which latent space is more disentangled?



Reconstruction Error (Xu et al. CVPR'21)

Space	MSE	FID
W space	0.0601	22.24
S space	0.0464	18.48

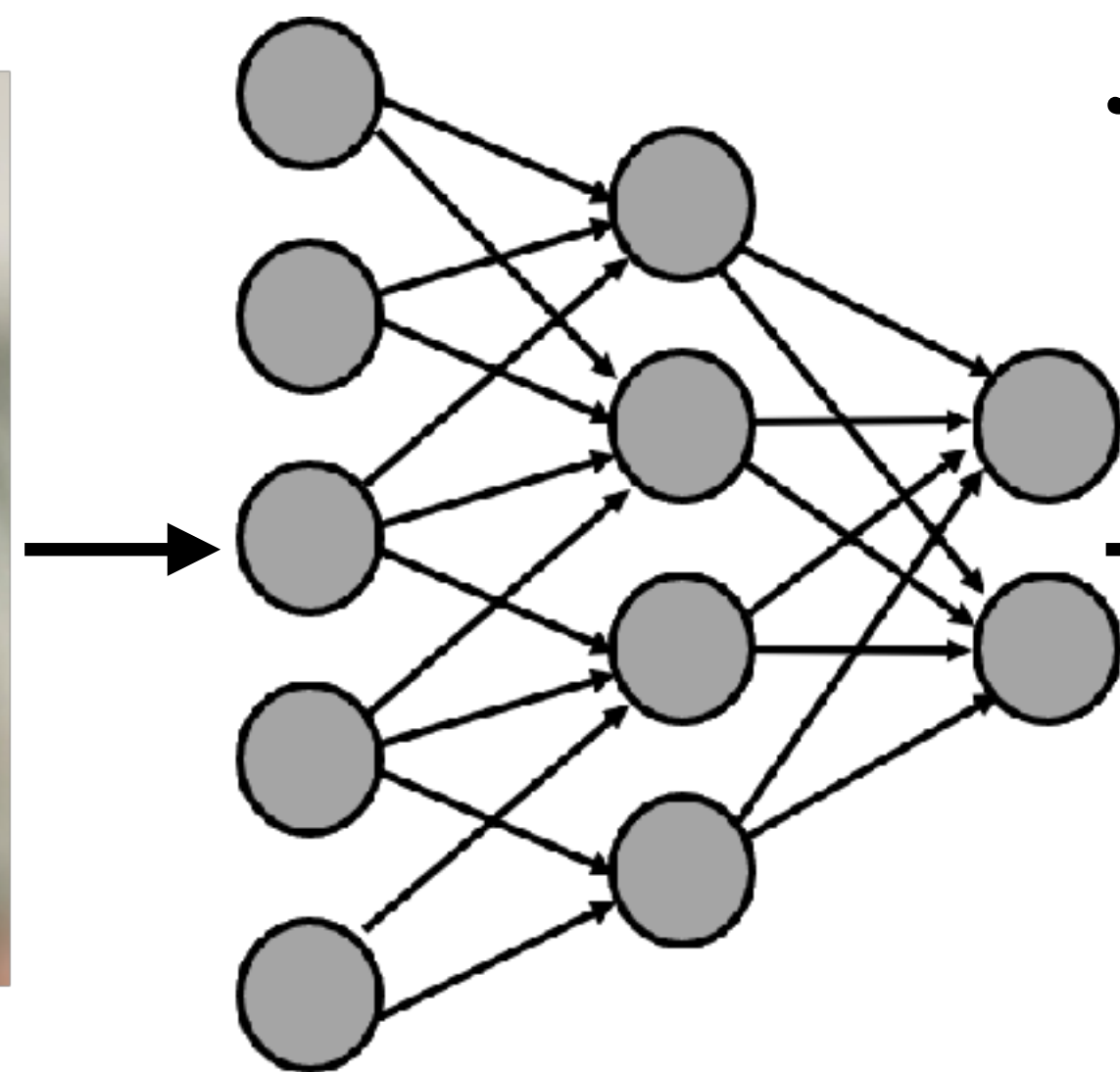
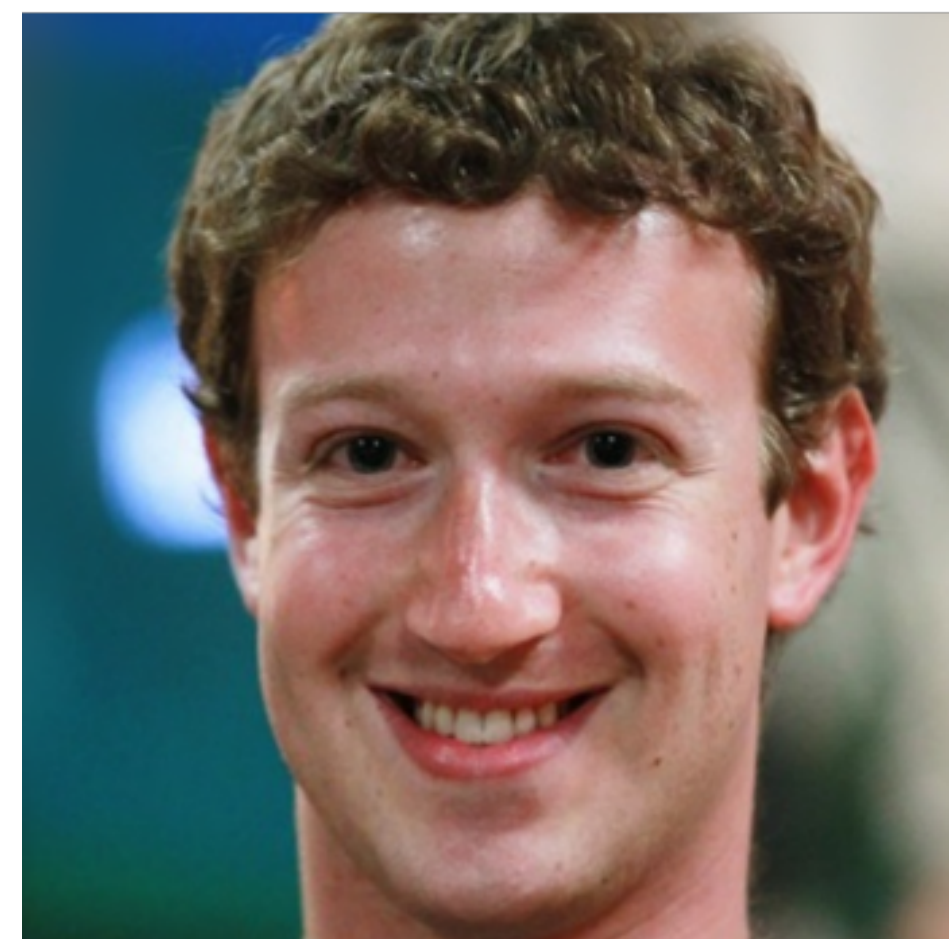
Disentanglement (Wu et al. CVPR'21)

Space	Disentanglement
Z space	0.31
W space	0.54
S space	0.75

Encoding Real Image into StyleGAN space

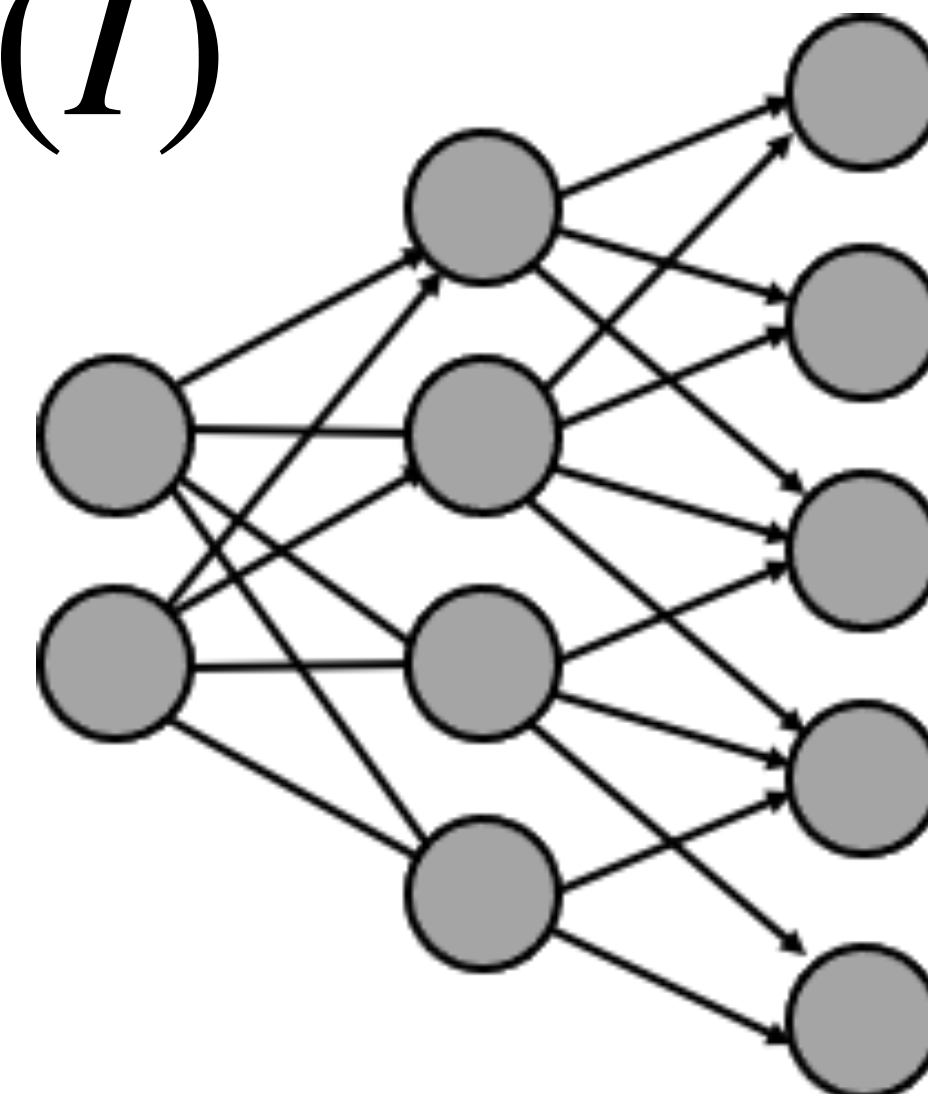
Which latent space to use?

I



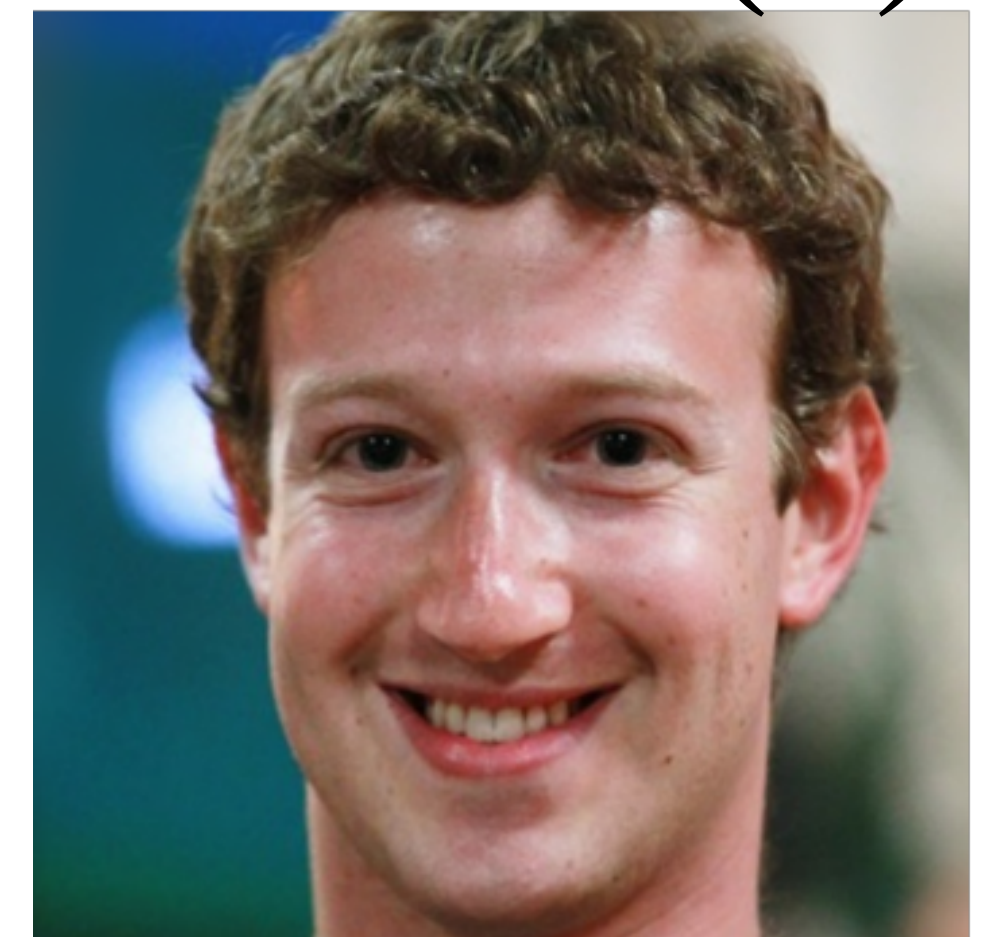
Encoder

$$x = E(I)$$



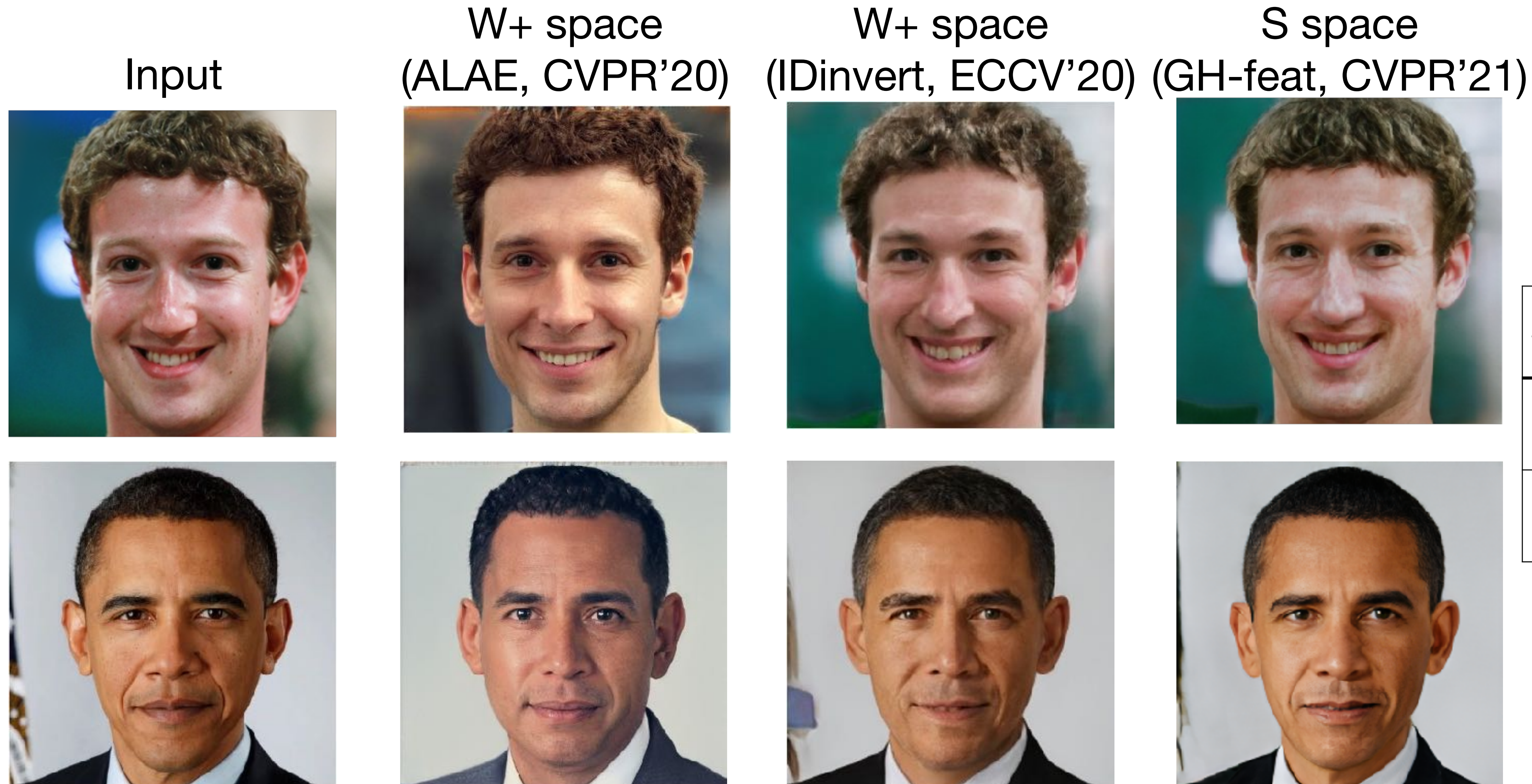
StyleGAN
Generator

$$I' = G(x)$$



Encoding Real Image into StyleGAN space

Which latent space to use?



Space	MSE	FID
W+ space	0.0601	22.24
S space	0.0464	18.48

Generative Image Prior

Applying the pretrained GAN model to image processing tasks

GAN inversion:

$$x^* = \operatorname{argmin}_x ||G(x) - I||$$

Colorization:

$$x^* = \operatorname{argmin}_x ||\operatorname{rgb2gray}(G(x)) - I_{\text{gray}}||$$

Super-resolution:

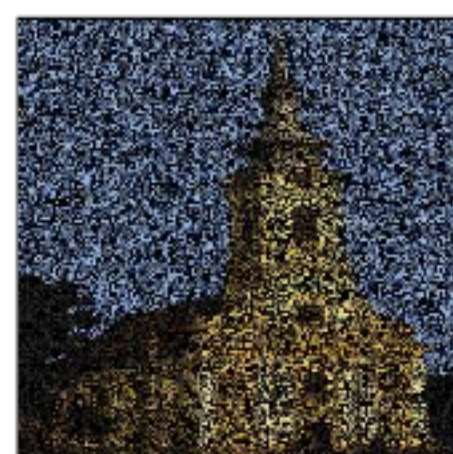
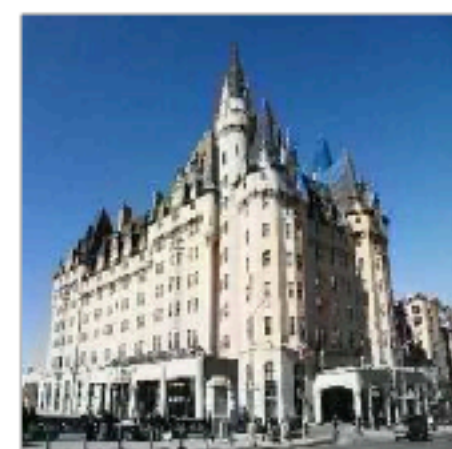
$$x^* = \operatorname{argmin}_x ||\operatorname{down}(G(x)) - I_{\text{small}}||$$



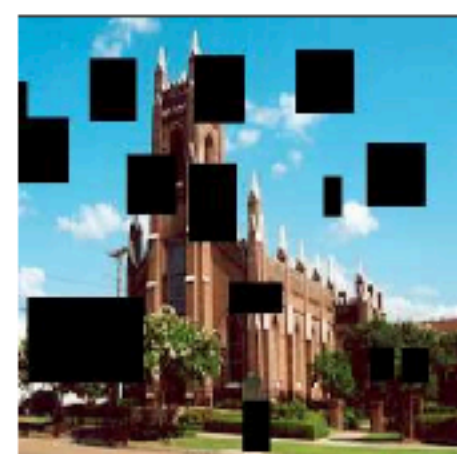
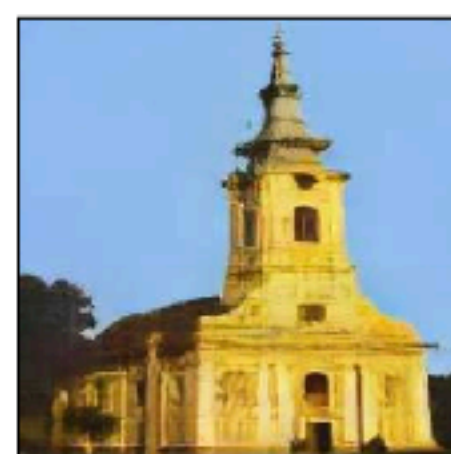
(a) Image Reconstruction



(b) Image Colorization



(d) Image Denoising

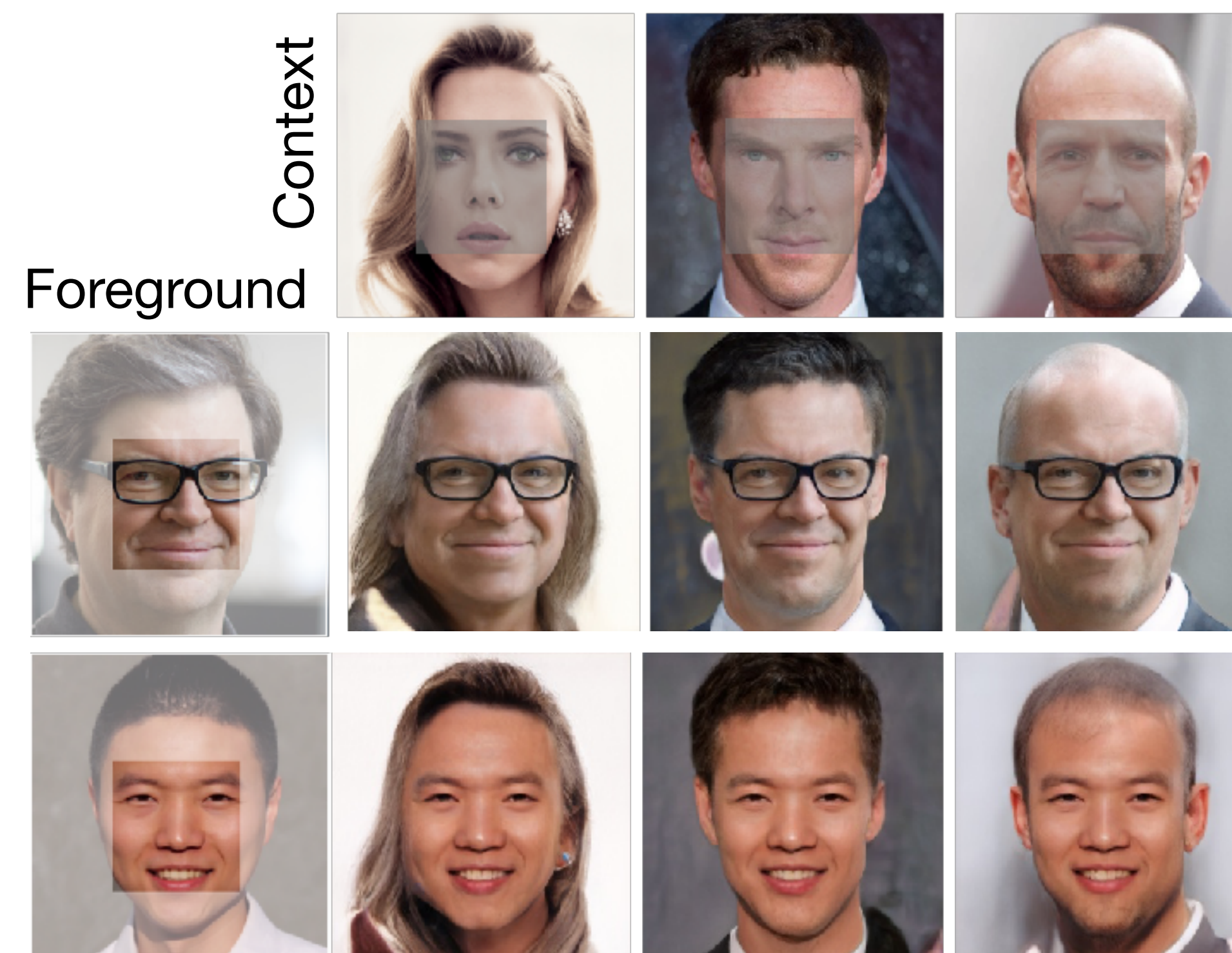


(e) Image Inpainting



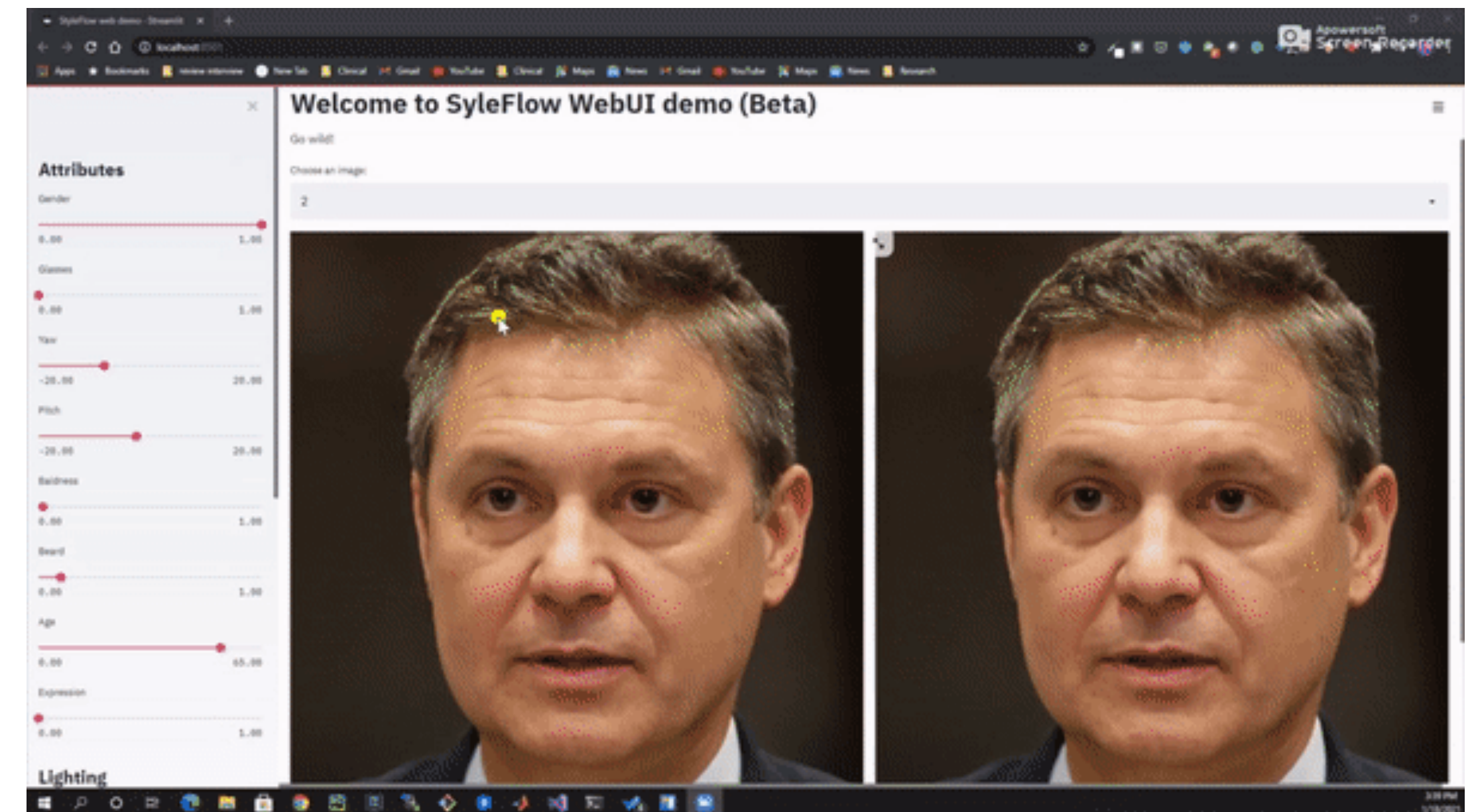
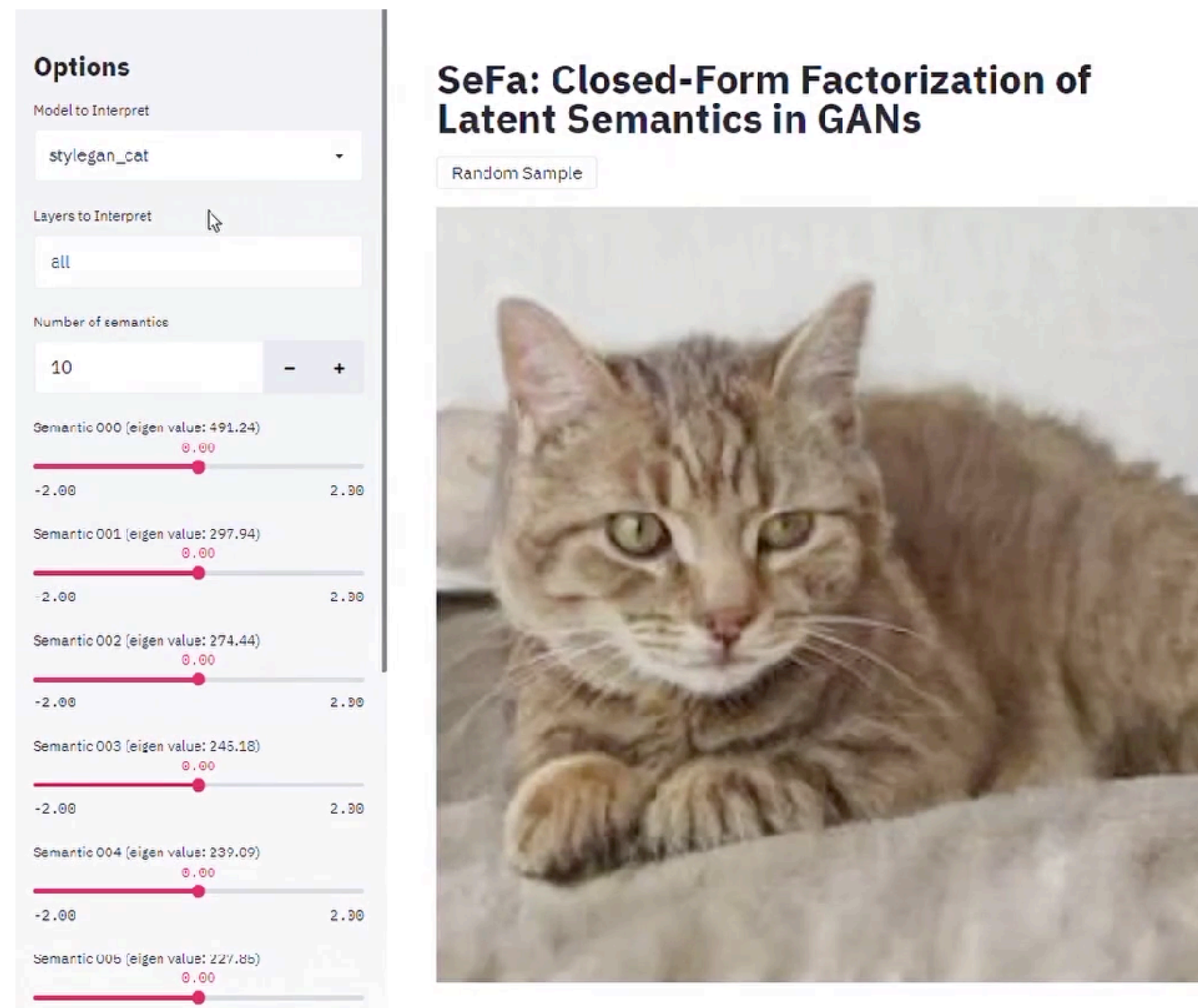
Masked optimization

$$x^* = \operatorname{argmin}_x ||m \cdot G(x) - m \cdot I_{\text{context}}||$$



Summary

Interpreting generative models facilitates interactive content creation and human-AI collaboration.



Please refer to the recent survey paper on GAN Inversion: <https://arxiv.org/pdf/2101.05278.pdf>