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

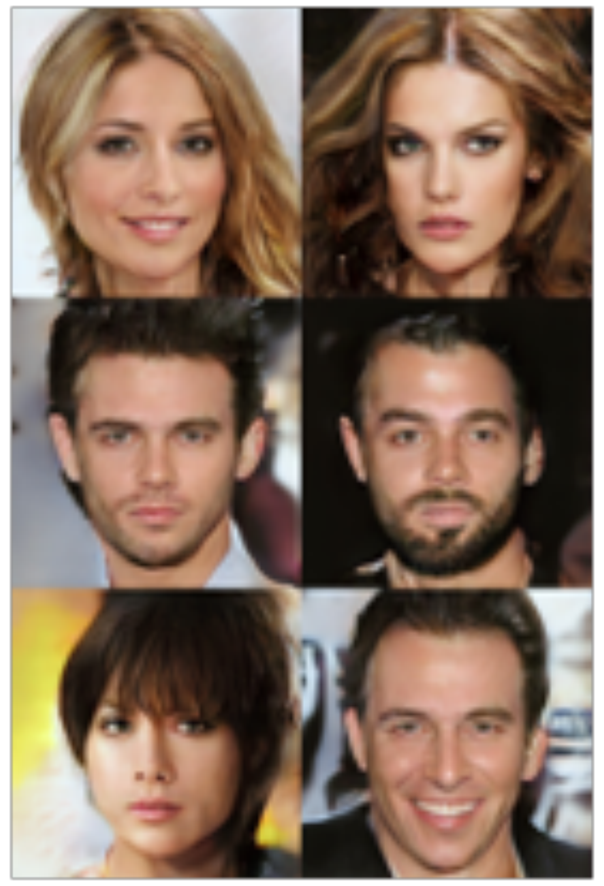
2014

GAN

2015

DCGAN

2017



PG-GAN

2018

StyleGAN

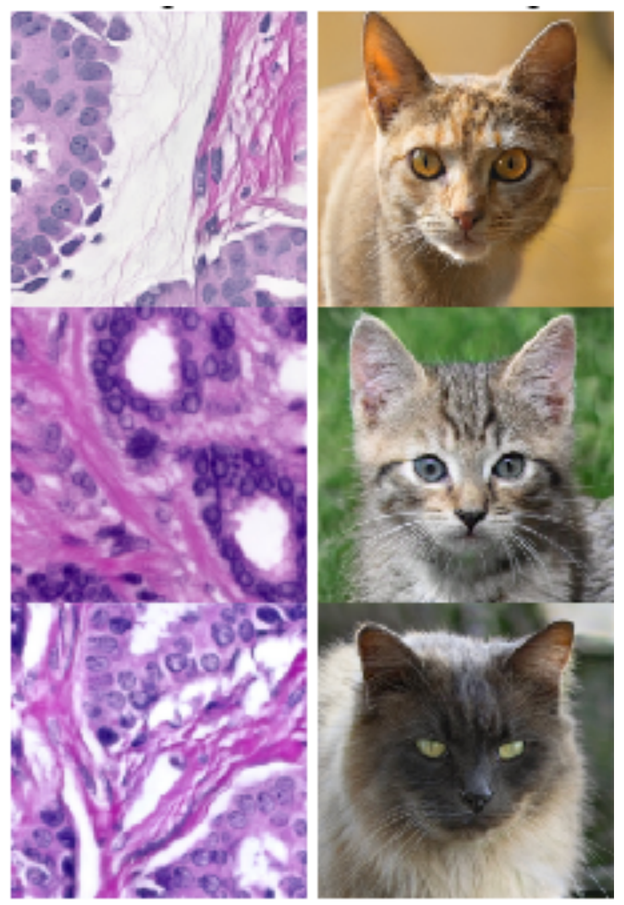
2018

BigGAN

2019

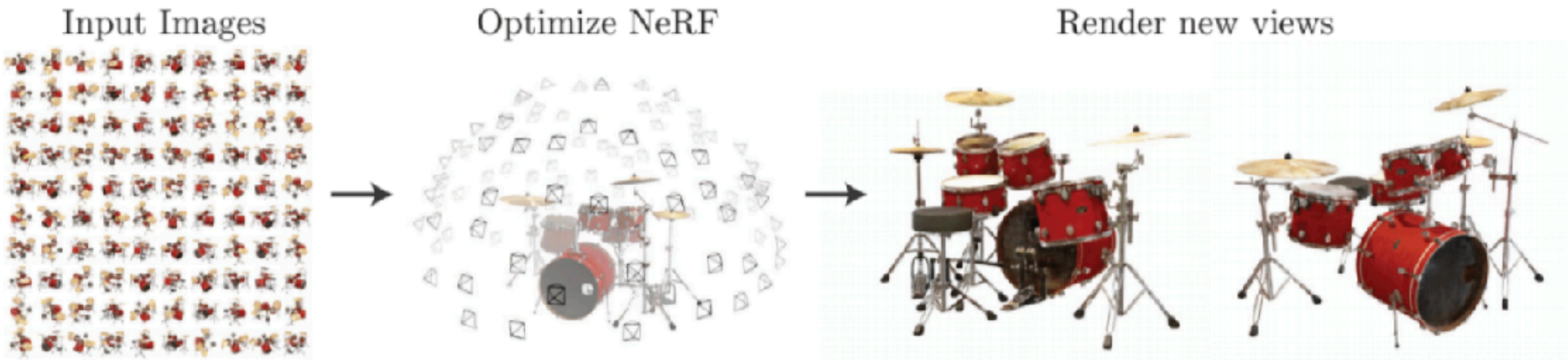
StyleGANv2

2020



StyleGAN-ADA

2020: NeRF (Neural Radiance Fields)

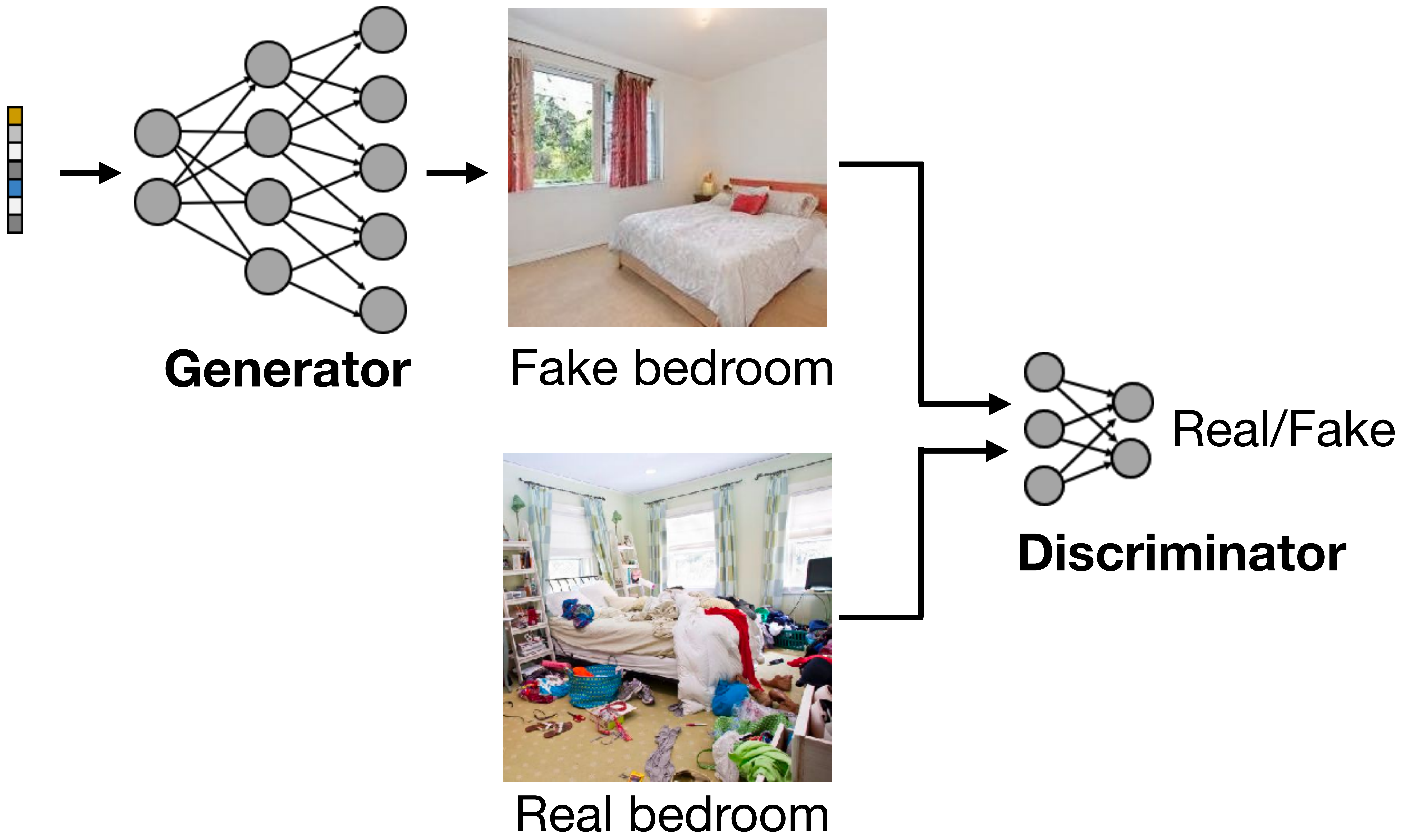


2021: OpenAI DALLE (VQ-VAE)

Arm chair in shape of avocado

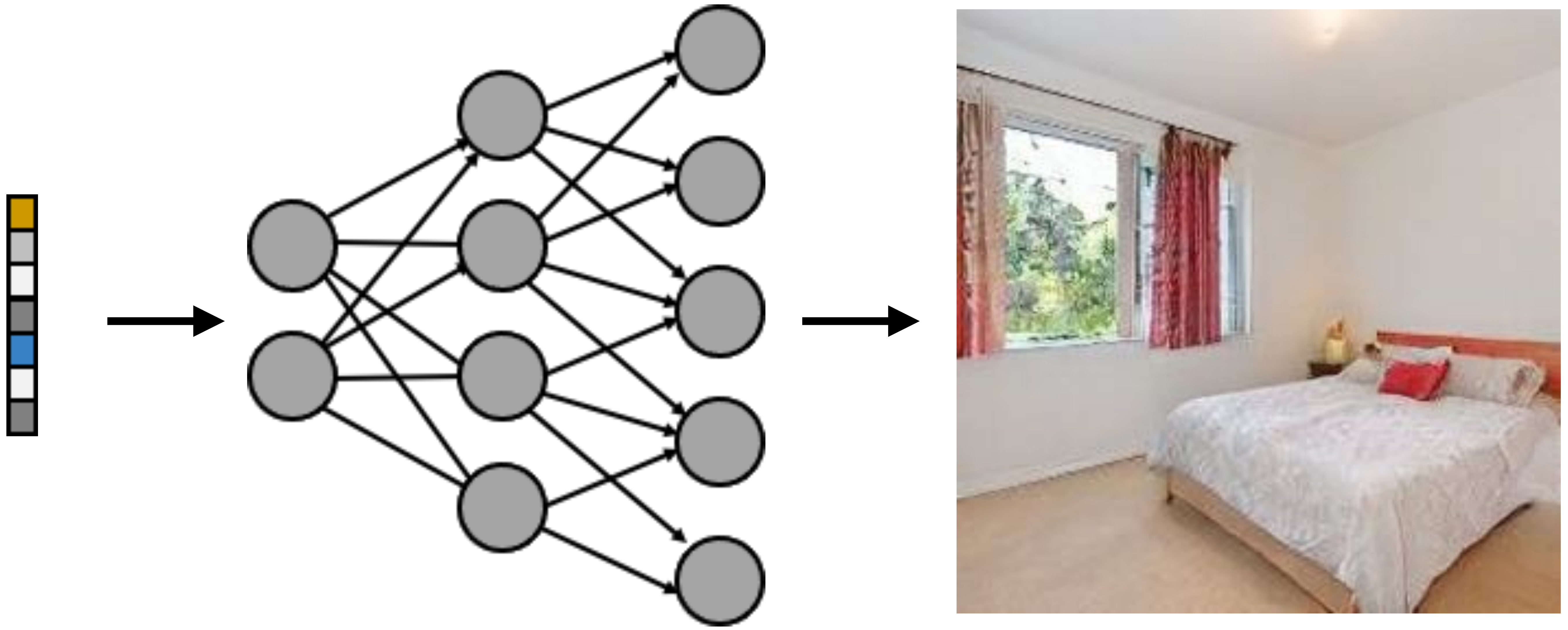
Generative Adversarial Networks (GANs)

Adversarial Training



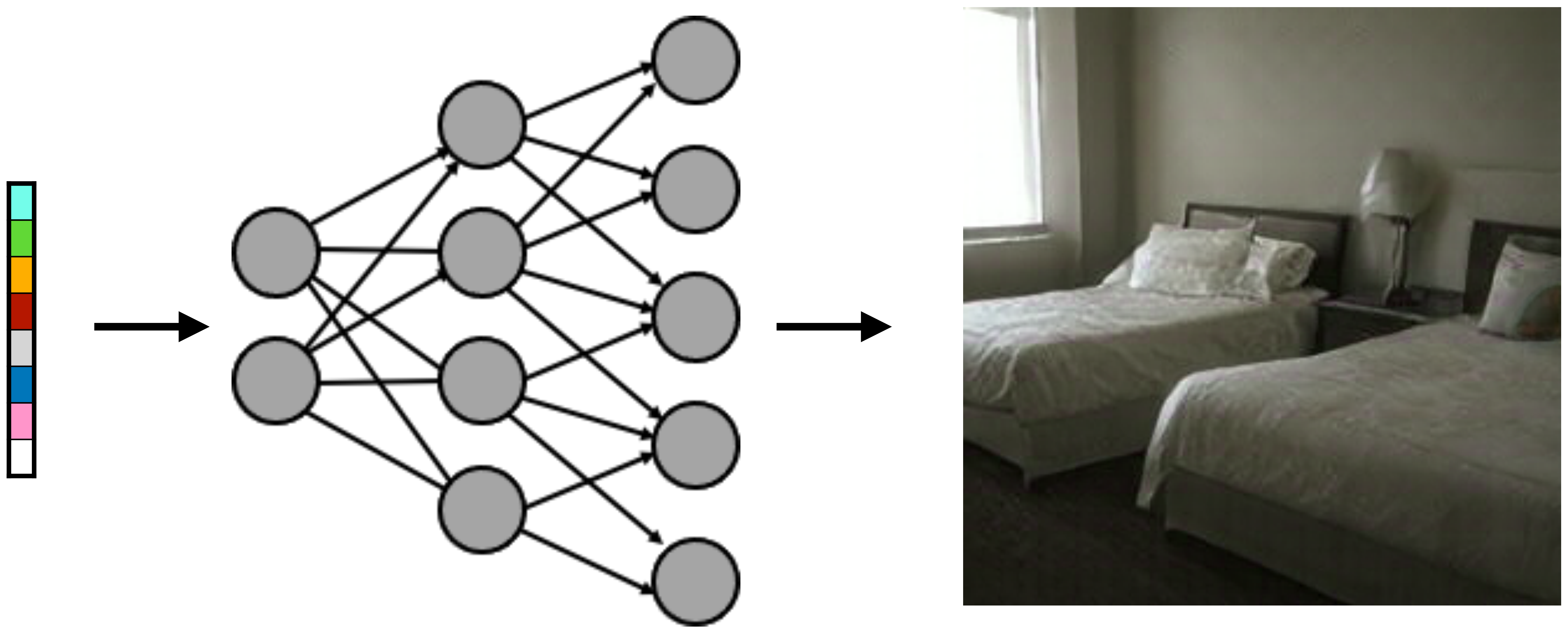
Goodfellow et al. NeurIPS'14

Neural Image Generation

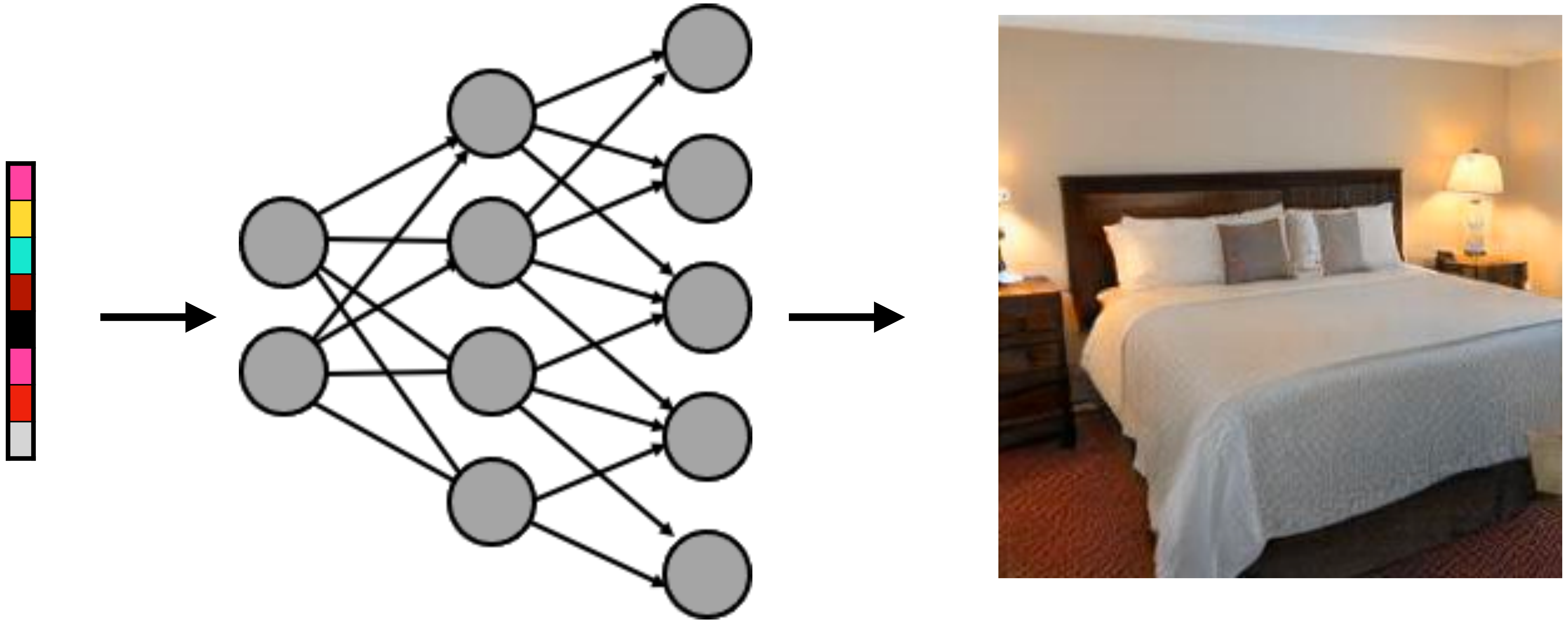


Goodfellow et al. NeurIPS'14

Neural Image Generation

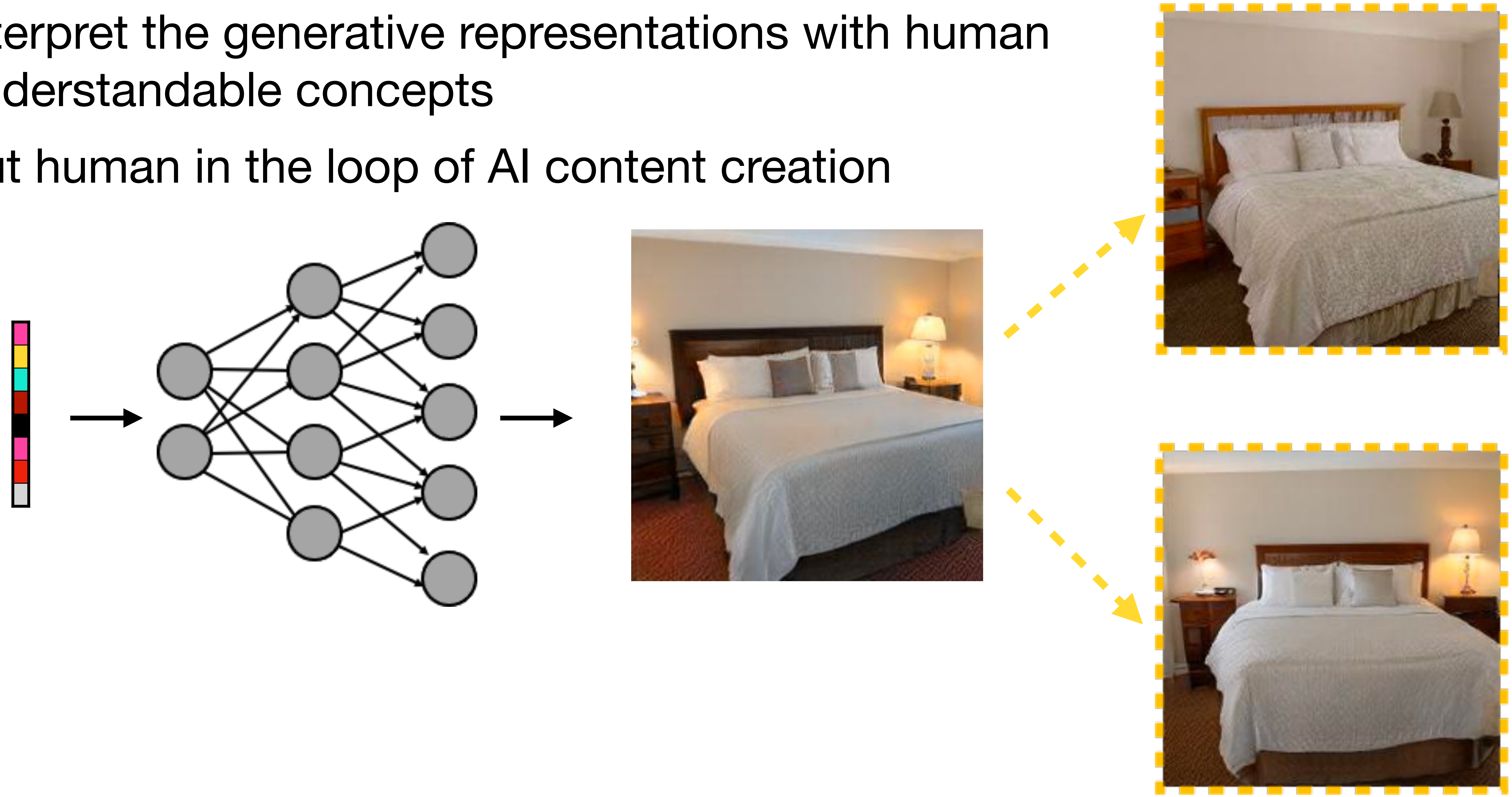


Neural Image Generation

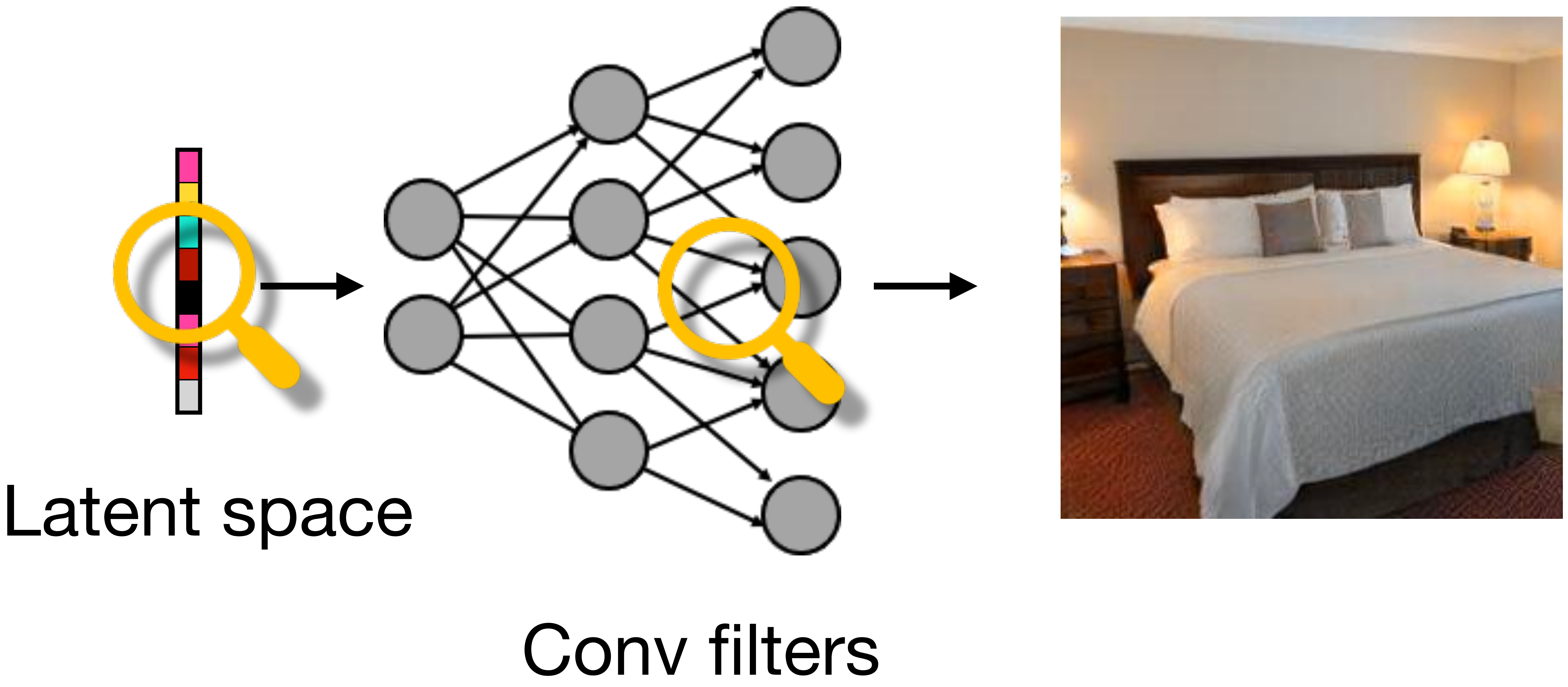


How to Steer Neural Image Generation?

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



Deep Generative Representations



Interpretation Approaches

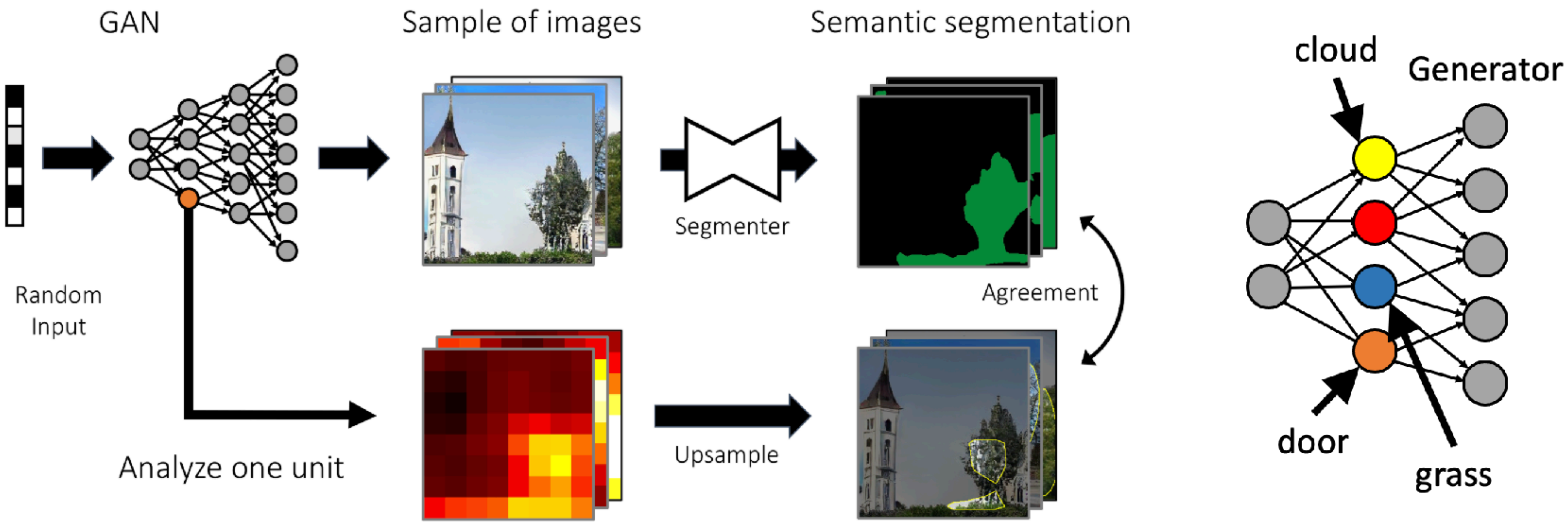
- **Supervised approach:** use labels or trained classifiers to probe the representation of the generator
- **Unsupervised approach:** identify the controllable dimensions of generator without labels/classifiers
- **Zero-shot approach:** align language embedding with generative representations

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

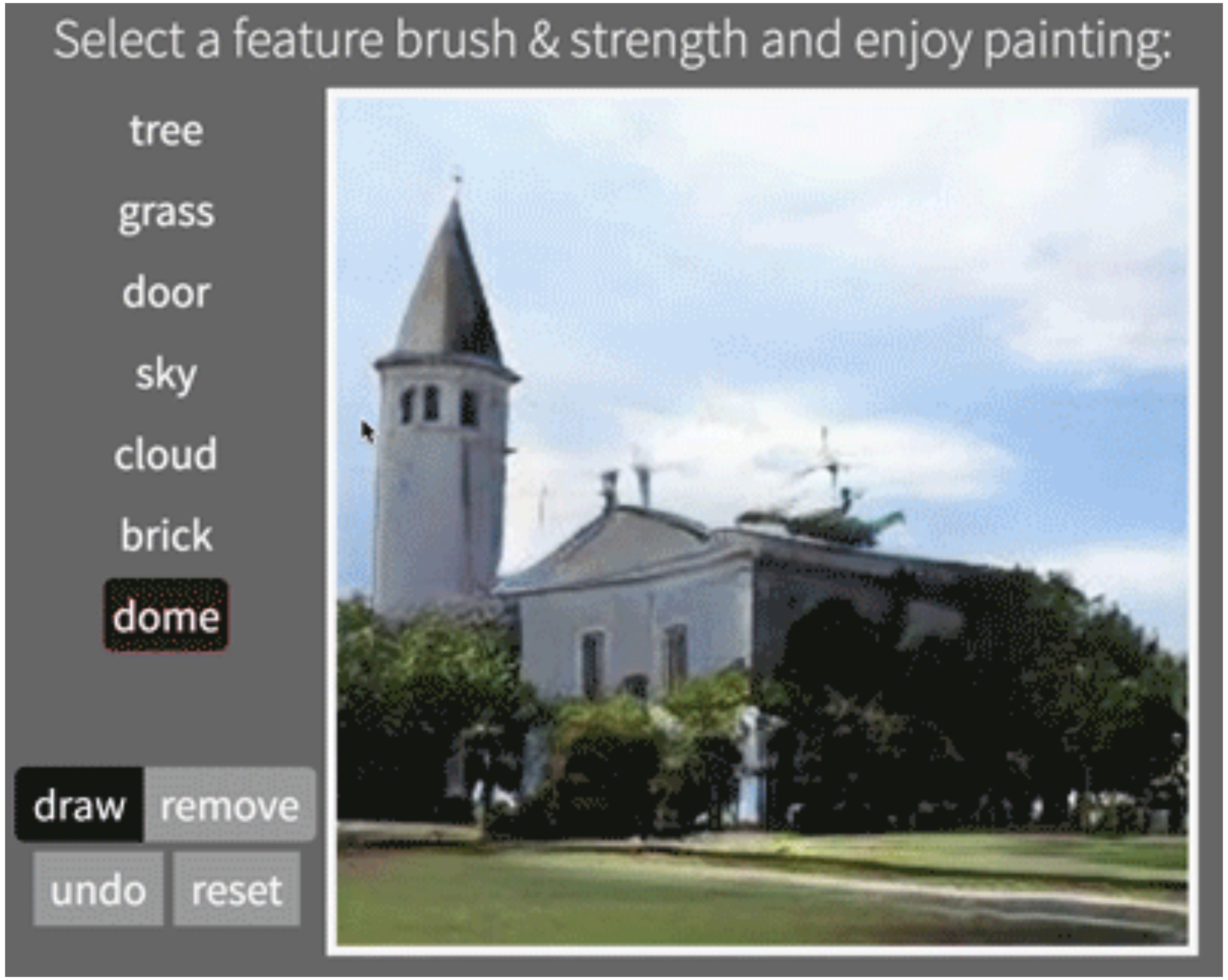
GAN Dissection: Aligning semantic segmentation with GAN feature map

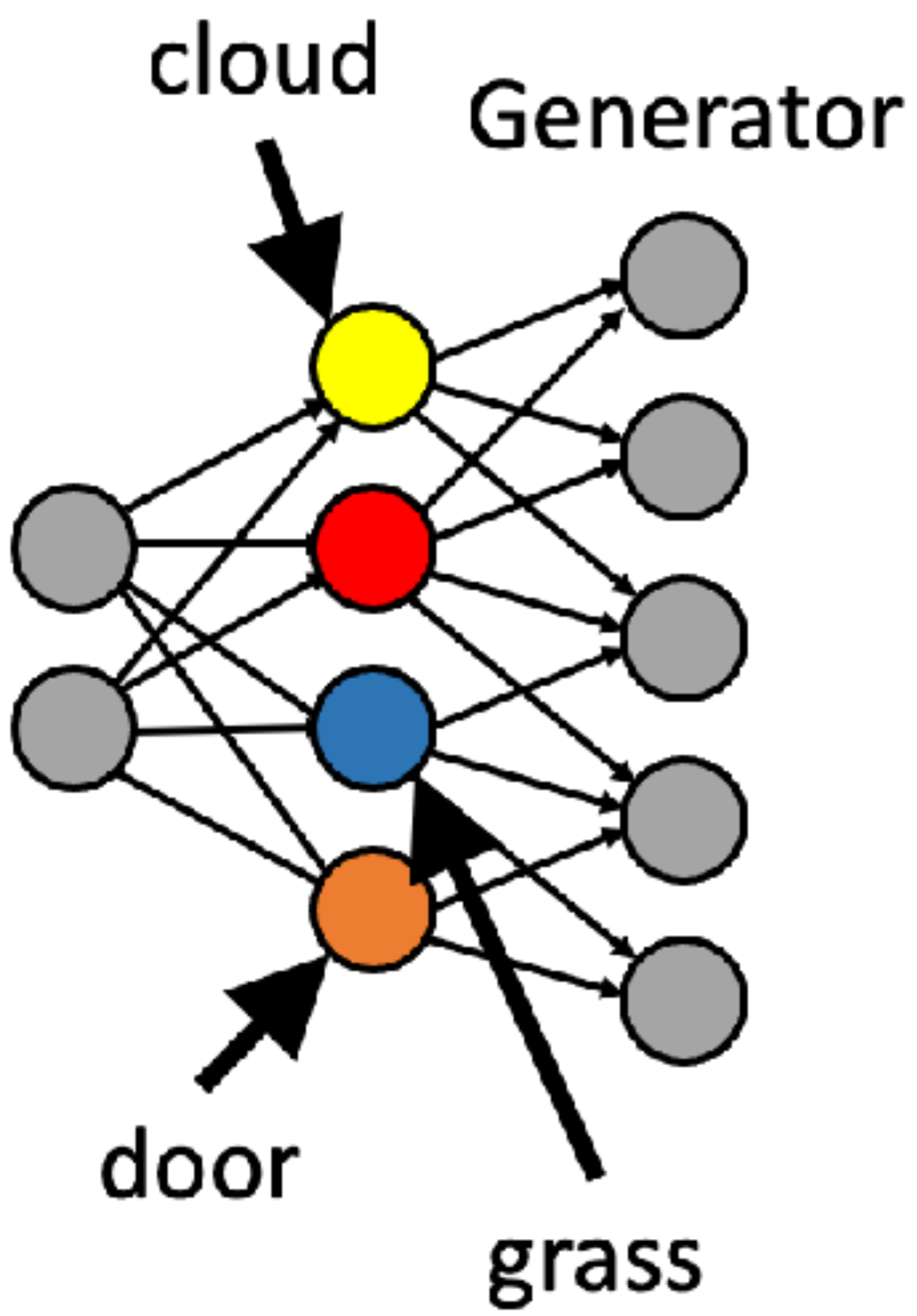


Bau, Zhu, Strobelt, Zhou, Tenenbaum, Freeman, Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. ICLR’19

Supervised Approach

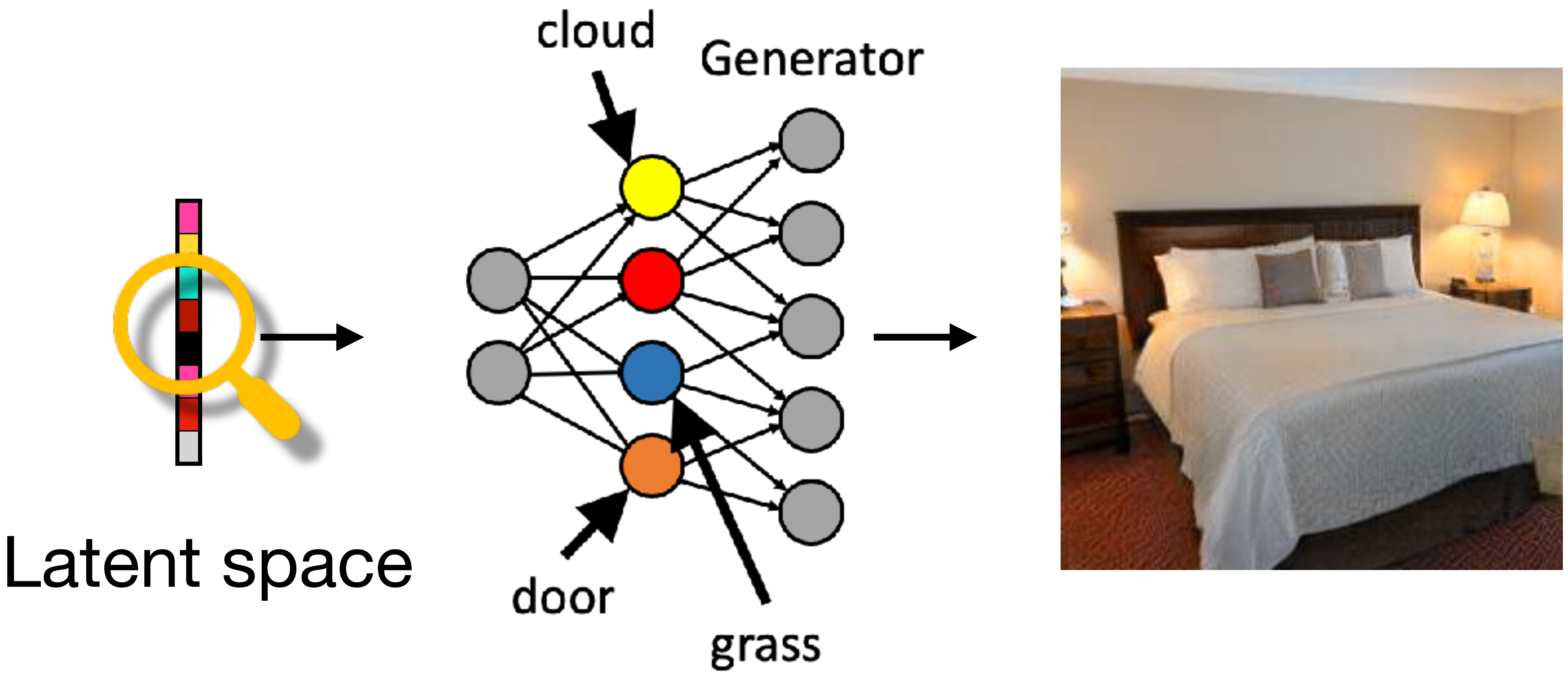
GAN Dissection: Aligning semantic segmentation with GAN feature map





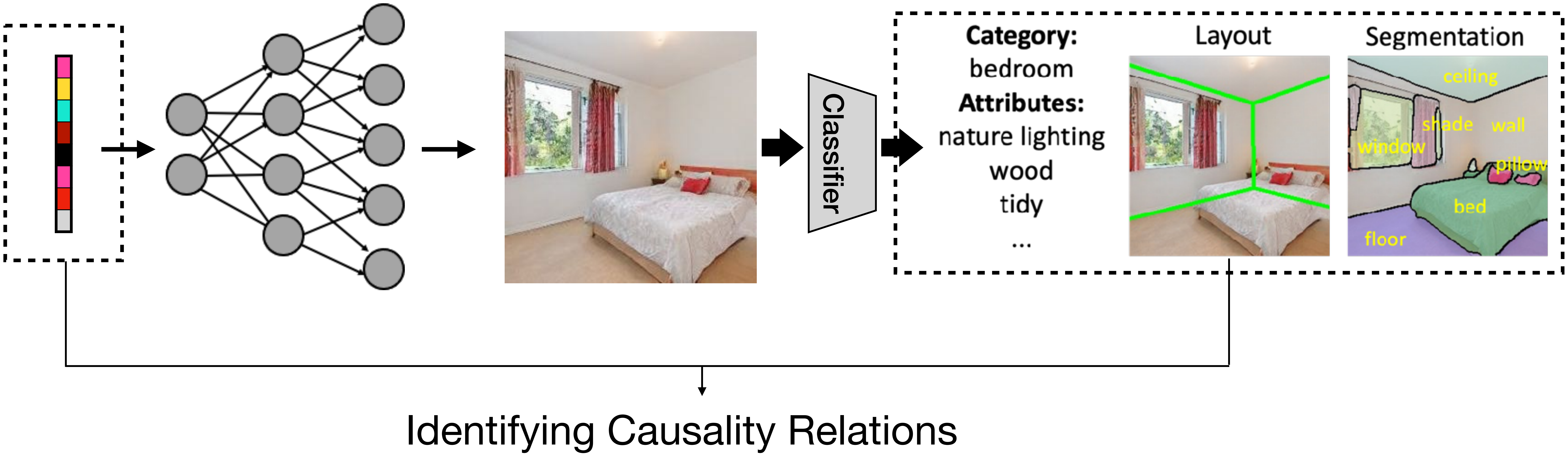
Bau, Zhu, Strobelt, Zhou, Tenenbaum, Freeman, Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. ICLR'19

Representations of GAN’s Generator



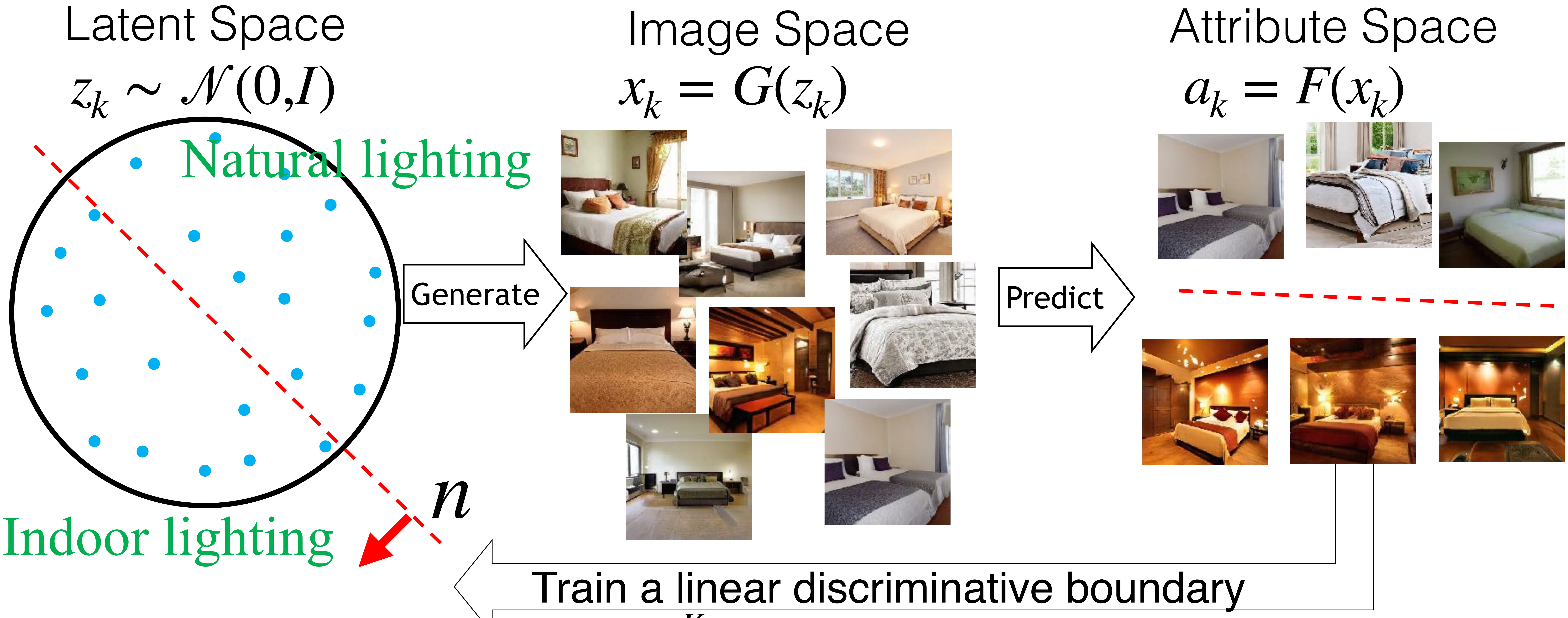
Supervised Approach

Probing latent space with linear classifier



Yang, Shen, Zhou. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis. IJCV, 2020

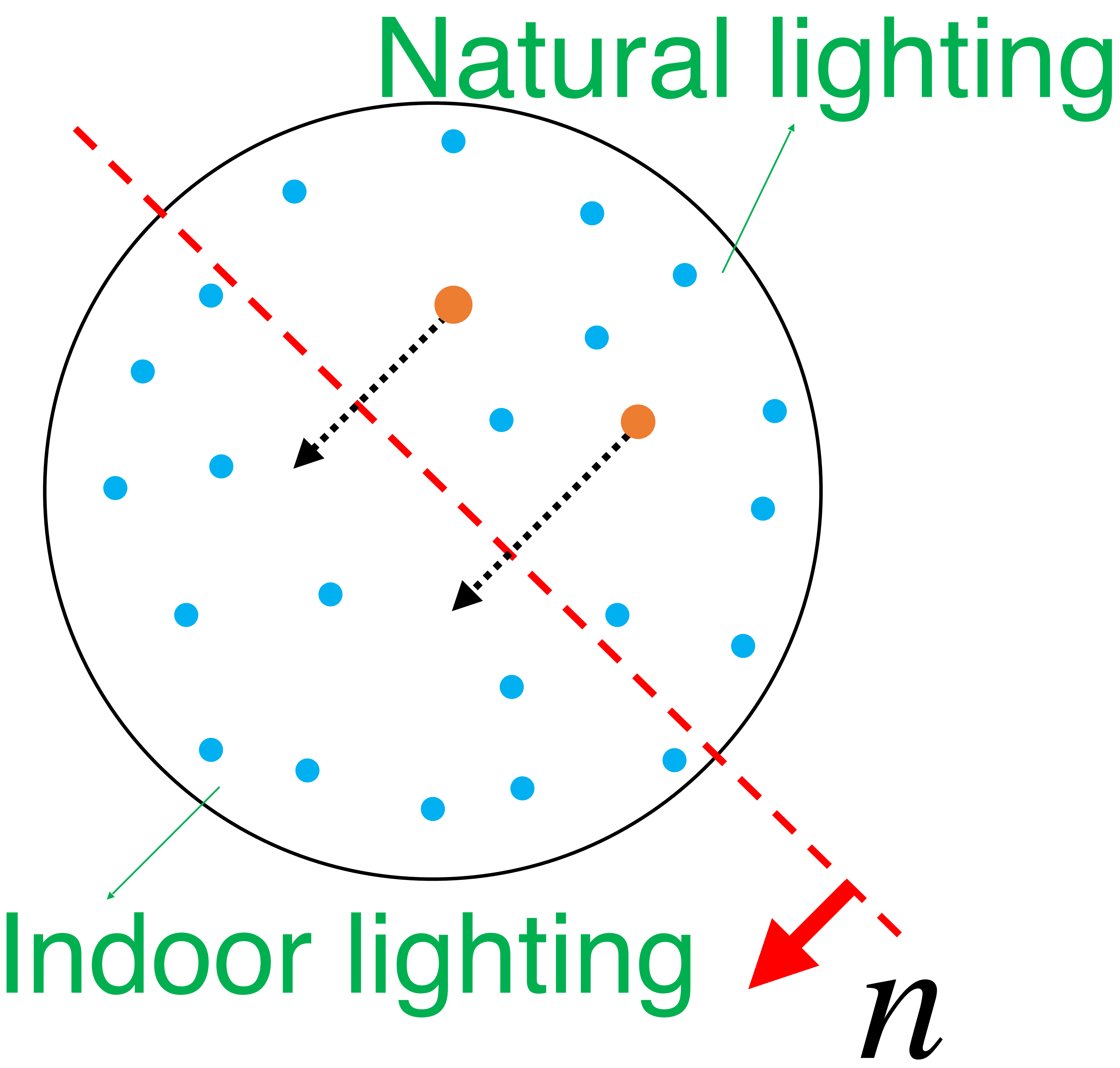
Identifying Causality Relations



Counterfactual Verification: $$\Delta a = \frac{1}{K} \sum_{k=1}^{K} \max(F(G(z_k + n)) - F(G(z_k)), 0)$$

Yang, Shen, Zhou. IJCV, 2020

Linear Manipulation on Latent Code

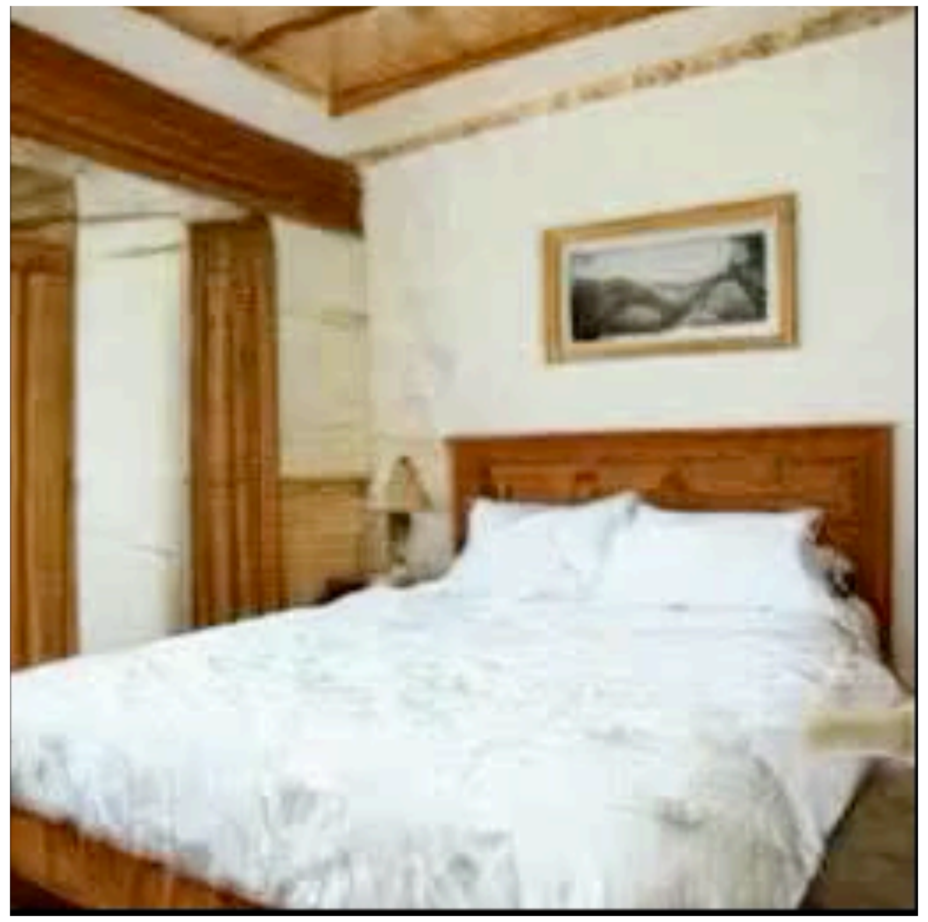


$$G(z_k) \dashrightarrow G(z_k + \lambda n)$$

Yang, Shen, Zhou. IJCV, 2020

Steering Generative Model

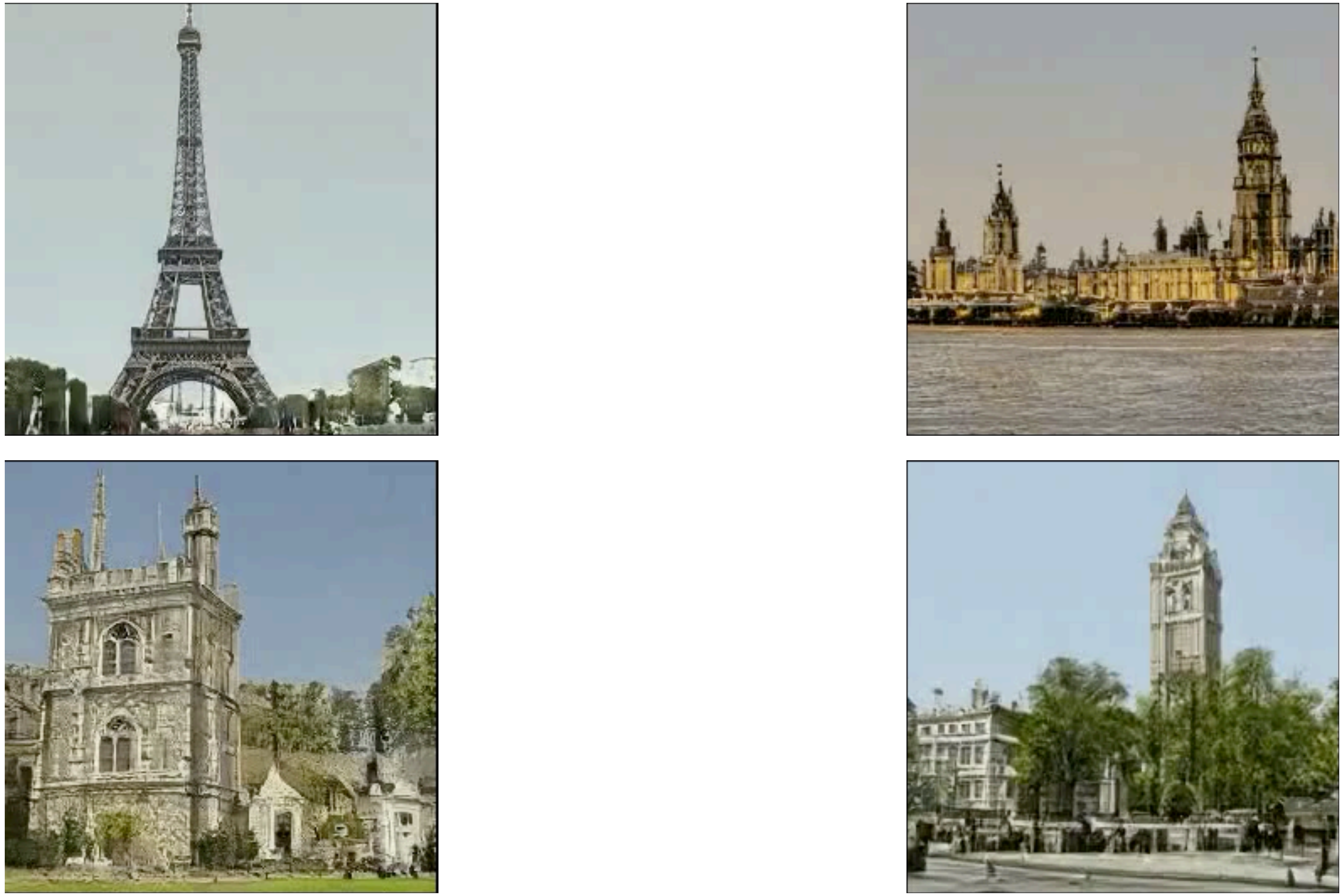
Changing Indoor lighting



Yang, Shen, Zhou. IJCV, 2020

Steering Generative Model

Adding clouds



Yang, Shen, Zhou. IJCV, 2020

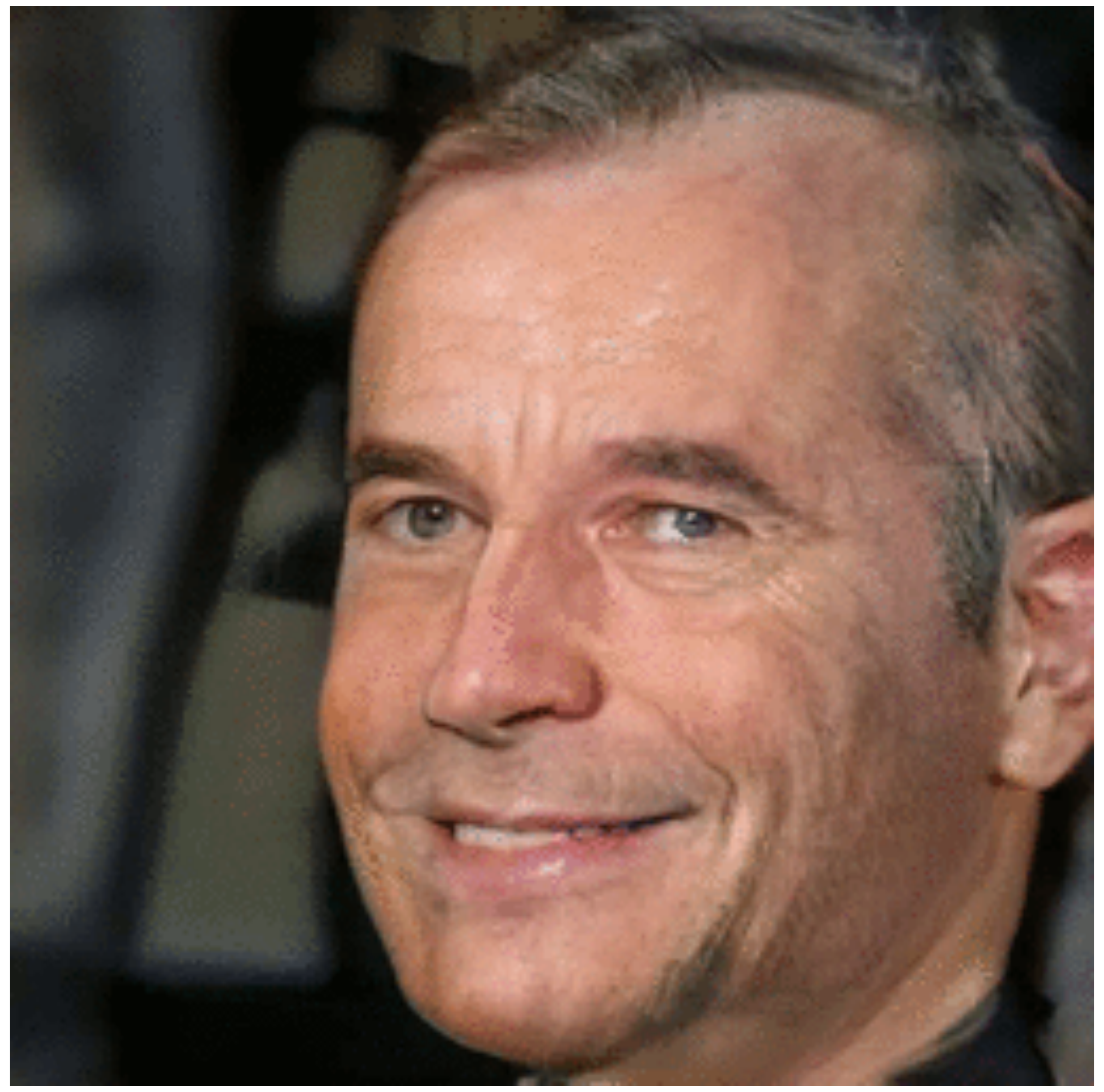
Supervised Approach

InterFaceGAN: Probing latent space of face GAN with linear classifier

Age

Gender

Pose



Artifact

Shen, Yang, Tang, Zhou. InterFaceGAN: interpreting the disentangled face representation learned by GANs. CVPR'20 and IEEE TPAMI

Supervised Approach

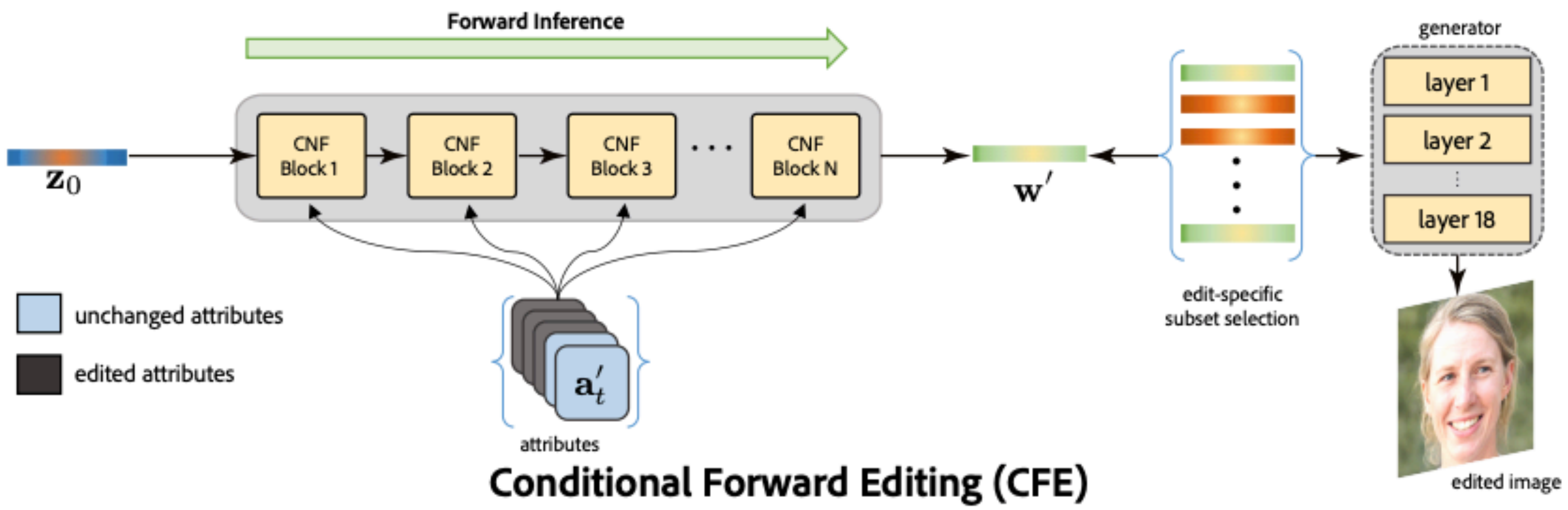
StyleFlow: StyleGAN + flow-based conditional model

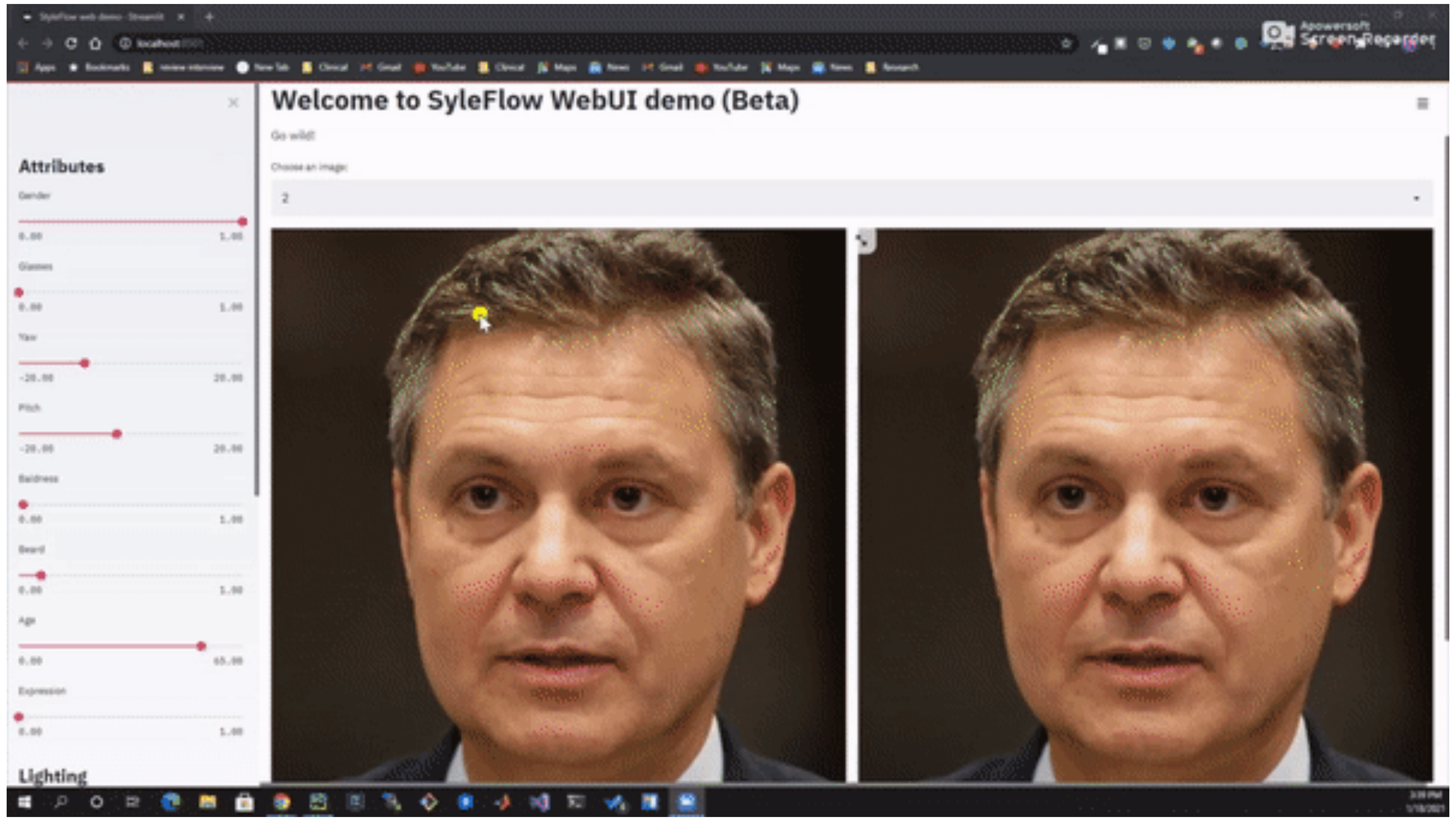
- Previous work assumes the linear manipulation model:

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

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

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





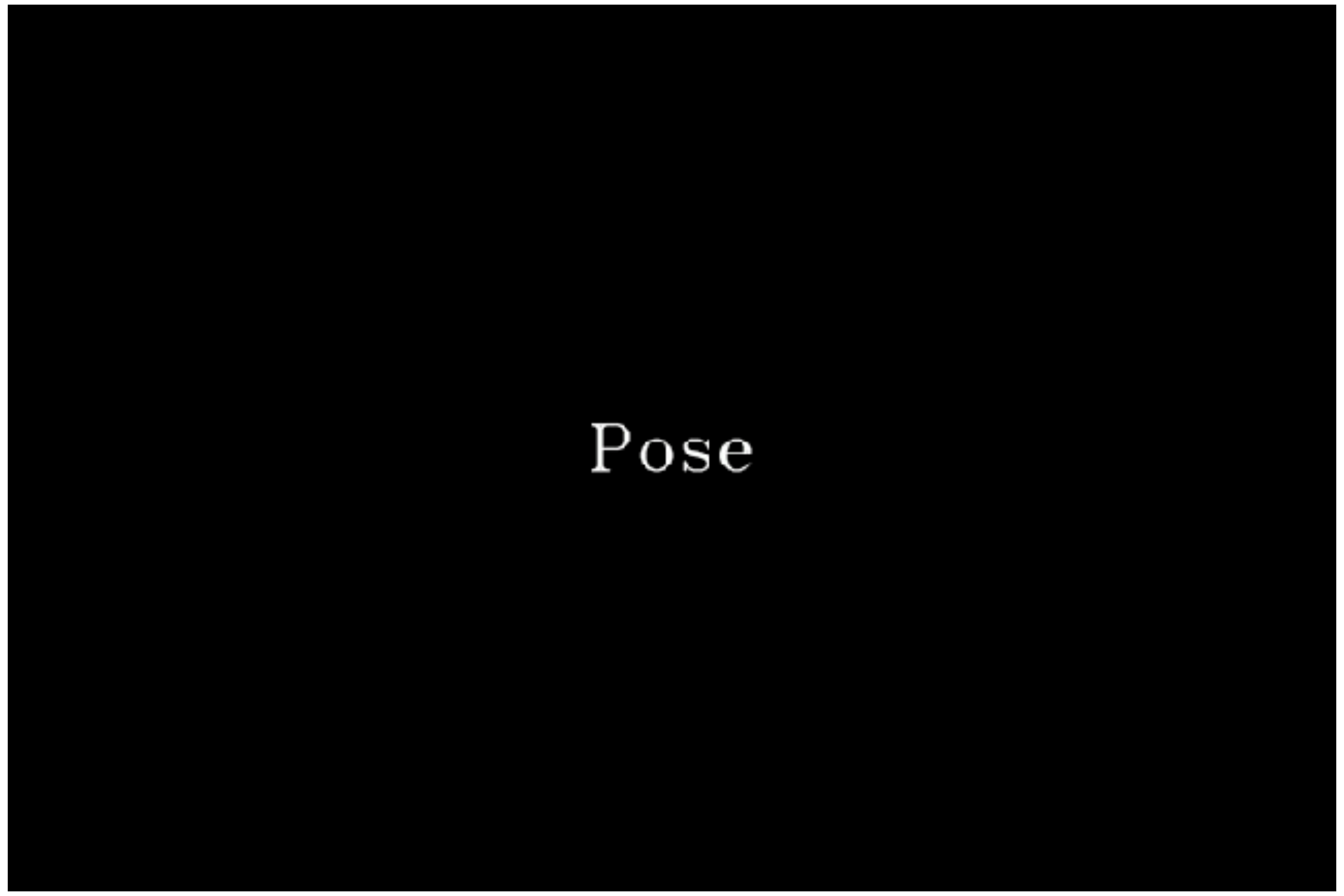
Abdal et al. StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. SIGGRAPH'21

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)

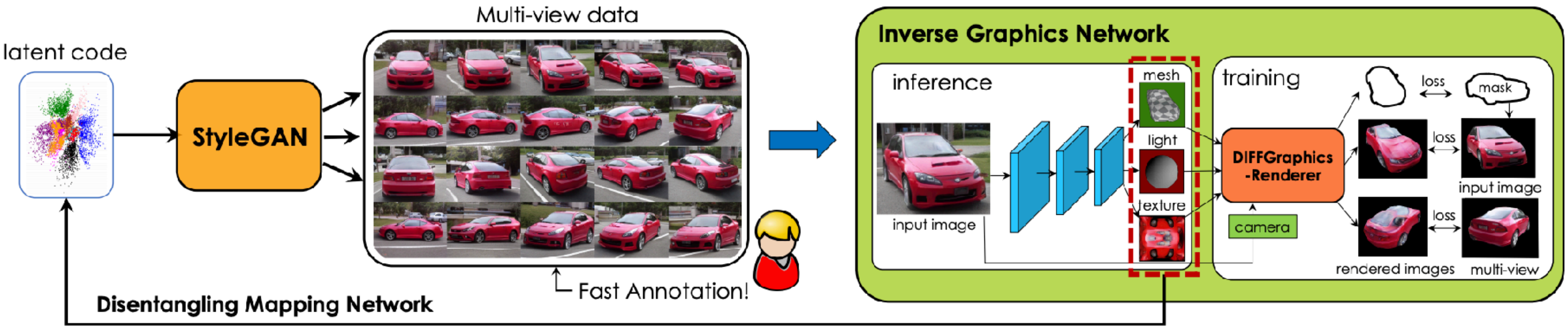


Supervised Approach

Parsing 3D Information from 2D Image Generator

Differentiable rendering for inverse graphics and interpretable 3D rendering

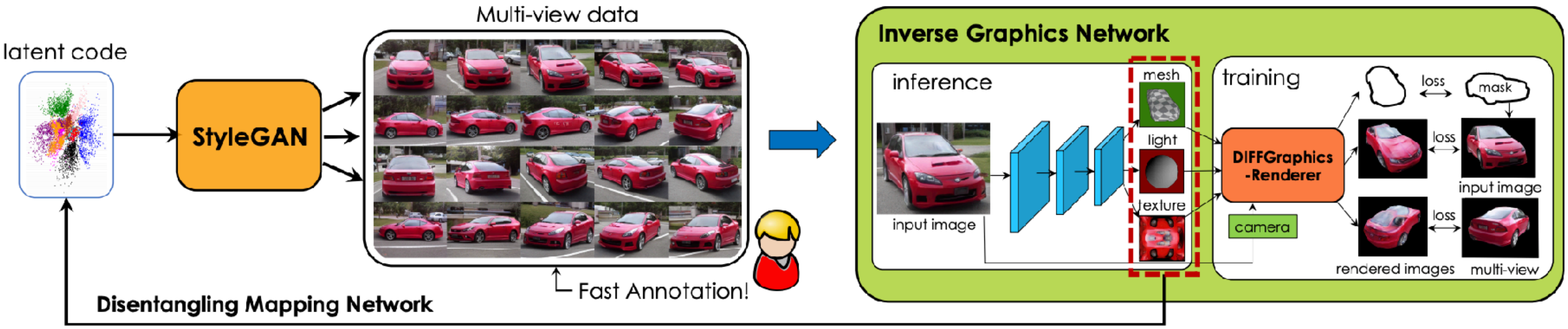
1. Use multi-view synthetic data to train inverse graphics network
2. Use trained inverse graphics net to train mapping network

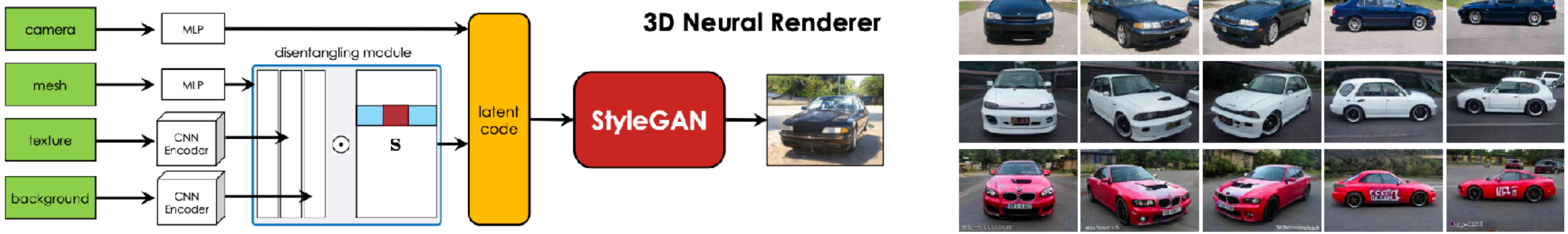


Zhang et al. Image GANs meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. ICLR'21

Supervised Approach

Parsing 3D Information from 2D Image Generator





Controllable output

Zhang et al. Image GANs meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. ICLR'21

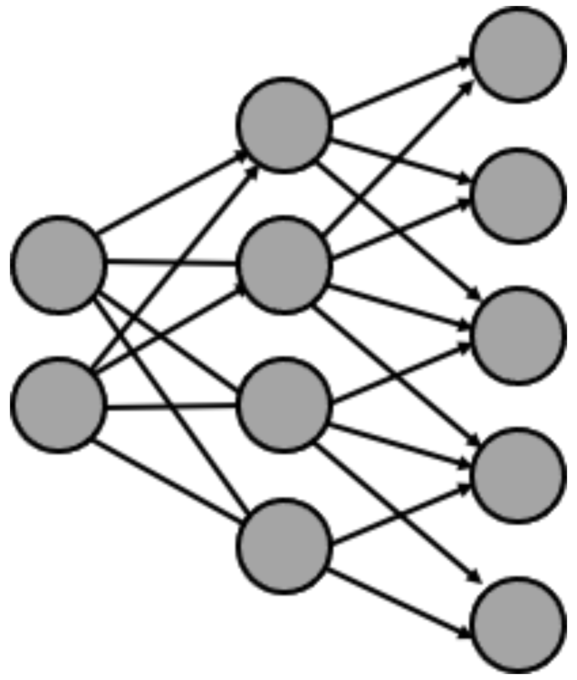
Challenges for Supervised Approach

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

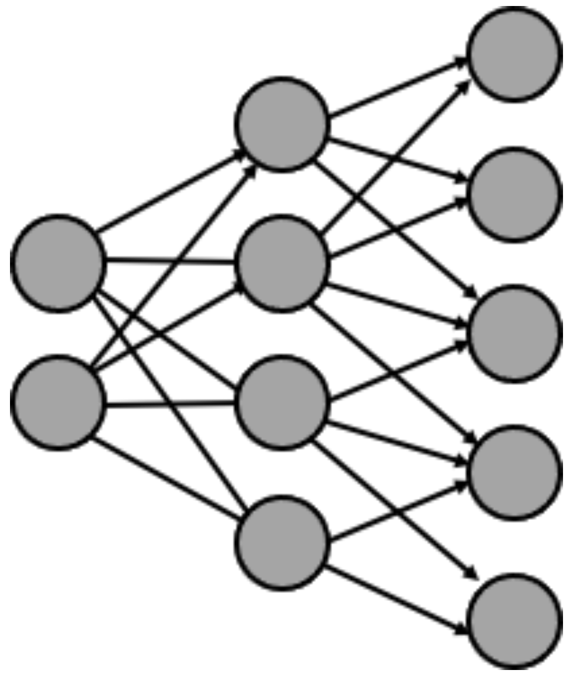
Interpretation Approaches

- **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



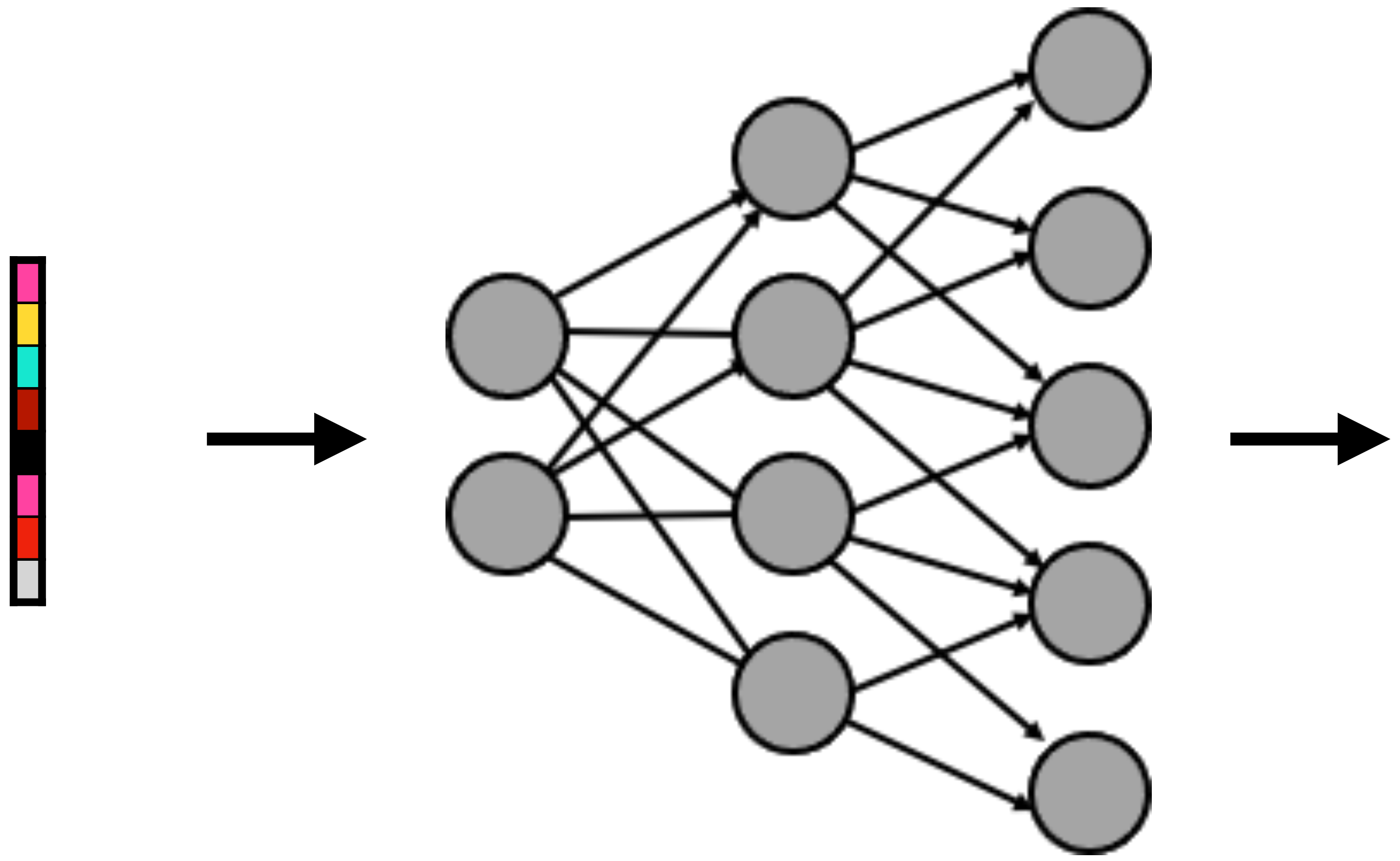
Generative model
for cats



Generative model
for cartoons

Unsupervised Approach

SeFa: Closed-form factorization of latent space in GANs



Shen and Zhou, CVPR'21

Unsupervised Approach

SeFa: Closed-form factorization of latent space in GANs

Intermediate activation:

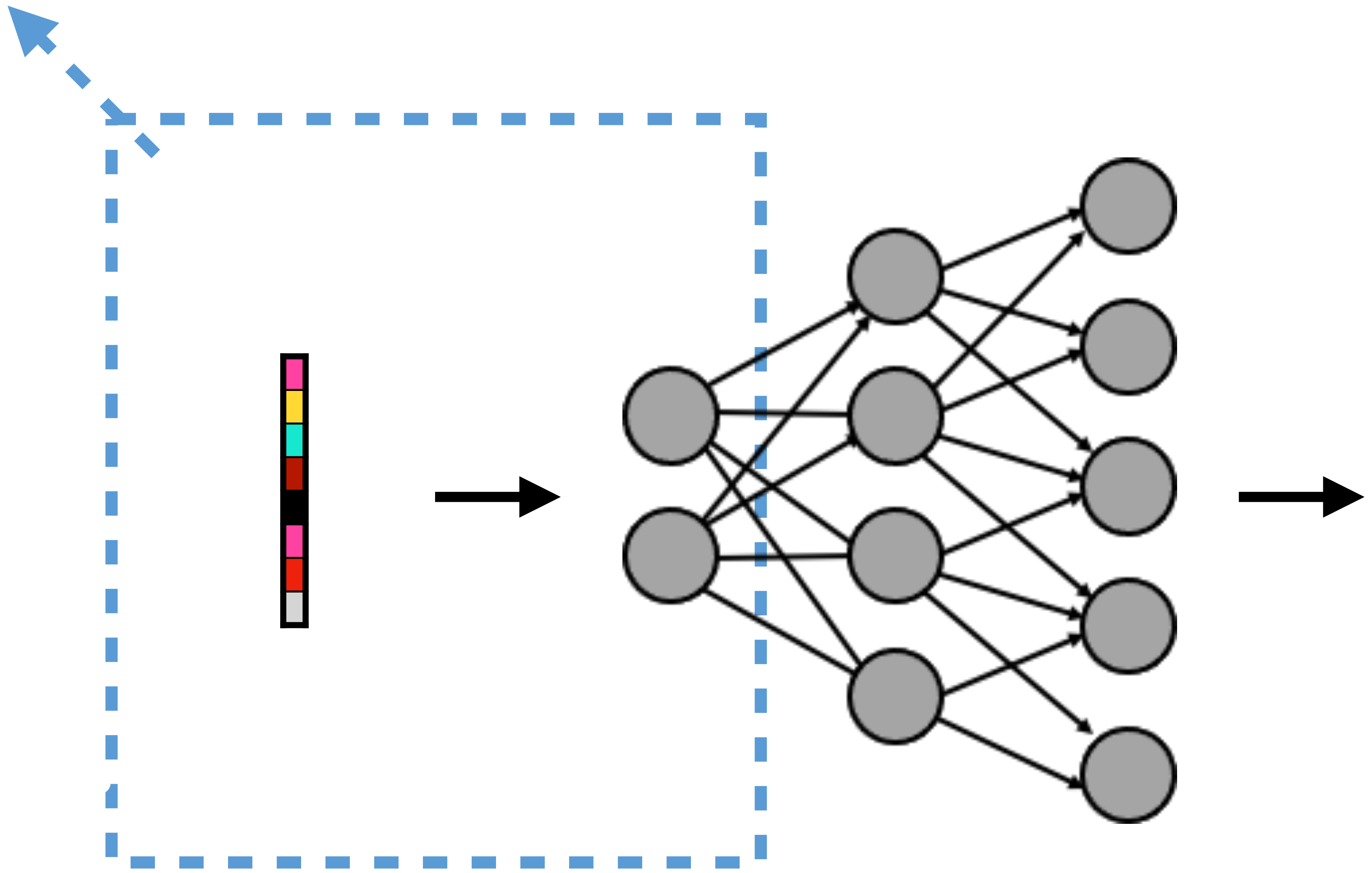
$$G_1(z) \triangleq y = Az + b$$

Feature difference after editing:

$$\Delta y = G_1(z + \lambda n) - G_1(z) = \lambda A n$$

Objective: to maximize variation of the difference

$$n^* = \mathrm{argmax}_{\{n \in R^d:\ n^T n = 1\}} ||An||_2^2$$



Shen and Zhou, CVPR'21

Unsupervised Approach
SeFa: Closed-form factorization of latent space in GANs

Human-in-the-loop
AI content creation

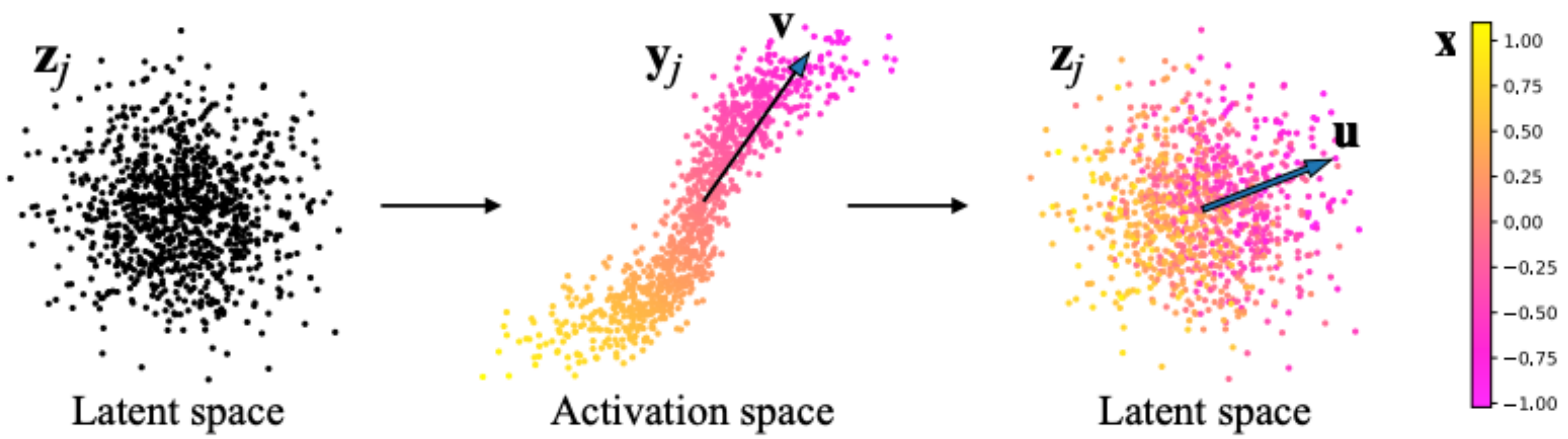
https://genforce.github.io/sefa

Shen and Zhou, CVPR'21

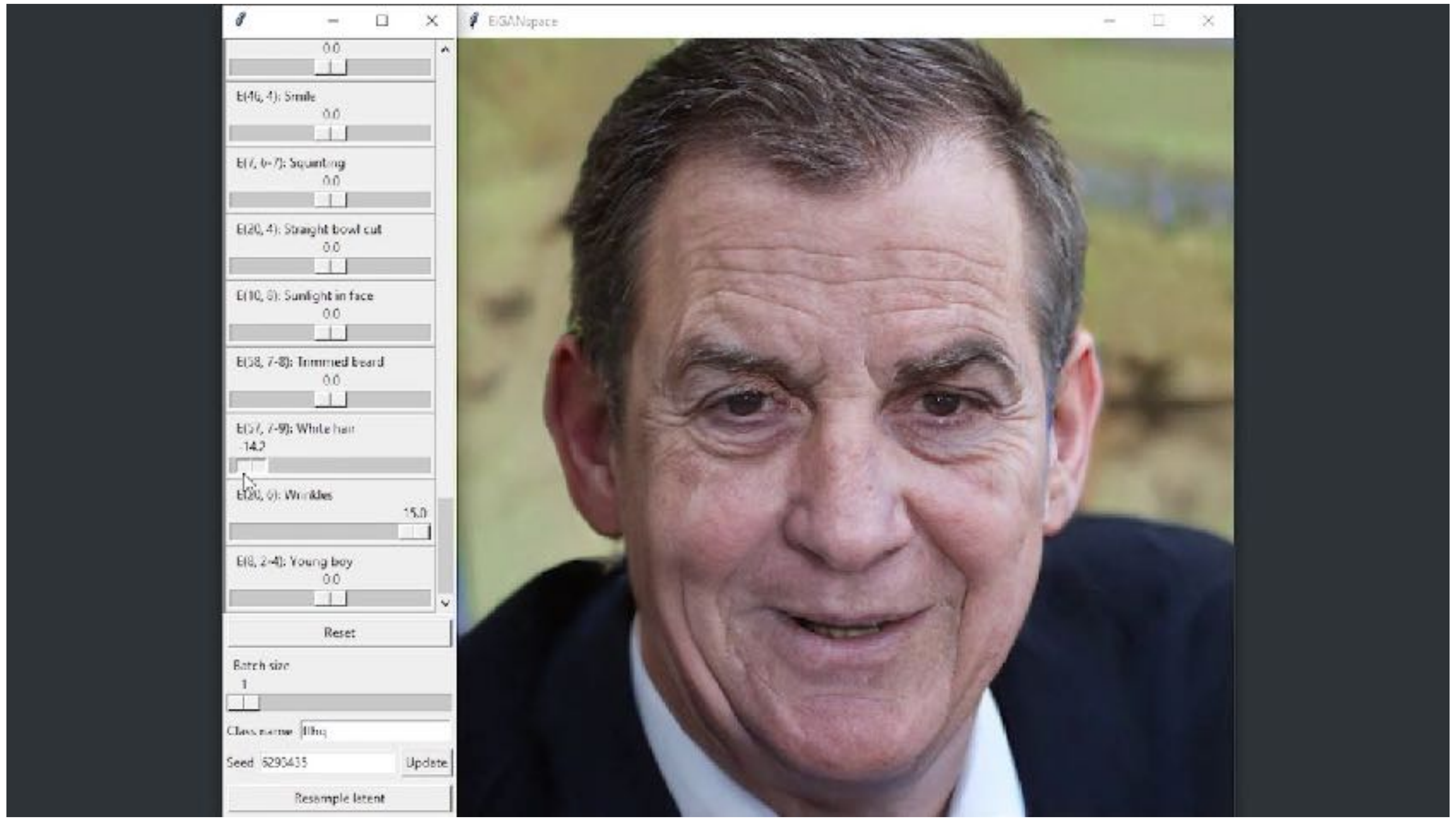
Unsupervised Approach

GANspace: PCA applied to the latent space of StyleGAN

PCA direction
in feature space



Regression from PCA
direction in latent space



Erik Harkonen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: discovering interpretable GAN controls. NeurIPS'20

Unsupervised Approach

Hessian Penalty: A weak prior for unsupervised disentanglement.

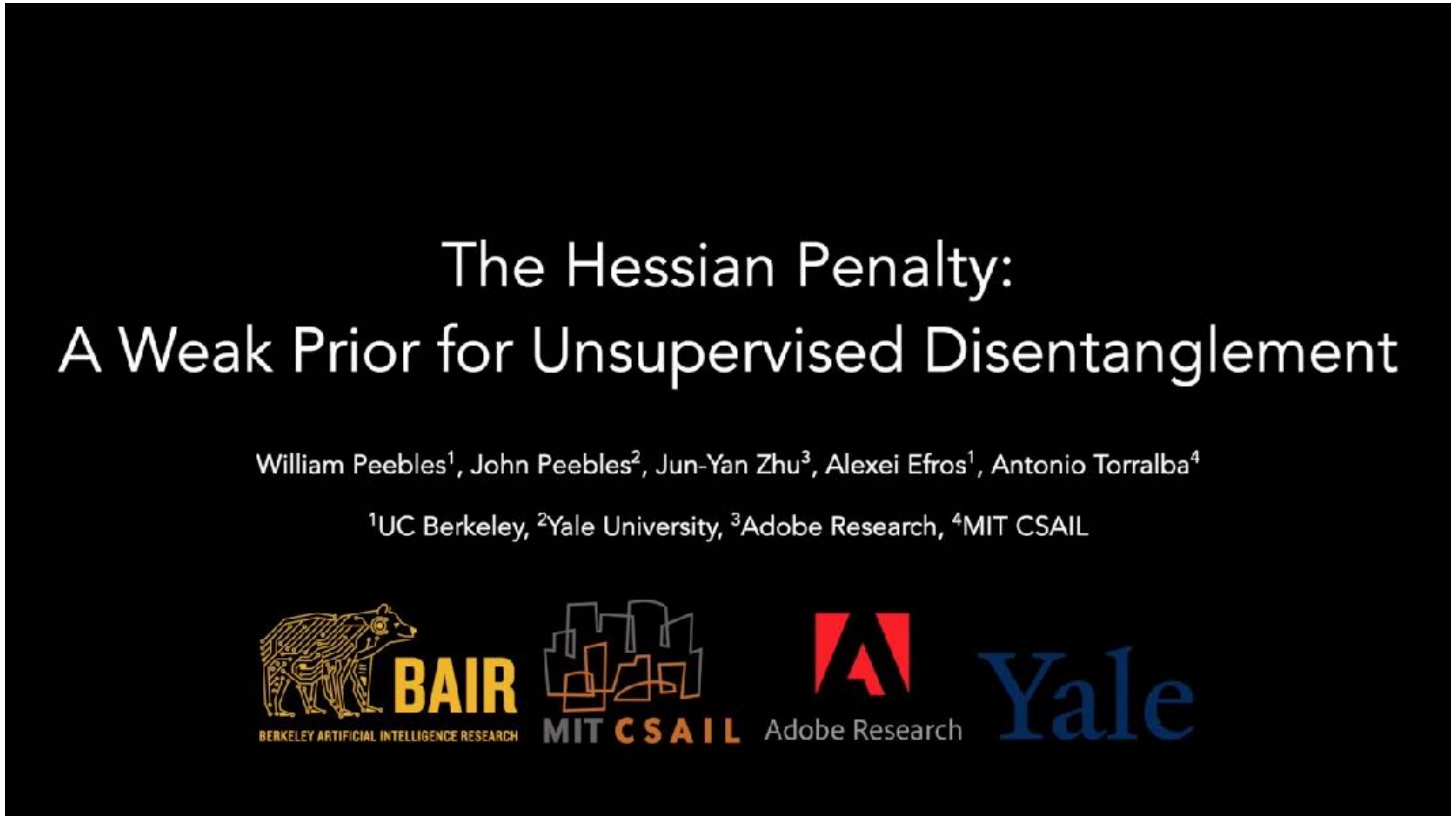
$$I = G(z)$$

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:

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

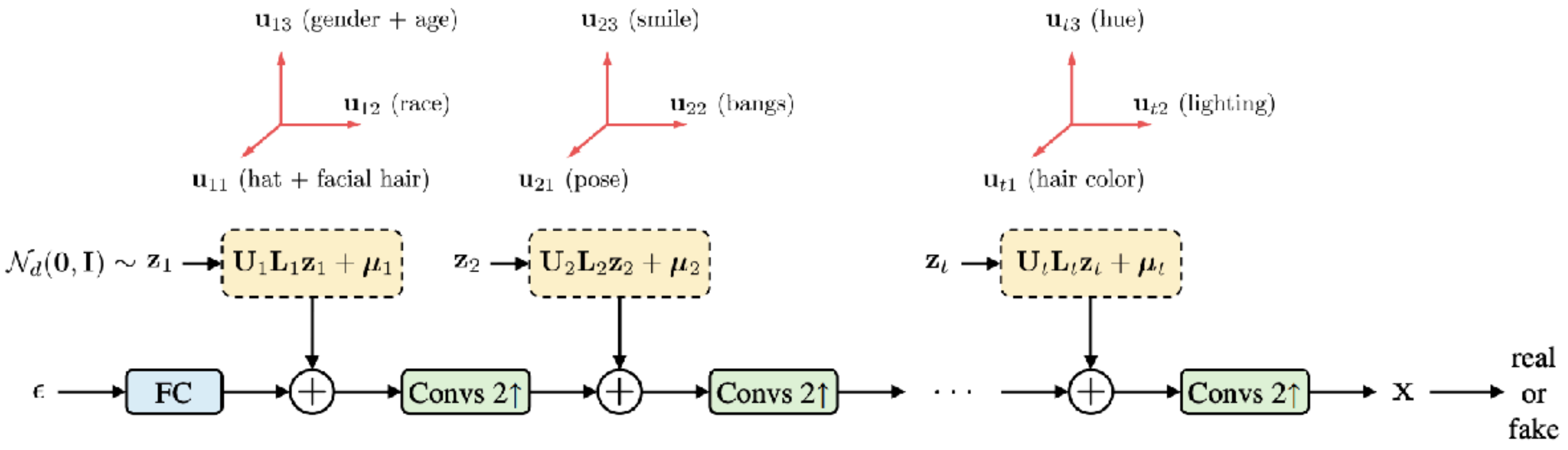


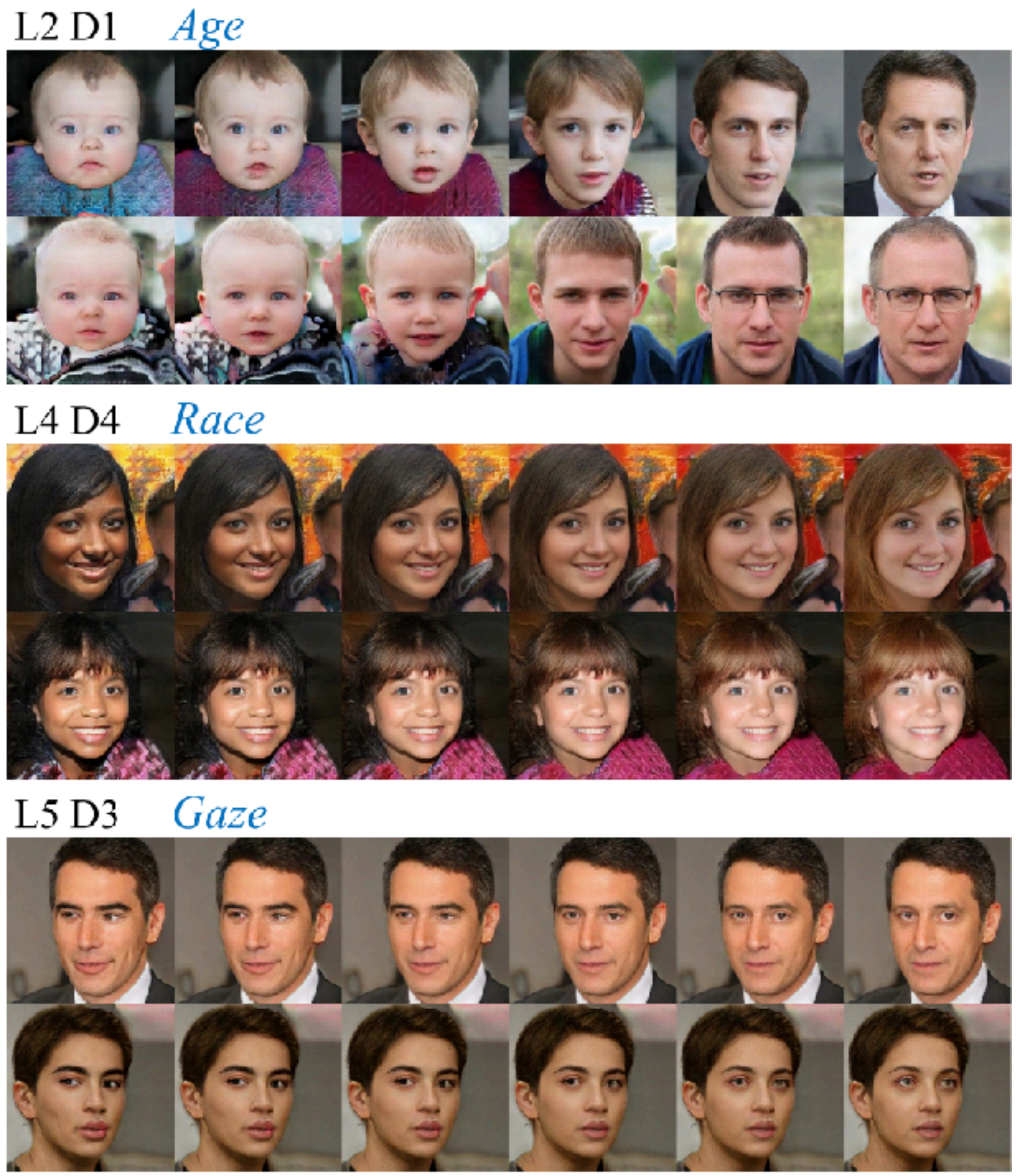
William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. ECCV'20

Unsupervised Approach

EigenGAN: Layer-Wise Eigen-Learning for GANs

Design inductive bias of disentanglement in the generator:





Zhenliang He, Meina Kan, Shiguang Shan. https://arxiv.org/pdf/2104.12476.pdf

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”

Patashnik, Wu, Shechtman, Cohen-Or, Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. https://arxiv.org/pdf/2103.17249.pdf

Zero-Shot Approach

StyleCLIP: CLIP + StyleGAN

Source:

Output:

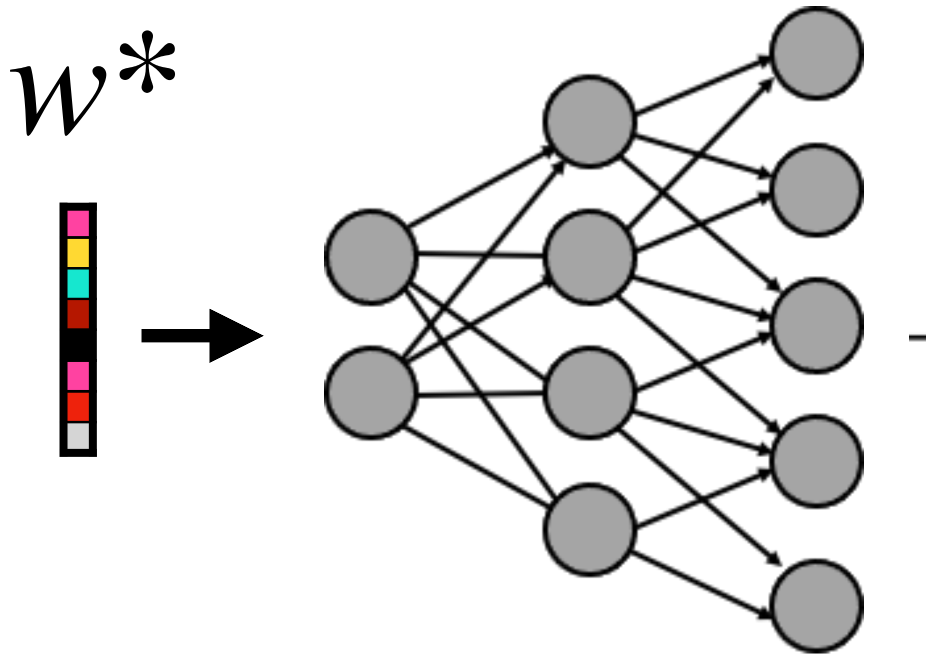
Text input: "Mohawk hairstyle" "Without makeup" "Cute cat" "Lion" "Gothic church"

Patashnik, Wu, Shechtman, Cohen-Or, Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. https://arxiv.org/pdf/2103.17249.pdf

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



$\mathcal{L}_{CLIP}(\text{"surprised"},\cdot)$

w_s

Source

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

CLIP's joint embedding space

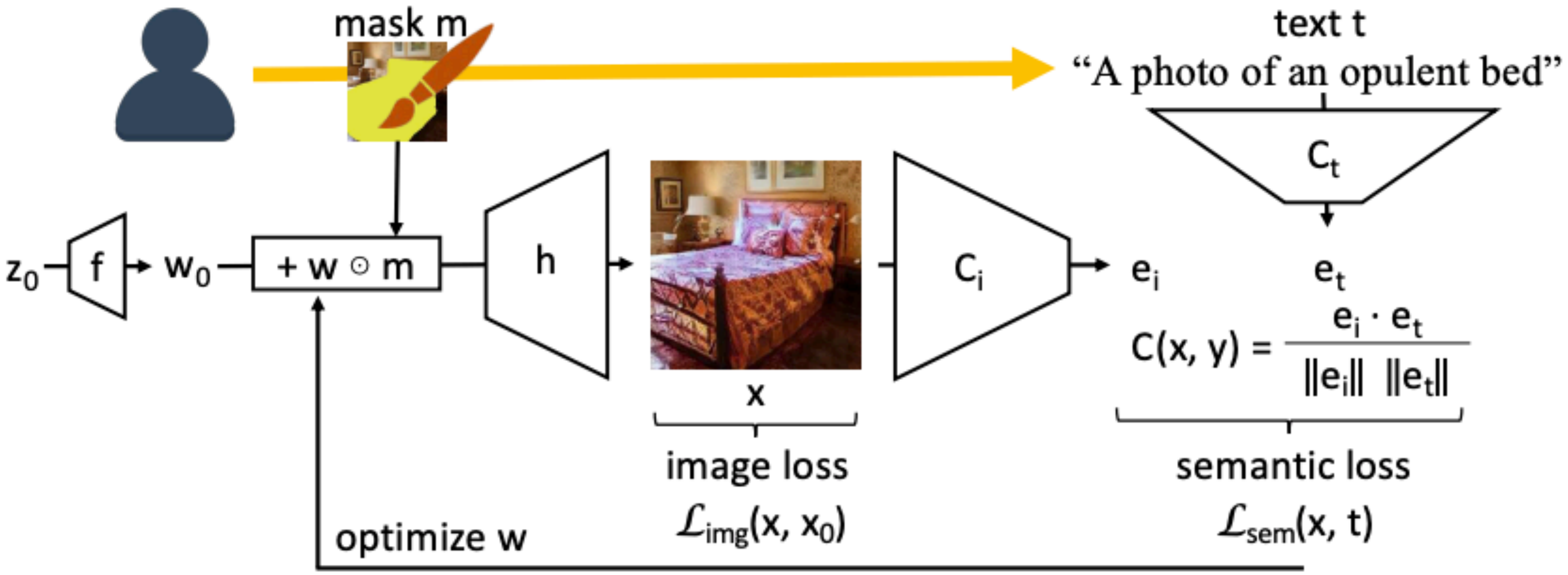
Regularizers on similarity and identity

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

Patashnik, Wu, Shechtman, Cohen-Or, Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. https://arxiv.org/pdf/2103.17249.pdf

Zero-Shot Approach

Paint by Word: CLIP + Region-based StyleGAN inversion



Bau, Andonian, Cui, Park, Jahanian, Oliva, Torralba. https://arxiv.org/pdf/2103.10951.pdf

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. a tapir with the texture of an 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

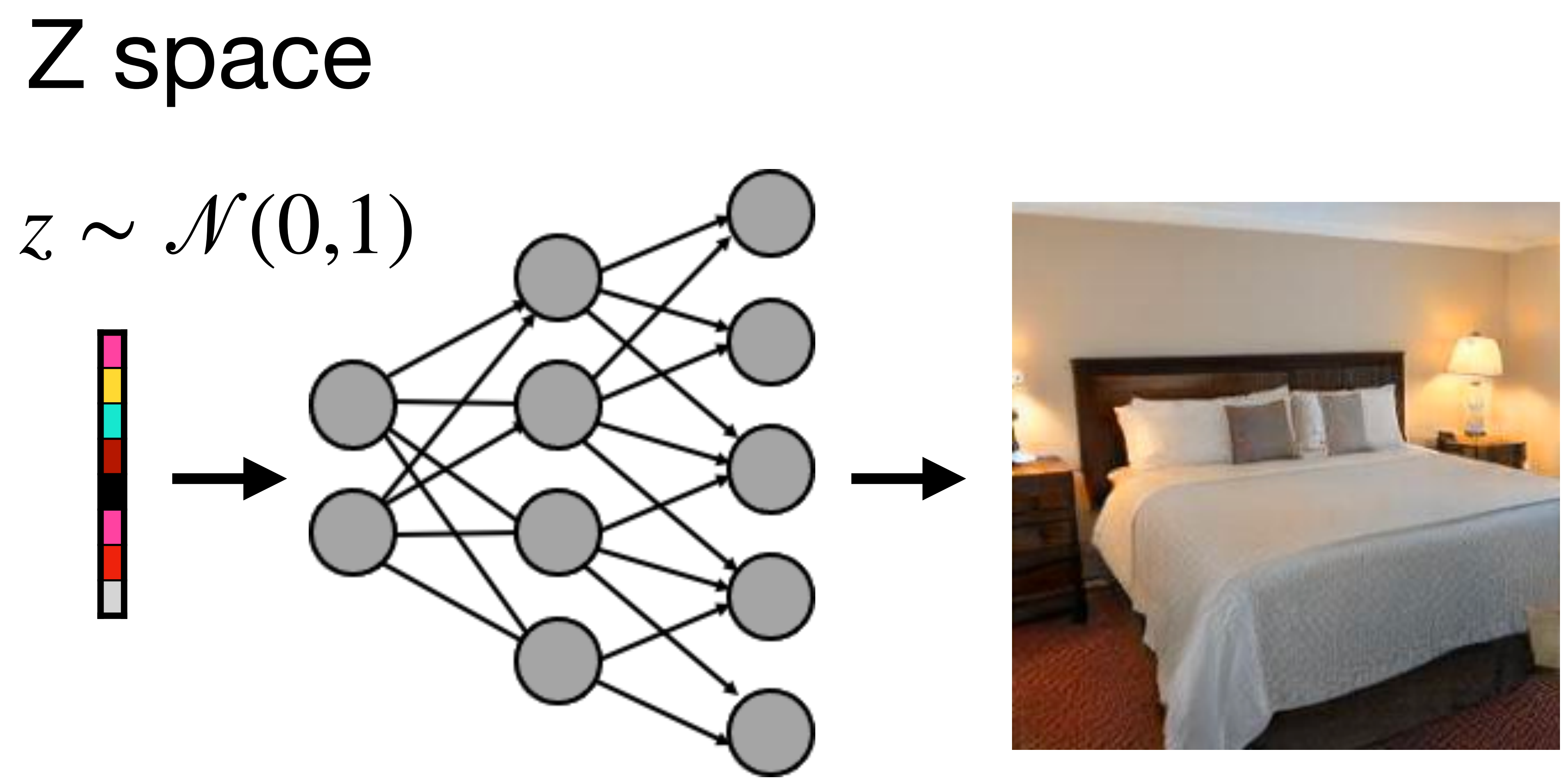
Ramesh et al. Zero-Shot Text-to-Image Generation. ICML'21

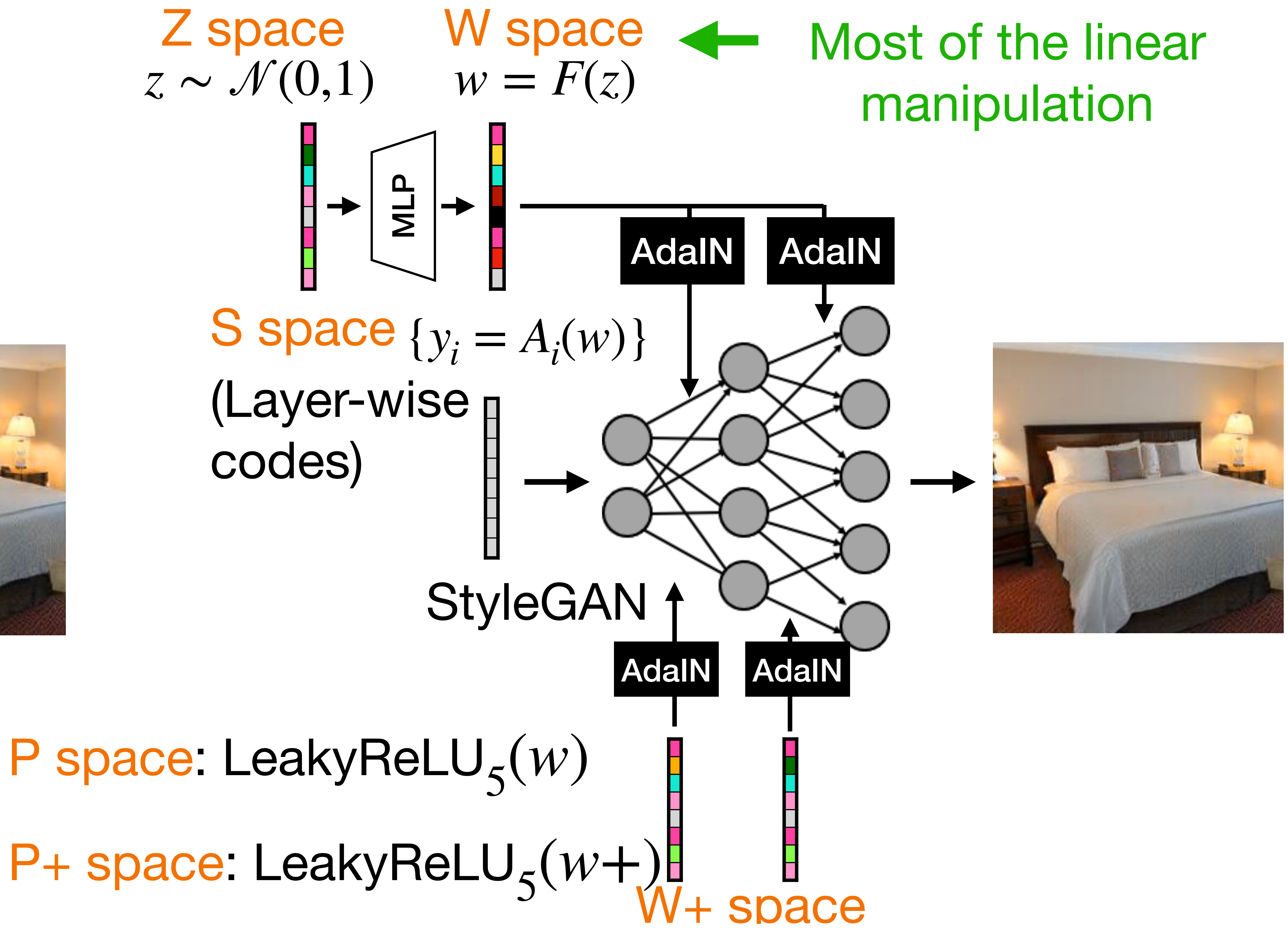
Interpretation Approaches

- **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

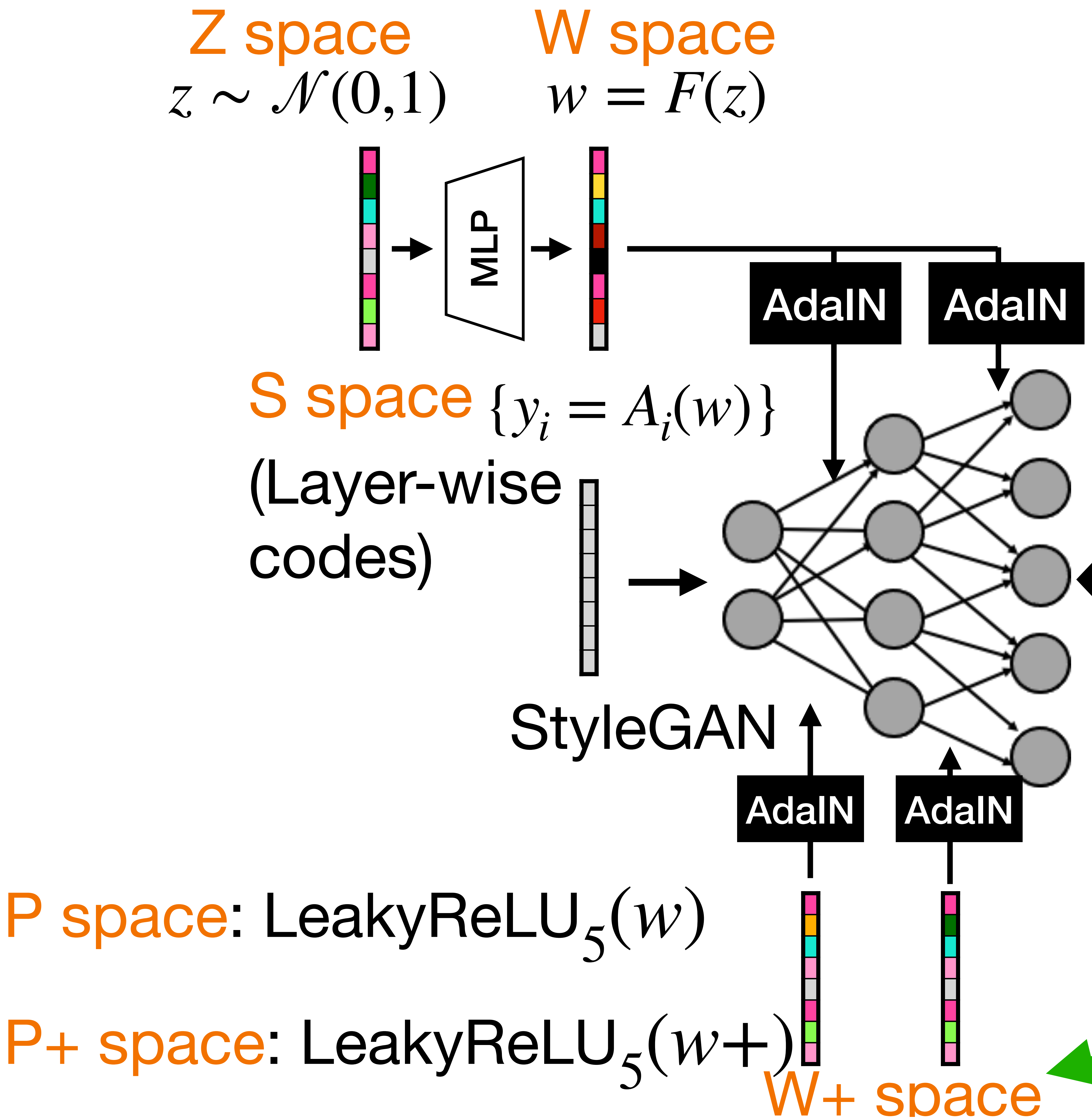




Latent Spaces of GAN’s Generator

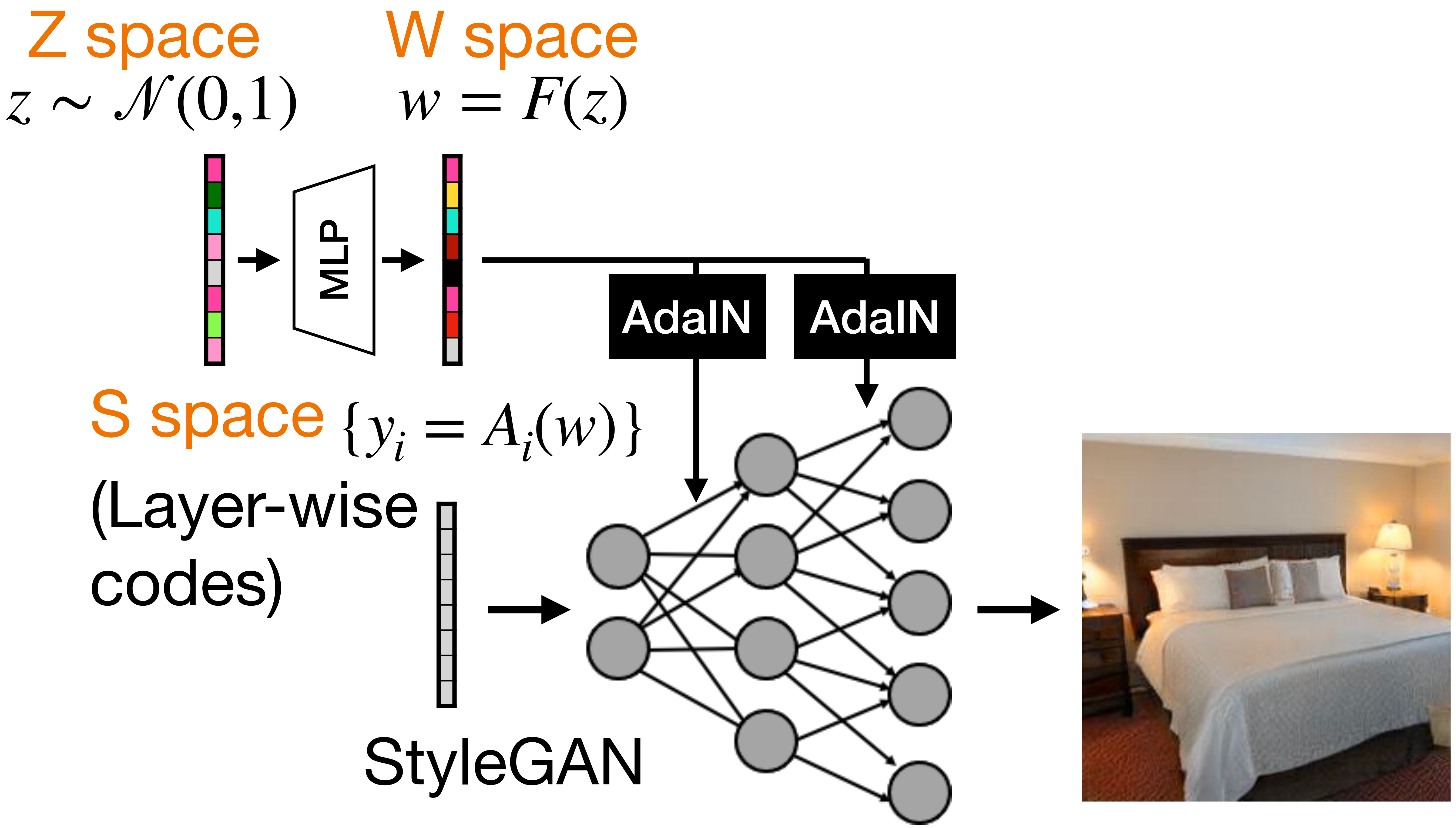
Z space, W space, StyleSpace (S space), W+ space, P/P+ space

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



Most of the GAN inversions happen in W+ 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**

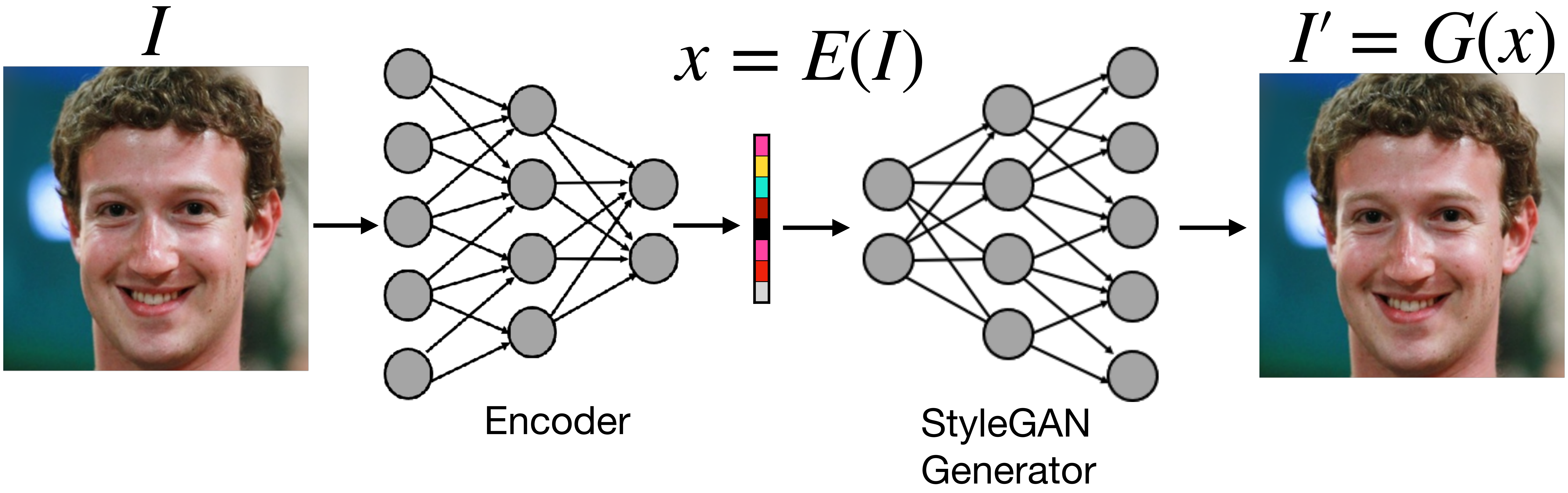
Disentanglement (Wu et al. CVPR'21)

Space	Disentanglement
Z space	0.31
W space	0.54
S space	**0.75**

Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, Bolei Zhou. Generative Hierarchical Features from Synthesizing Images. CVPR'21
Zongze Wu, Dani Lischinski, Eli Shechtman. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. CVPR'21

Encoding Real Image into StyleGAN space

Which latent space to use?



Encoding Real Image into StyleGAN space

Which latent space to use?

Input | W+ space (ALAE, CVPR'20) | W+ space (IDinvert, ECCV'20) | S space (GH-feat, CVPR'21)

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:

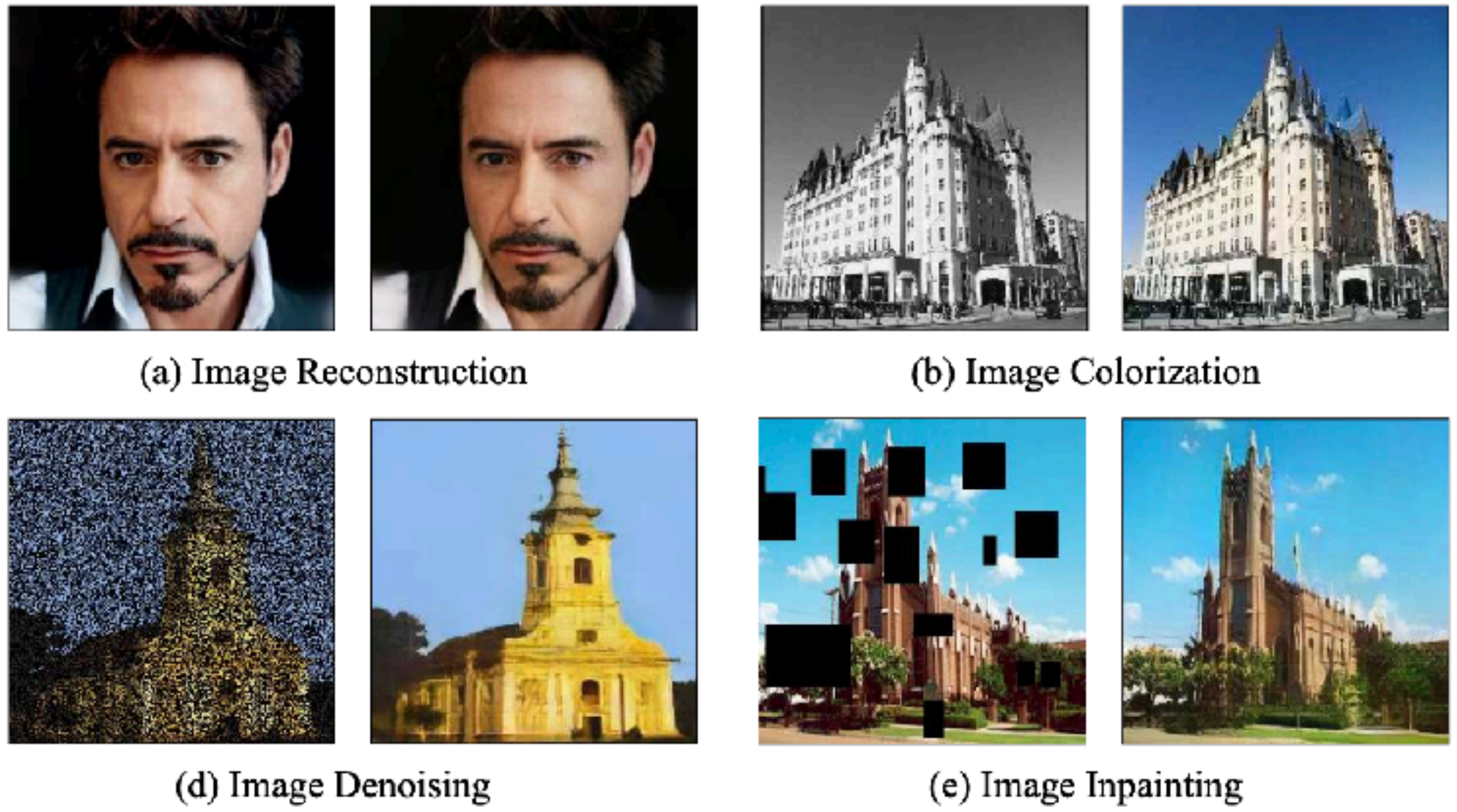
$$x^* = argmin_x ||G(x) - I||$$

Colorization:

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

Super-resolution:

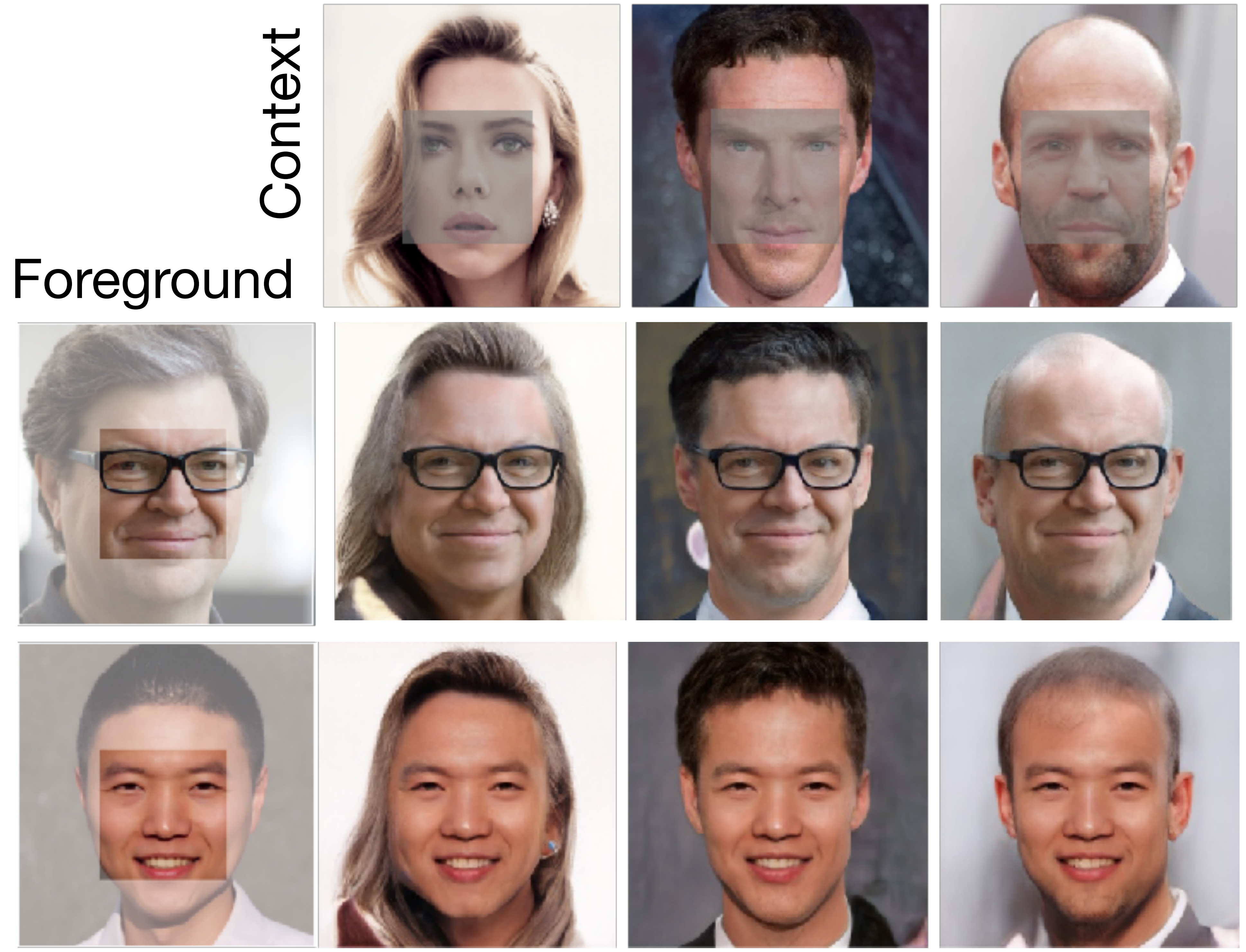
$$x^* = argmin_x ||\text{down}(G(x)) - I_{small}||$$



Gu, Shen, Zhou. Image Processing Using Multi-Code GAN Prior. CVPR'20

Masked optimization

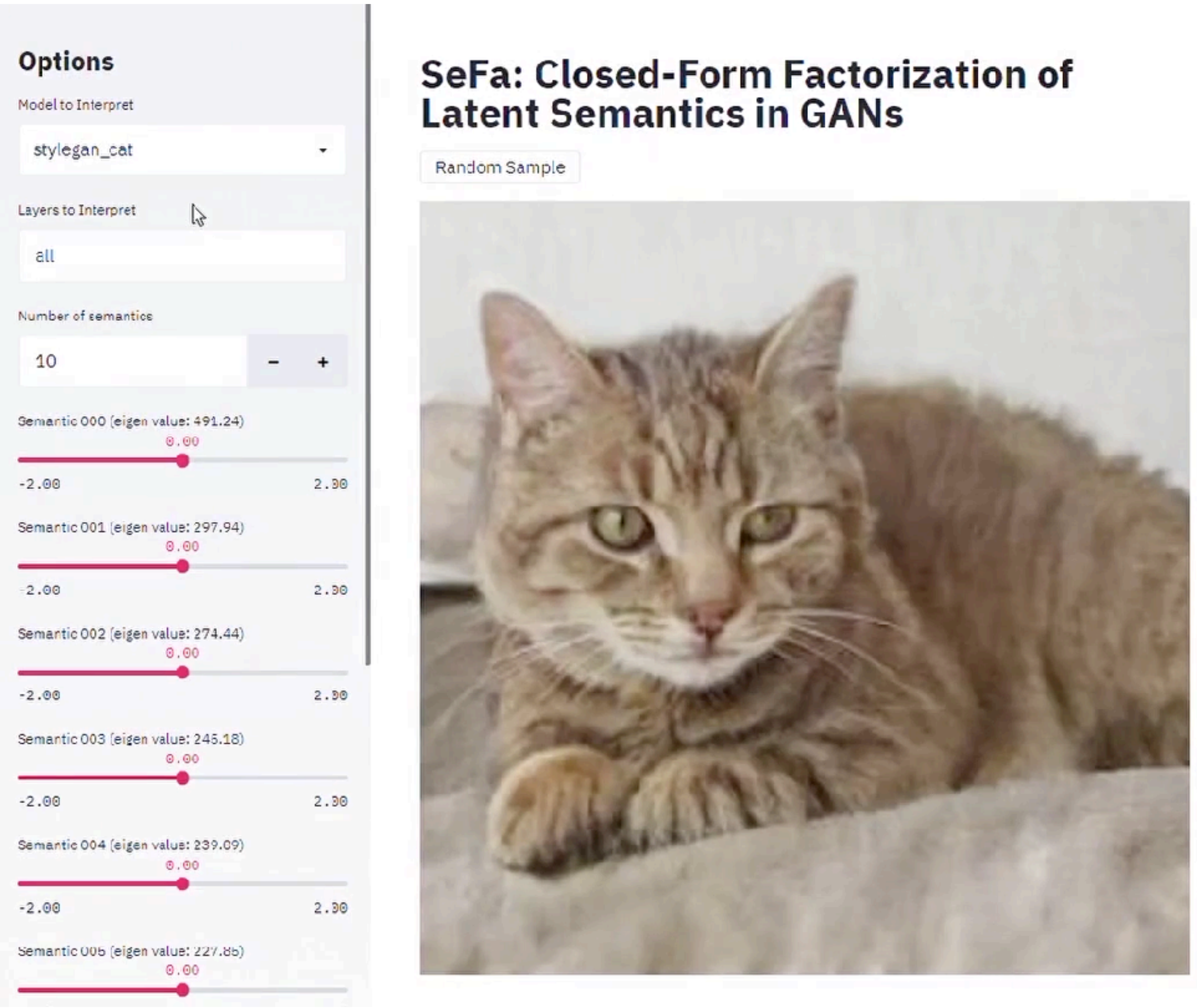
$$x^* = argmin_x ||m \cdot G(x) - m \cdot I_{context}||$$

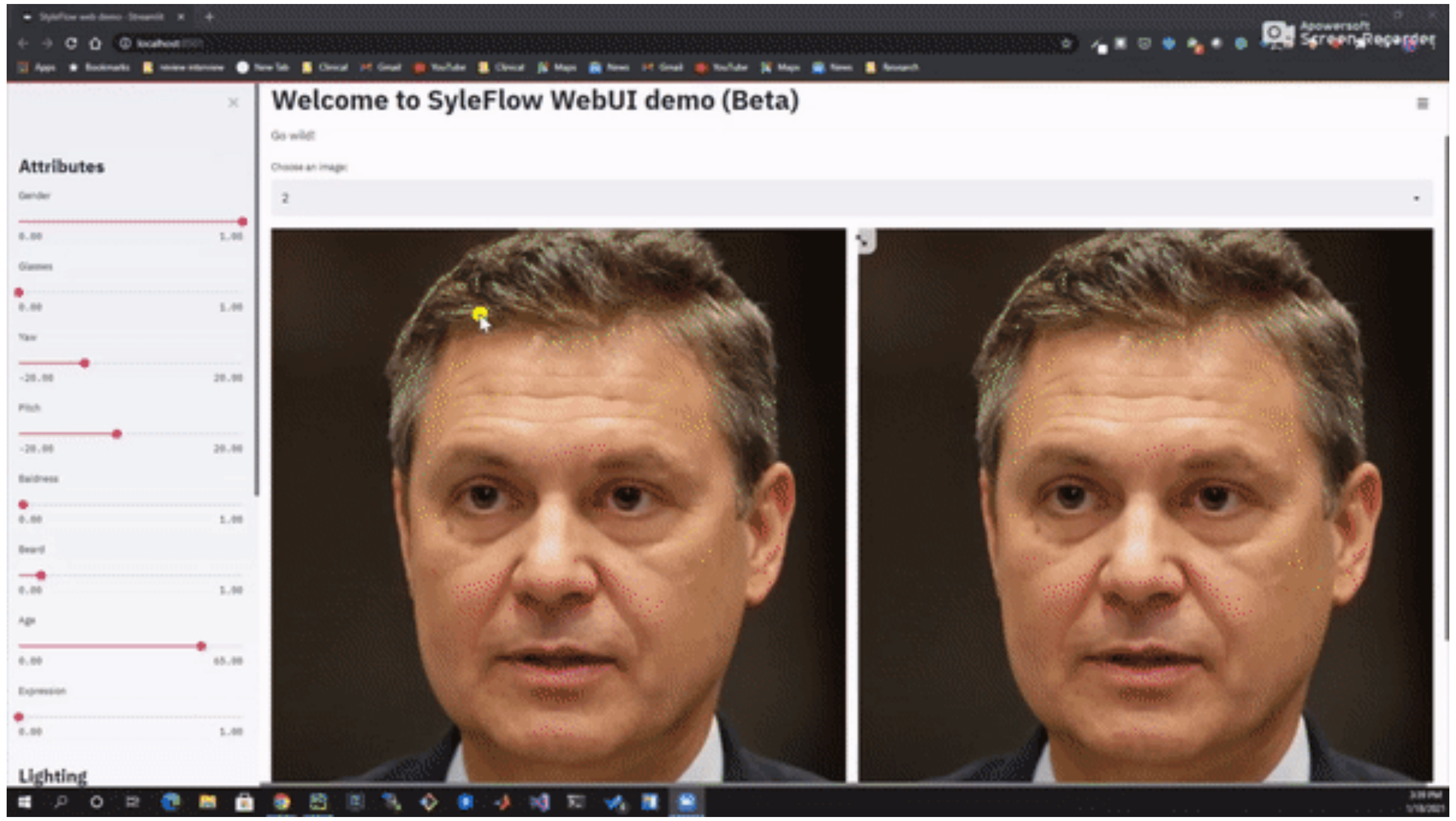


Zhu, Shen, Zhao, Zhou. In-domain GAN Inversion. ECCV'20

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