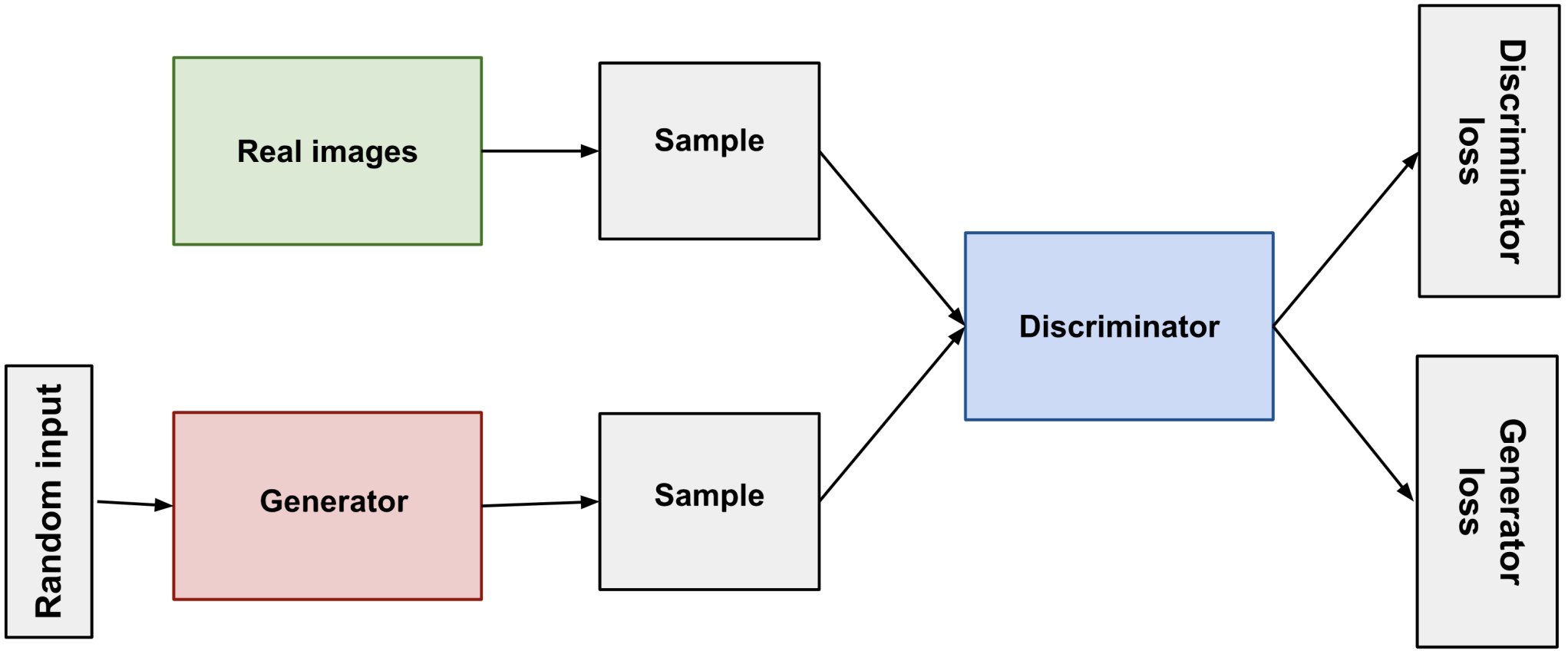
Intepretable GANs

Sergey Kim

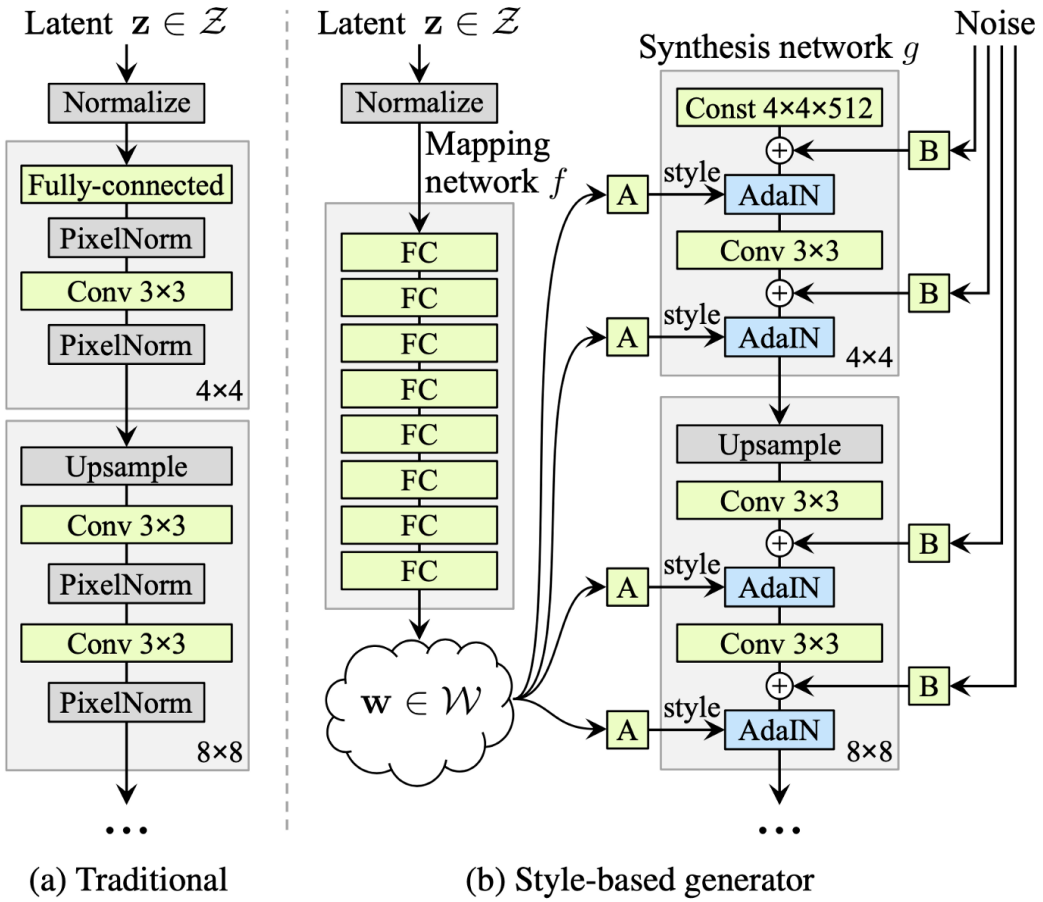
In the previous series

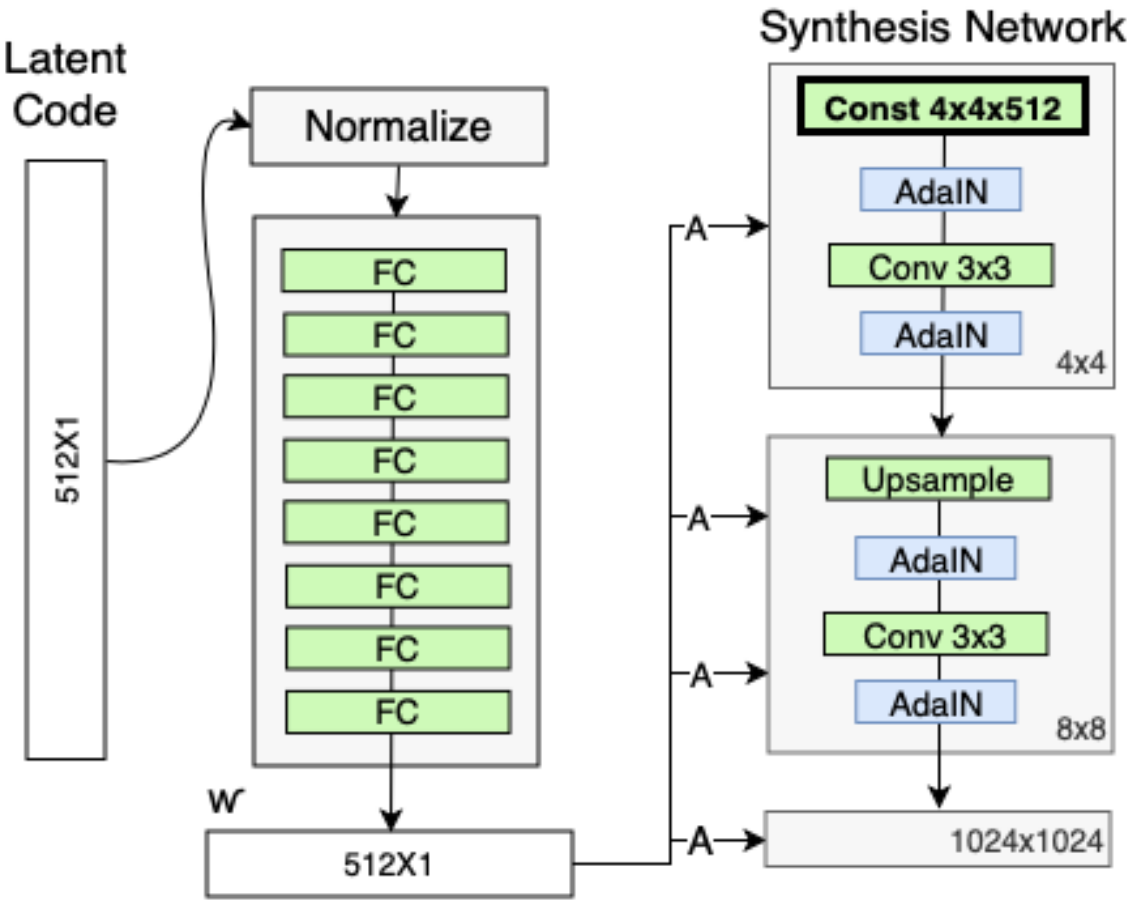
- GAN
- ProGAN
- StyleGAN
- StyleGAN2

GAN – Generative Adversarial Network



StyleGAN. Generator





Evolution

Oct 2017 – ProGAN (arxiv.org/abs/1710.10196) NVIDIA

Dec 2018 – StyleGAN (arxiv.org/abs/1812.04948) NVIDIA

Dec 2019 – StyleGAN2 (arxiv.org/abs/1912.04958) NVIDIA

Evolution

Oct 2017 – ProGAN (arxiv.org/abs/1710.10196)	NVIDIA
Dec 2018 – StyleGAN (arxiv.org/abs/1812.04948)	NVIDIA
Dec 2019 – StyleGAN2 (arxiv.org/abs/1912.04958)	NVIDIA
Jul 2019 – Steerability of GANs (arxiv.org/abs/1907.07171)	MIT
Jul 2019 – Interpreting GANs (arxiv.org/abs/1907.10786)	Hong Kong
Apr 2020 – GAN Inversion (arxiv.org/abs/2004.00049)	Hong Kong

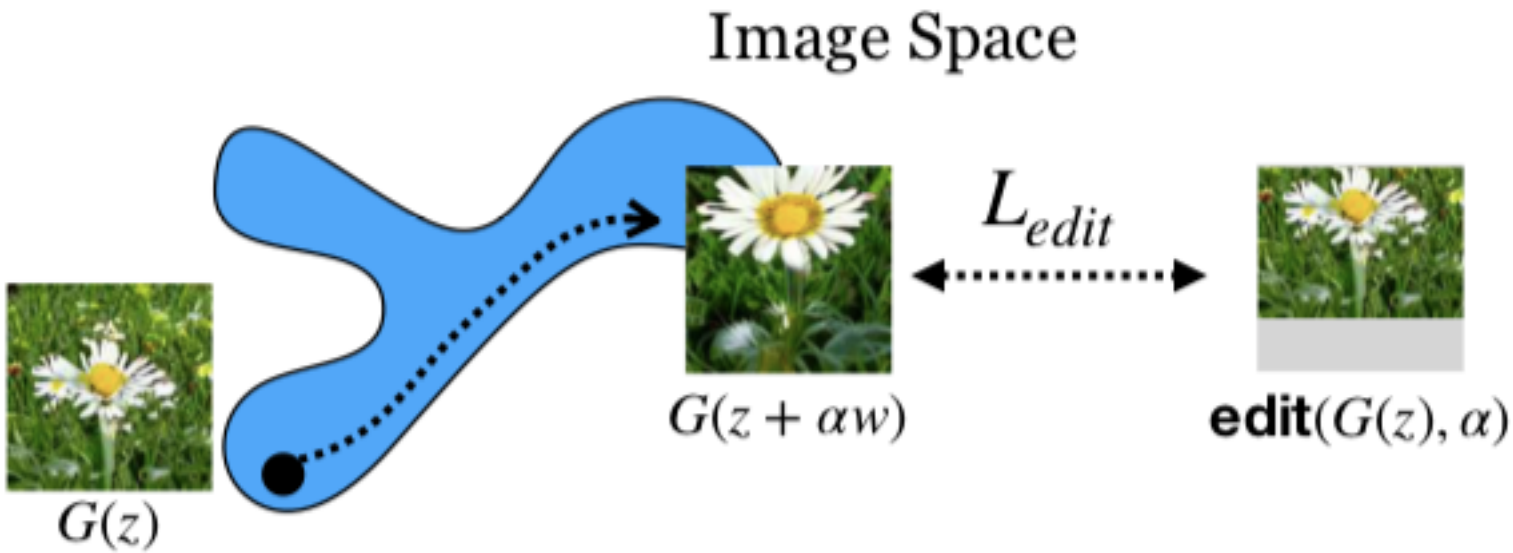
Evolution

Oct 2017 – ProGAN (arxiv.org/abs/1710.10196)	NVIDIA
Dec 2018 – StyleGAN (arxiv.org/abs/1812.04948)	NVIDIA
Dec 2019 – StyleGAN2 (arxiv.org/abs/1912.04958)	NVIDIA
Jul 2019 – Steerability of GANs (arxiv.org/abs/1907.07171)	MIT
Jul 2019 – Interpreting GANs (arxiv.org/abs/1907.10786)	Hong Kong
Apr 2020 – GAN Inversion (arxiv.org/abs/2004.00049)	Hong Kong
Dec 2020 – GAN Latent Discovery (arxiv.org/abs/2002.03754)	Yandex
Jun 2020 – Big GANs Are Watching (arxiv.org/abs/2006.04988)	Yandex
Nov 2020 – Navigating GAN Space (arxiv.org/abs/2011.13786)	Yandex

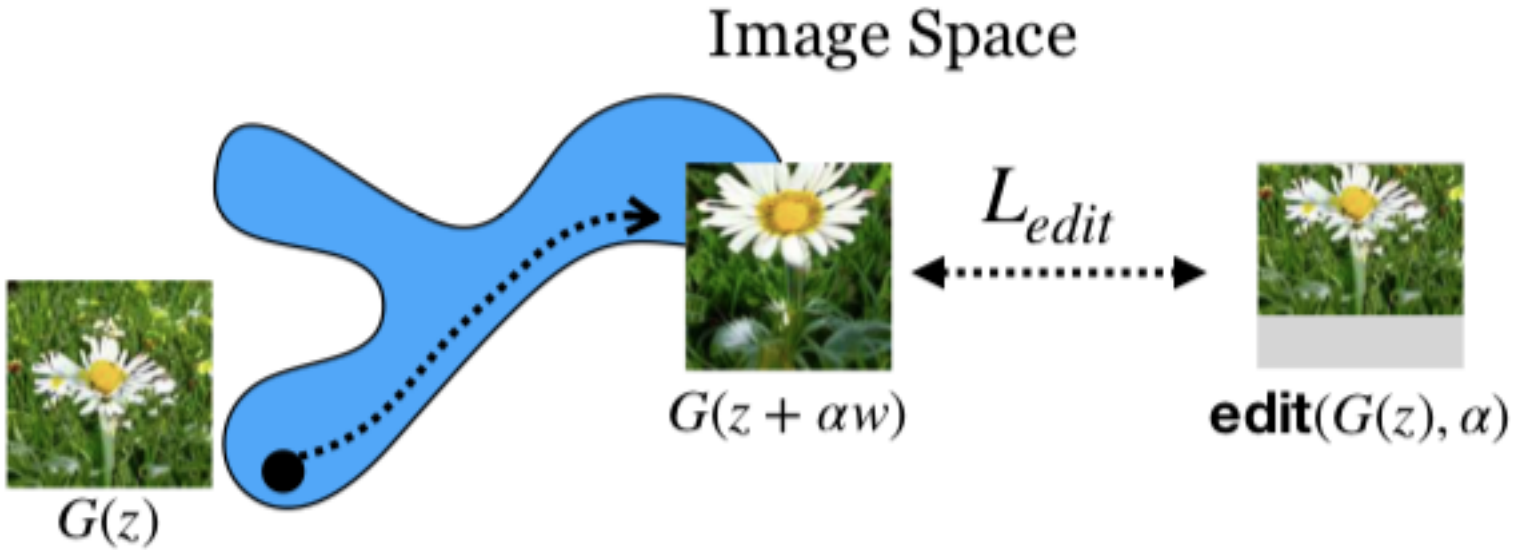
Massachusetts Institute of Technology

- On the ”Steerability" of Generative Adversarial Networks

Steerability of GANs

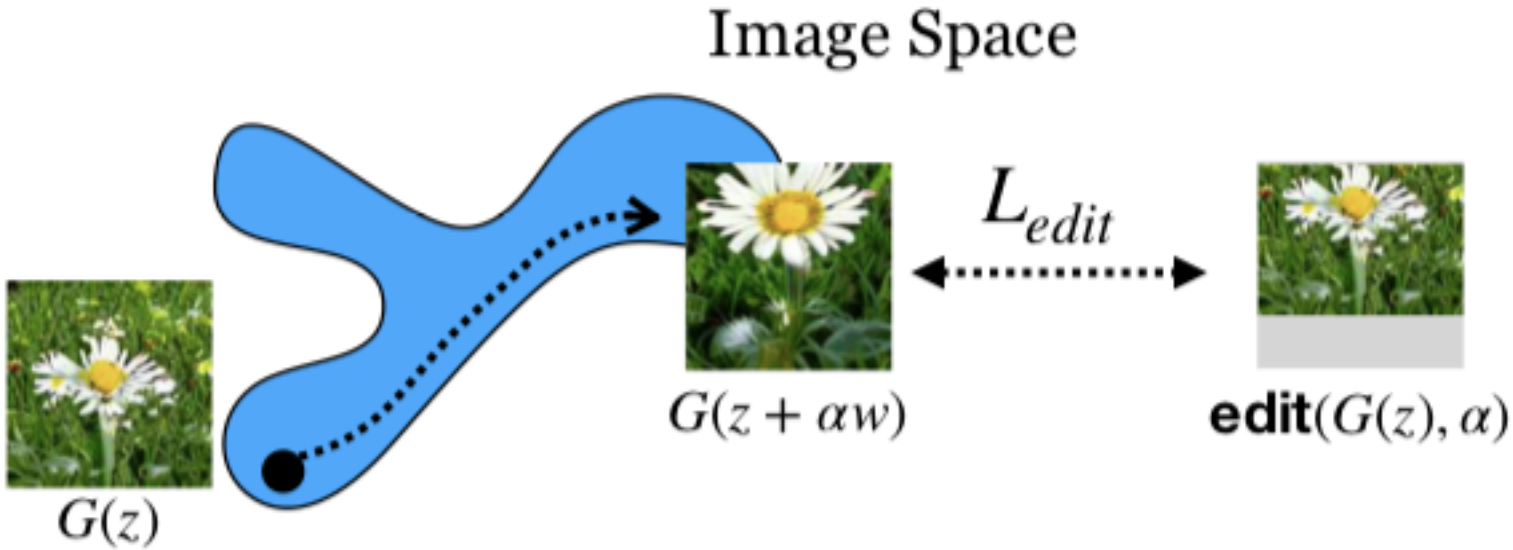


Steerability of GANs



$$w^* = \arg\min_w \mathbb{E}_{z,\alpha}[\mathcal{L}(G(z+\alpha w), \texttt{edit}(G(z), \alpha))].$$

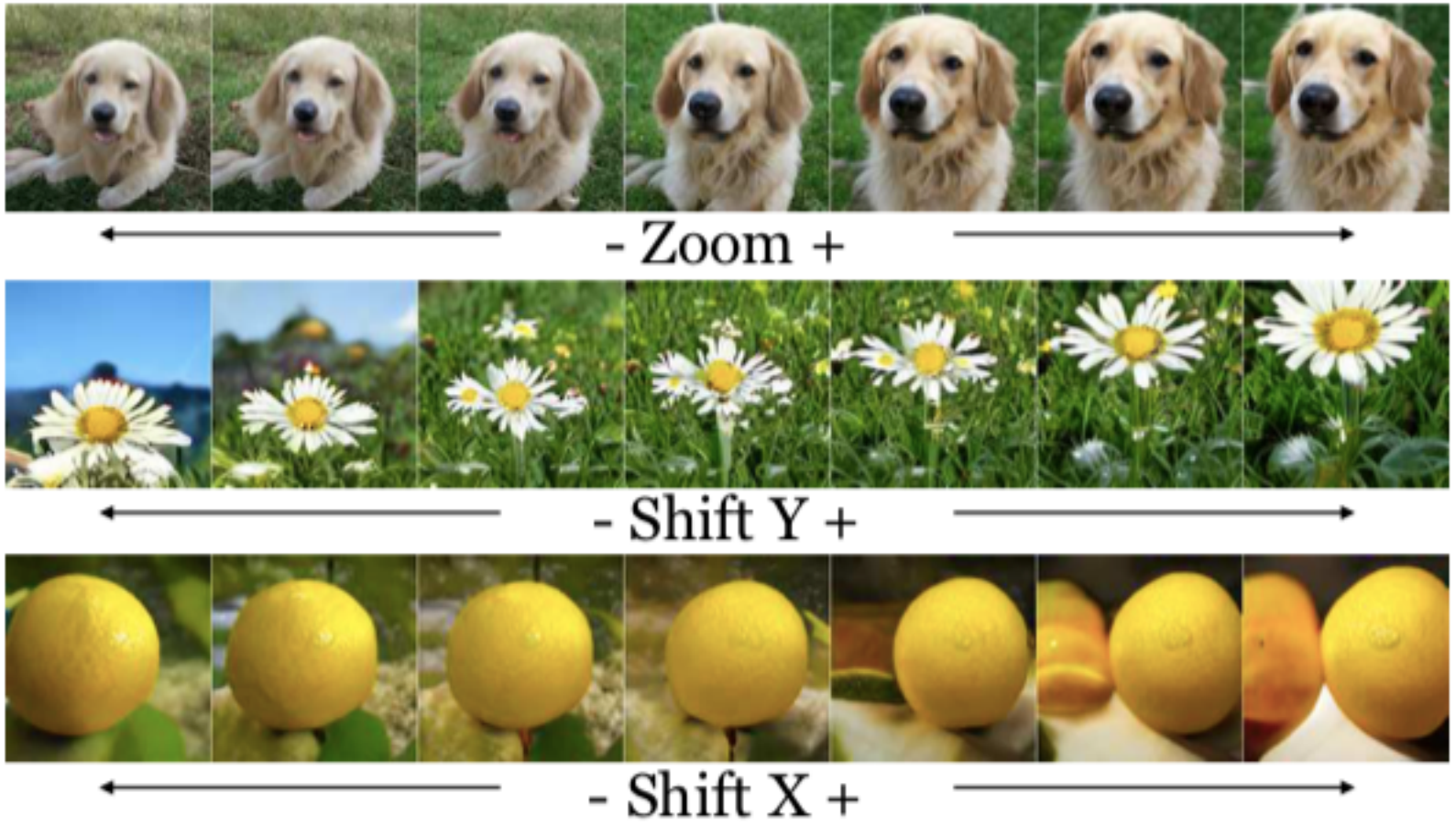
Steerability of GANs



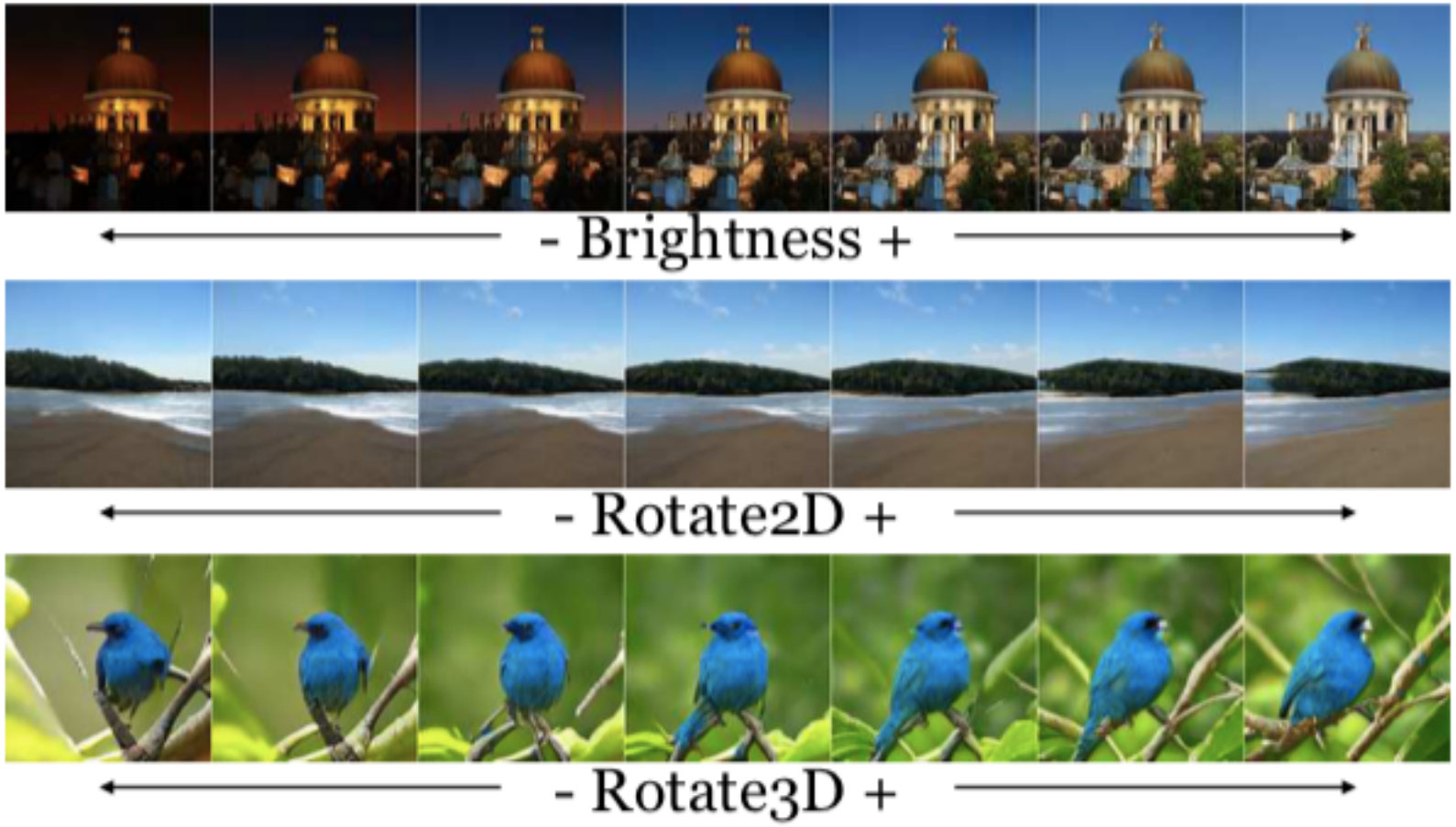
$$w^* = \arg\min_w \mathbb{E}_{z,\alpha}[\mathcal{L}(G(z+\alpha w), \texttt{edit}(G(z), \alpha))].$$

$$\mathcal{L} = \mathbb{E}_{z,n}[||G(f^n(z)) - \texttt{edit}(G(z), n\epsilon))||],$$

Steerability of GANs



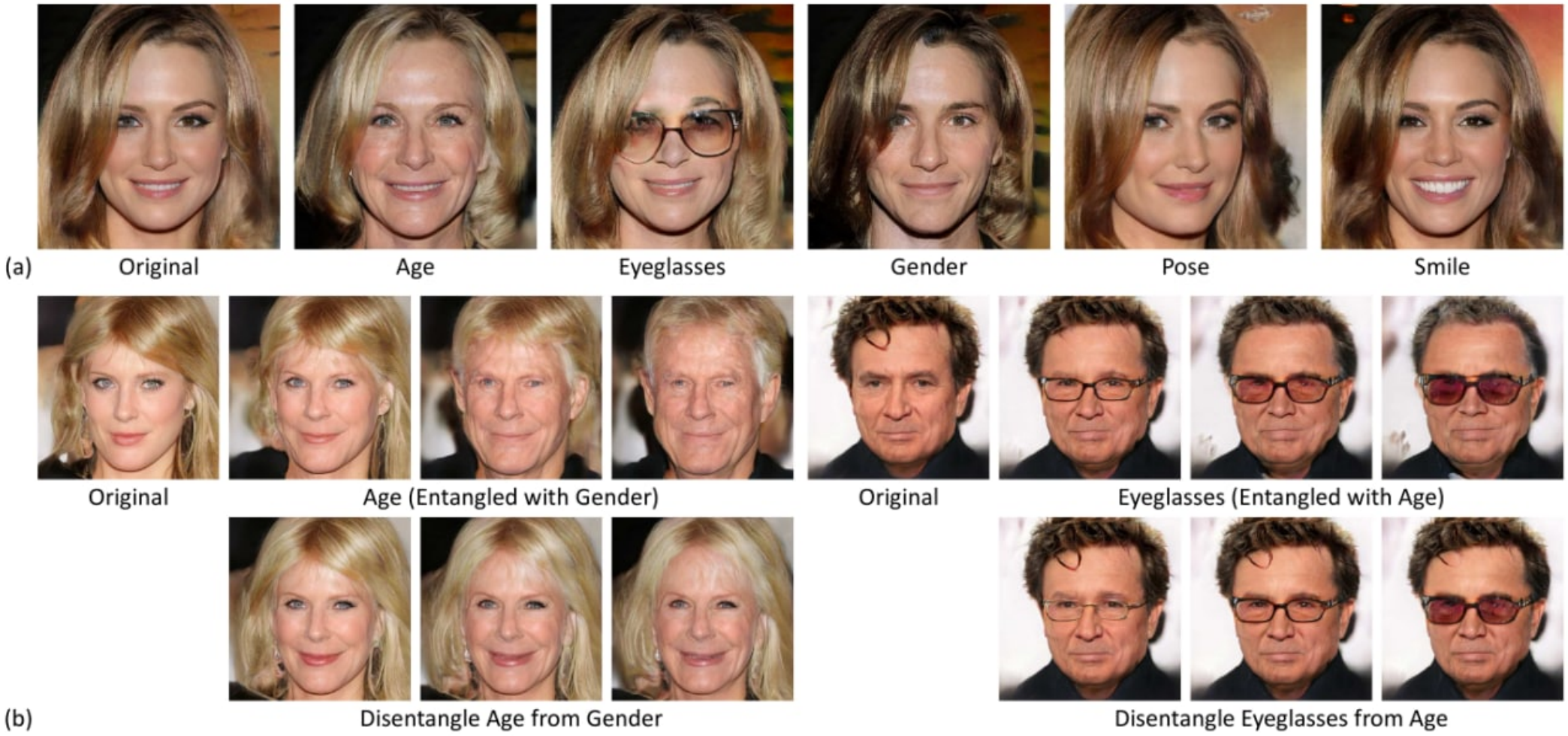
Steerability of GANs



Chinese University of Hong Kong

- Interpreting the Latent Space of GANs for Semantic Face Editing
- In-Domain GAN Inversion for Real Image Editing

Interpreting GANs



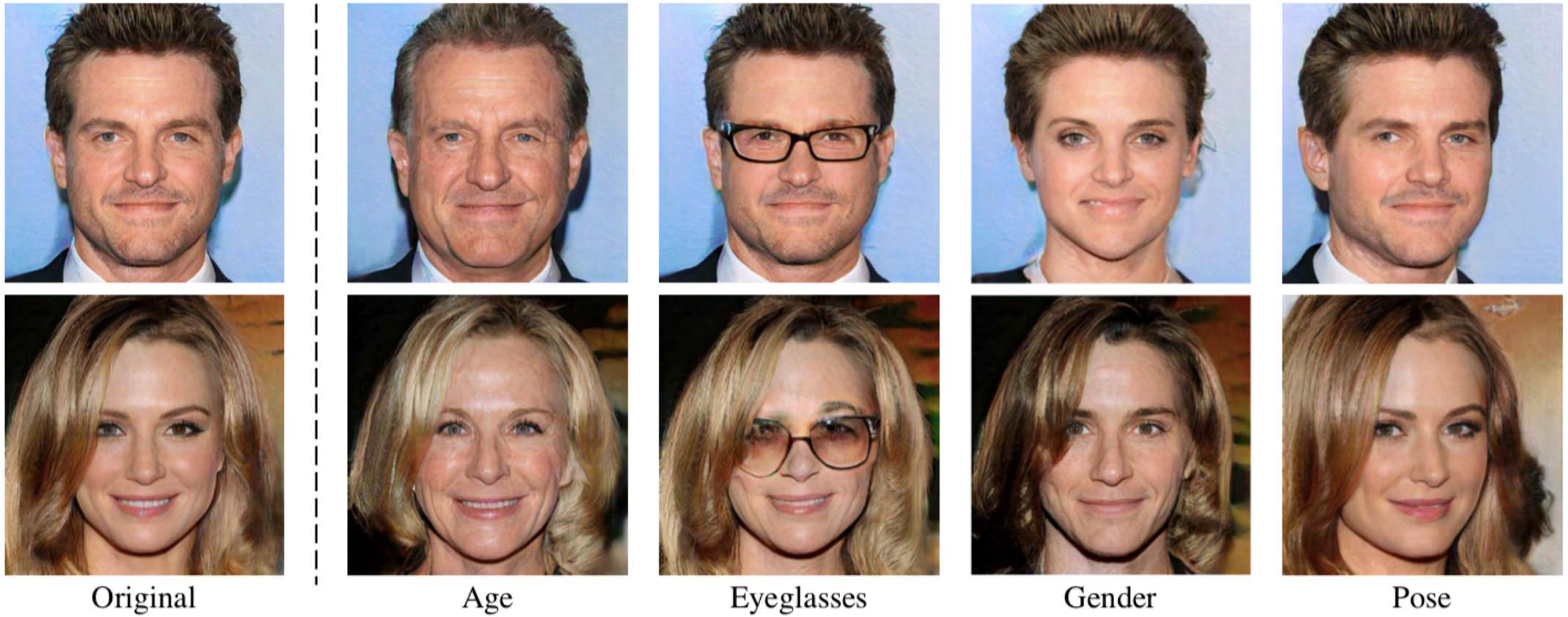
Interpreting GANs

1) StyleGAN inteprolation
2) Disentaglement problem

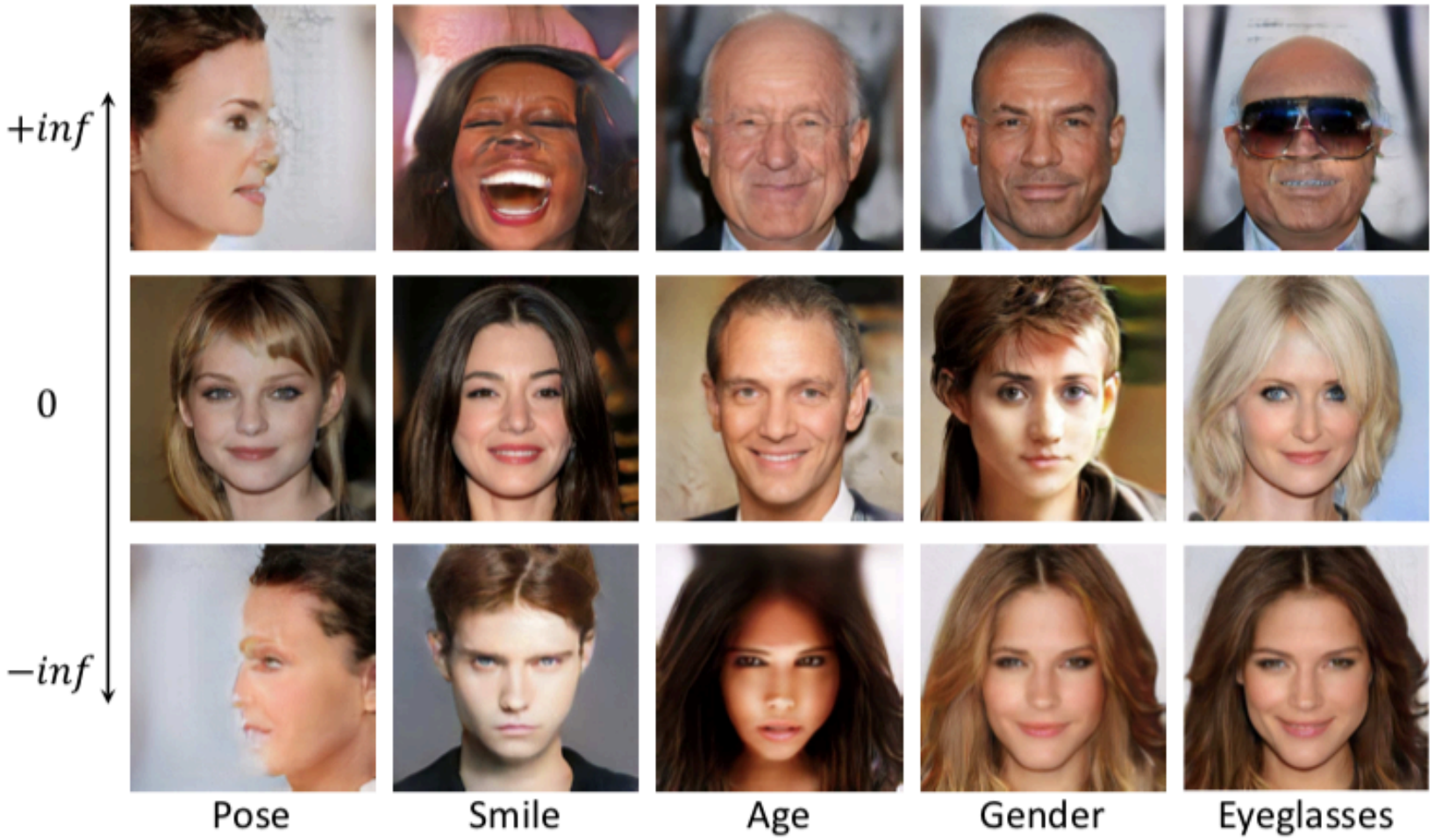
Interpreting GANs

1) StyleGAN inteprolation
2) Disentaglement problem
3) SVM for 5 attributes

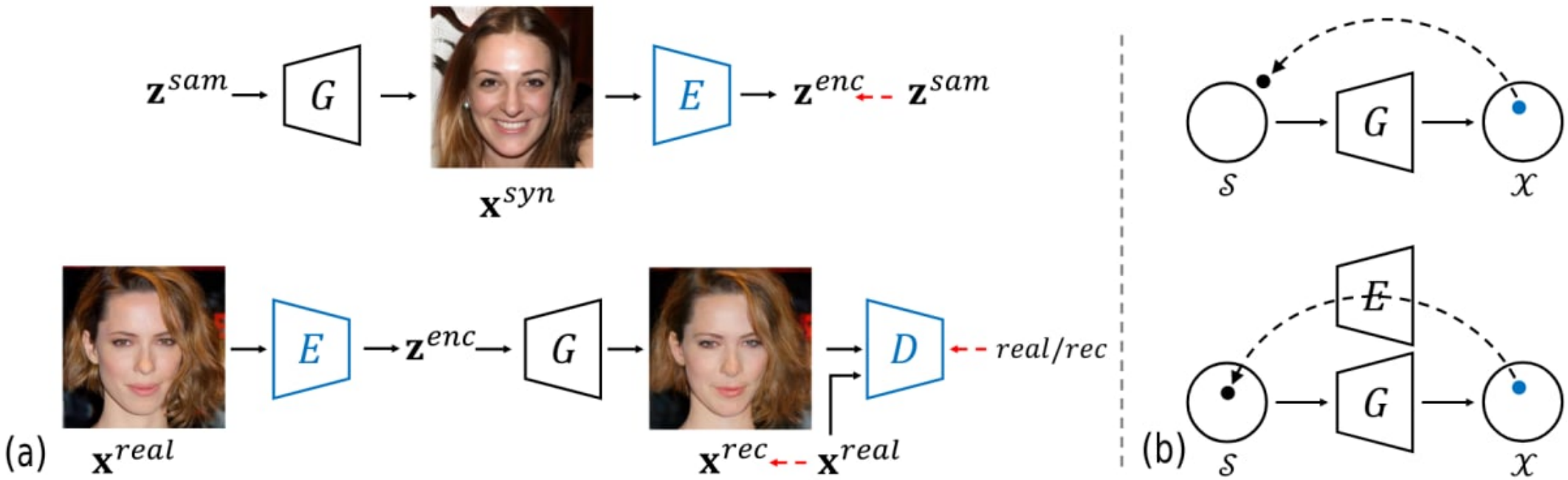
Interpreting GANs



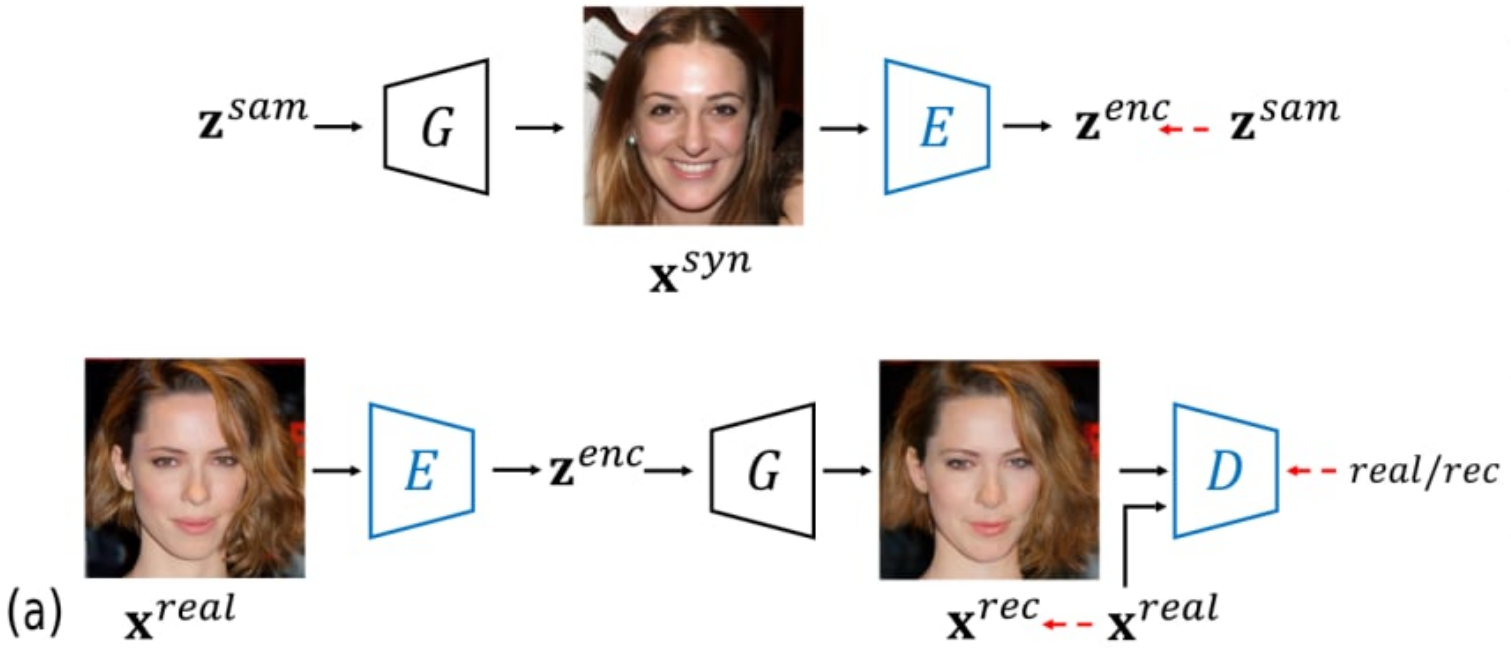
Interpreting GANs

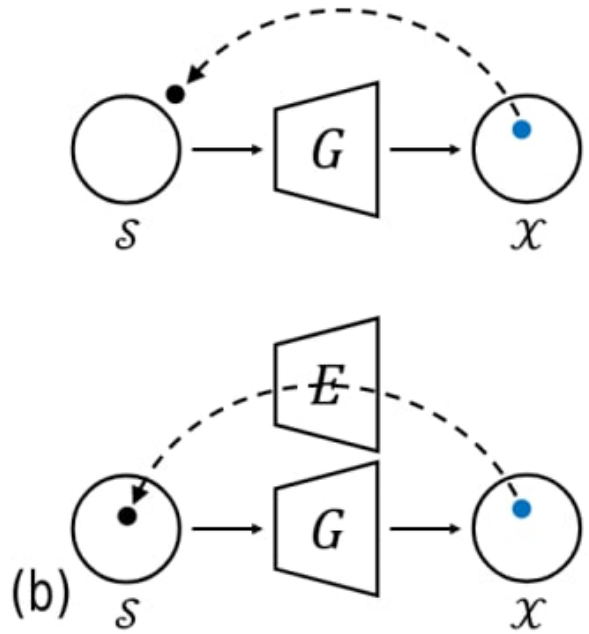


Gan Inversion



Gan Inversion





$$\min_{\Theta_E} \mathcal{L}_E = ||\mathbf{z}^{sam} - E(G(\mathbf{z}^{sam}))||_2, \quad (1)$$

Gan Inversion

$$\min_{\Theta_E} \mathcal{L}_E = ||\mathbf{z}^{sam} - E(G(\mathbf{z}^{sam}))||_2, \qquad (1)$$

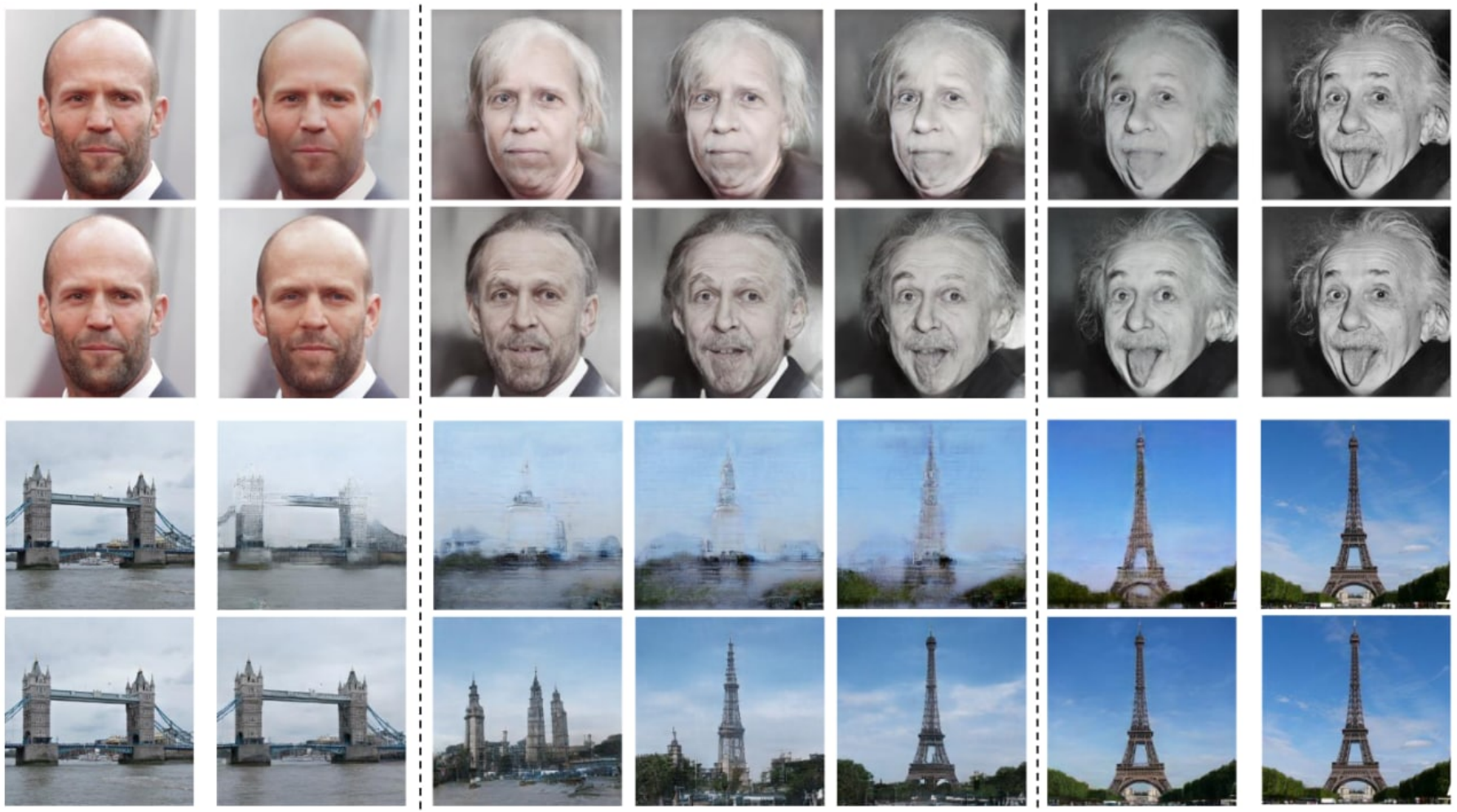
Gan Inversion

$$\min_{\Theta_E} \mathcal{L}_E = ||\mathbf{z}^{sam} - E(G(\mathbf{z}^{sam}))||_2, \tag{1}$$

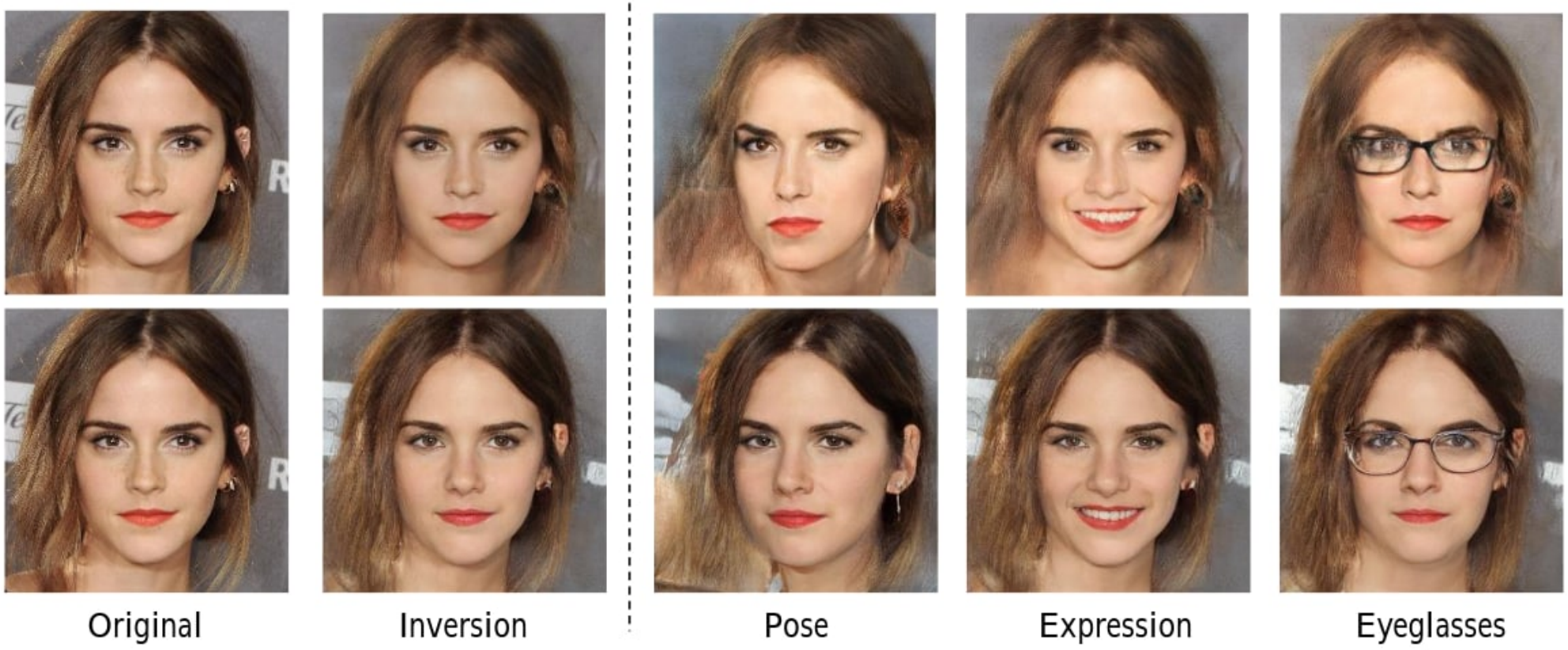
$$\begin{aligned}\min_{\Theta_E} \mathcal{L}_E = ||\mathbf{x}^{real} - G(E(\mathbf{x}^{real}))||_2 \ &+ \lambda_1 ||F(\mathbf{x}^{real}) - F(G(E(\mathbf{x}^{real})))||_2 \\ &- \lambda_2 \mathop{\mathbb{E}}_{\mathbf{x}^{real} \sim P_{data}} [D(G(E(\mathbf{x}^{real})))],\end{aligned} \tag{2}$$

$$\begin{aligned}\min_{\Theta_D} \mathcal{L}_D = \mathop{\mathbb{E}}_{\mathbf{x}^{real} \sim P_{data}} [D(G(E(\mathbf{x}^{real})))] \ &- \mathop{\mathbb{E}}_{\mathbf{x}^{real} \sim P_{data}} [D(\mathbf{x}^{real})] \\ &+ \frac{\gamma}{2} \mathop{\mathbb{E}}_{\mathbf{x}^{real} \sim P_{data}} [||\nabla_{\mathbf{x}} D(\mathbf{x}^{real})||_2^2],\end{aligned} \tag{3}$$

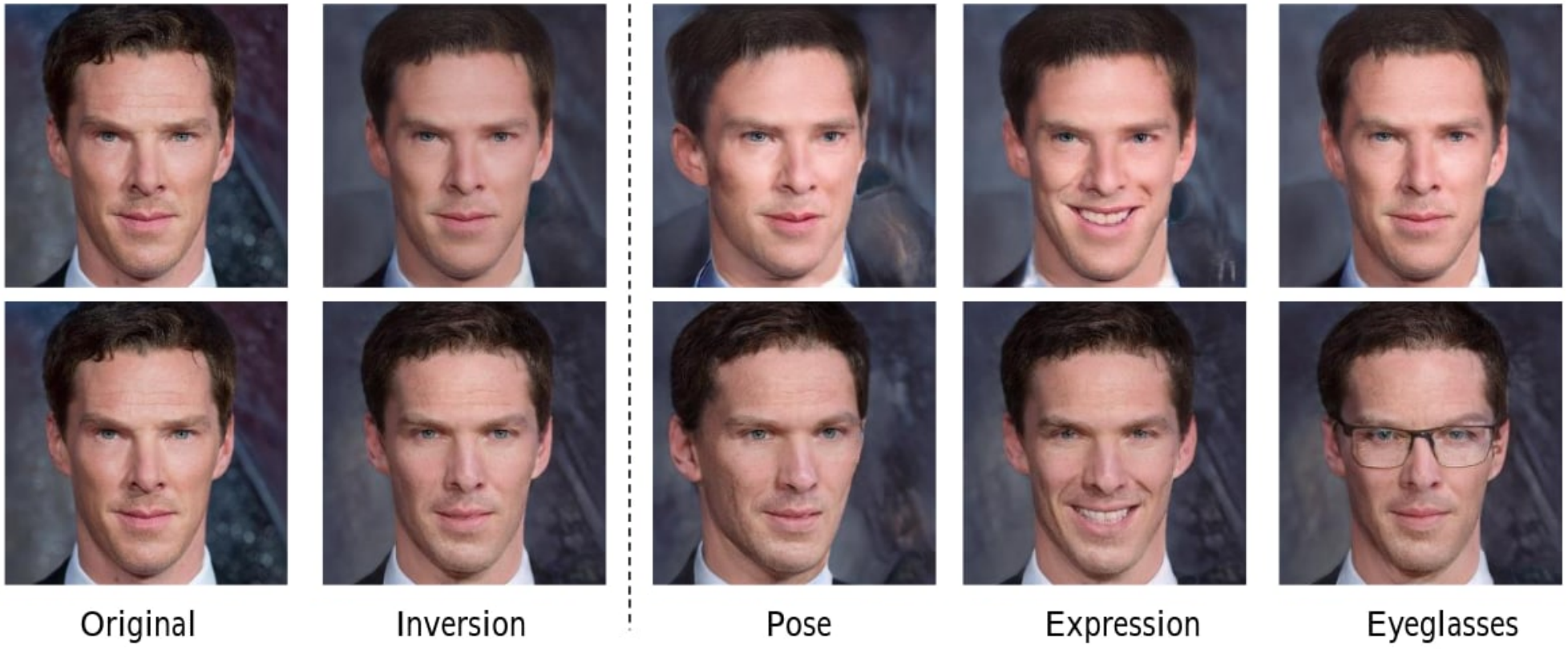
Gan Inversion



Gan Inversion



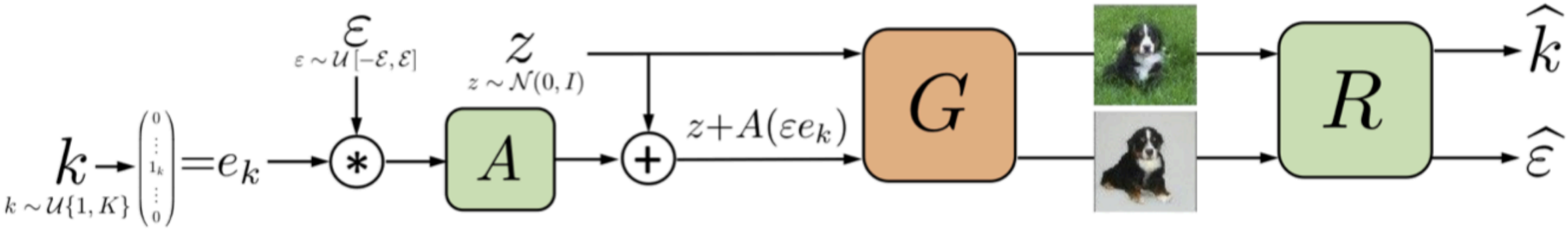
Gan Inversion



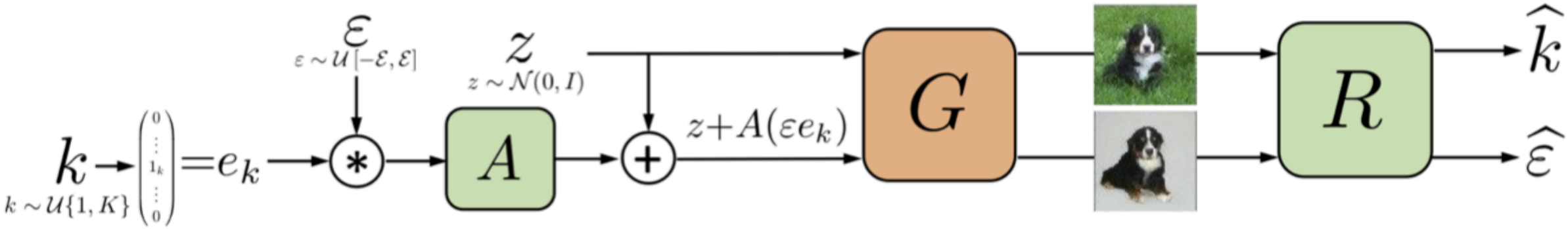
Yandex Research

- Unsupervised Discovery of Interpretable Directions in the GAN Latent Space
- Navigating the GAN Parameter Space for Semantic Image Editing
- Big GANs Are Watching You: Towards Unsupervised Object Segmentation with Off-the-Shelf Generative Models

GAN Latent Discovery

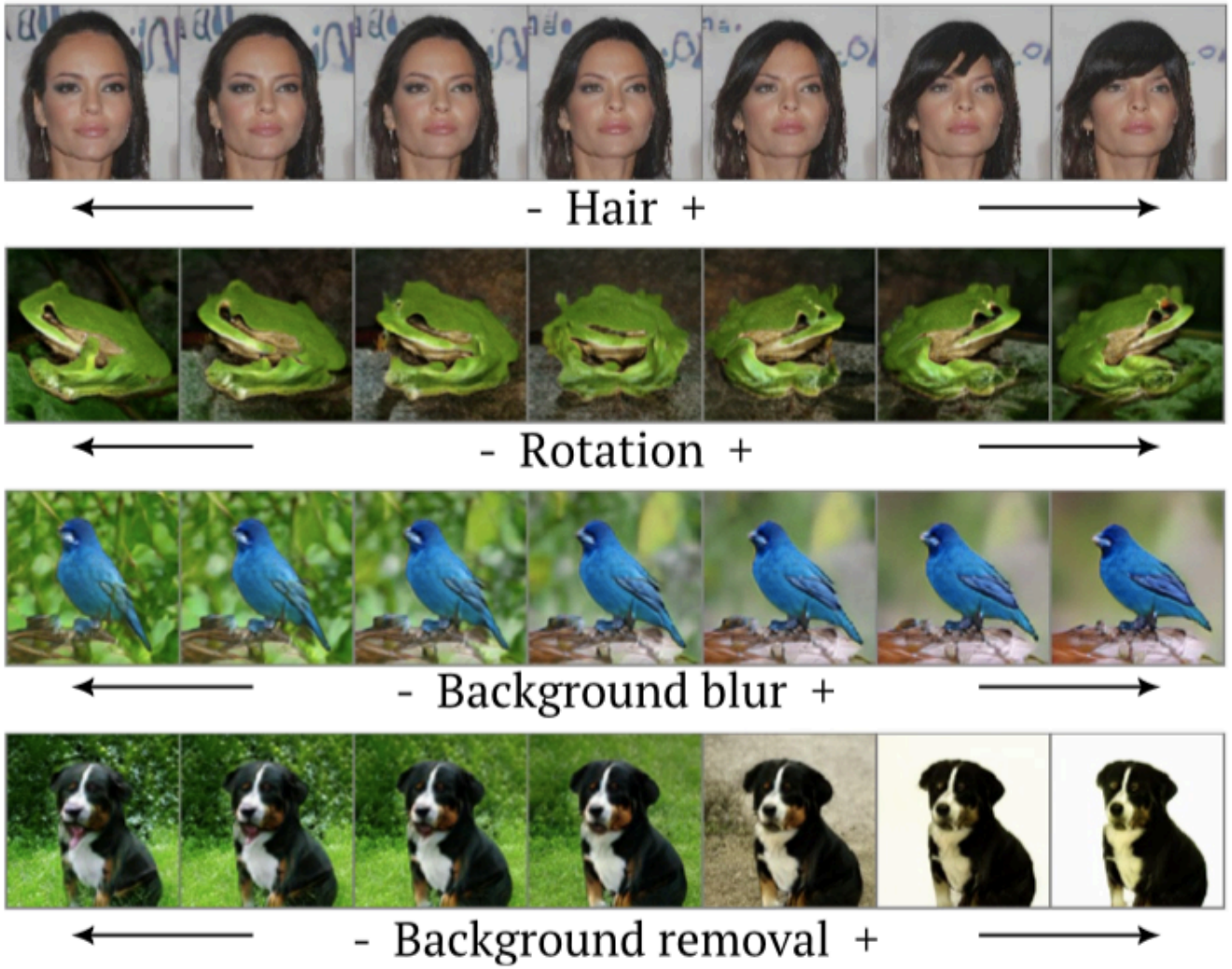


GAN Latent Discovery

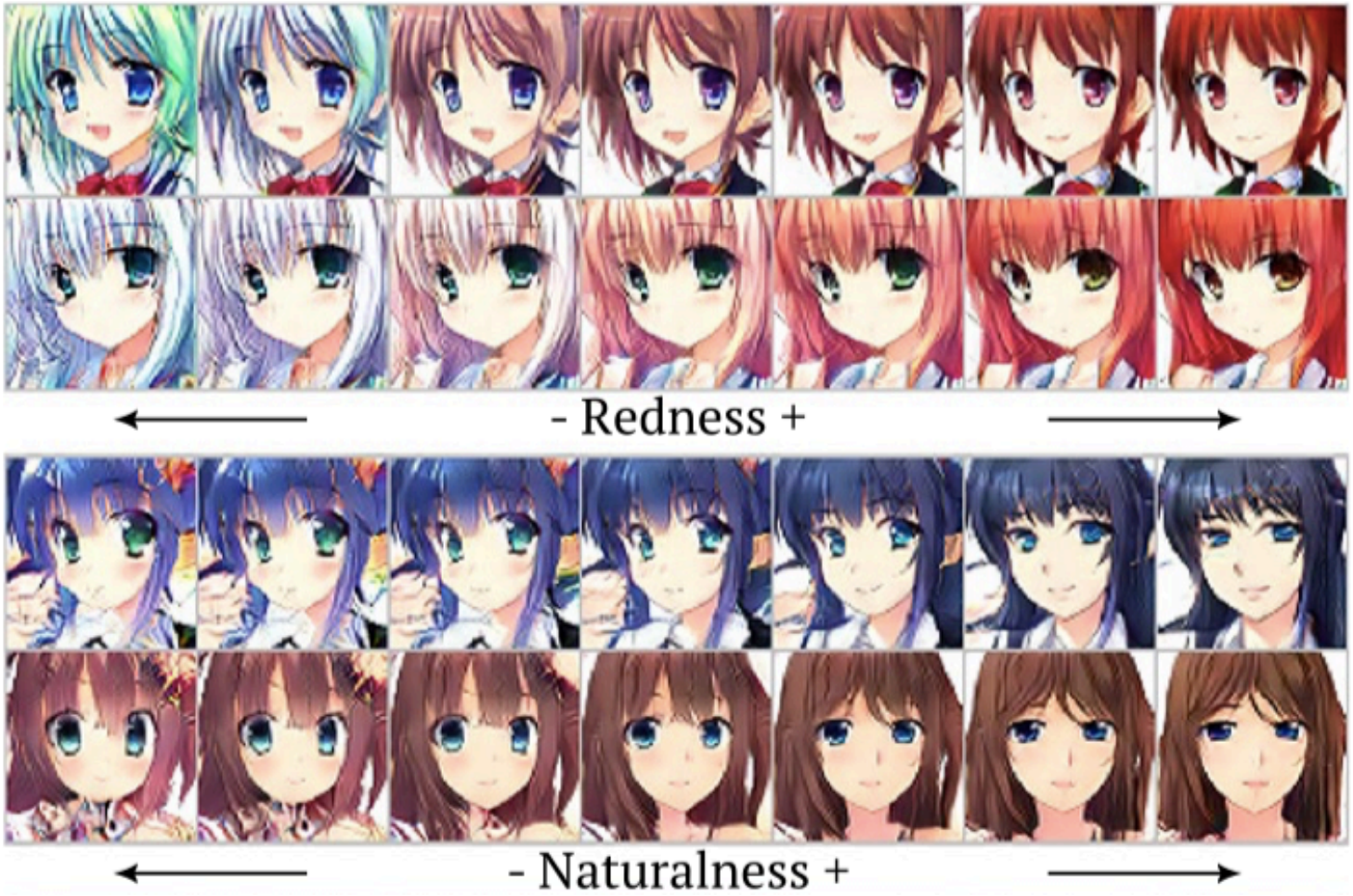


$$\min_{A,R} \mathop{\mathbb{E}}_{z,k,\varepsilon} L(A,R) = \min_{A,R} \mathop{\mathbb{E}}_{z,k,\varepsilon} \left[L_{cl}(k, \widehat{k}) + \lambda L_r(\varepsilon, \widehat{\varepsilon}) \right]$$

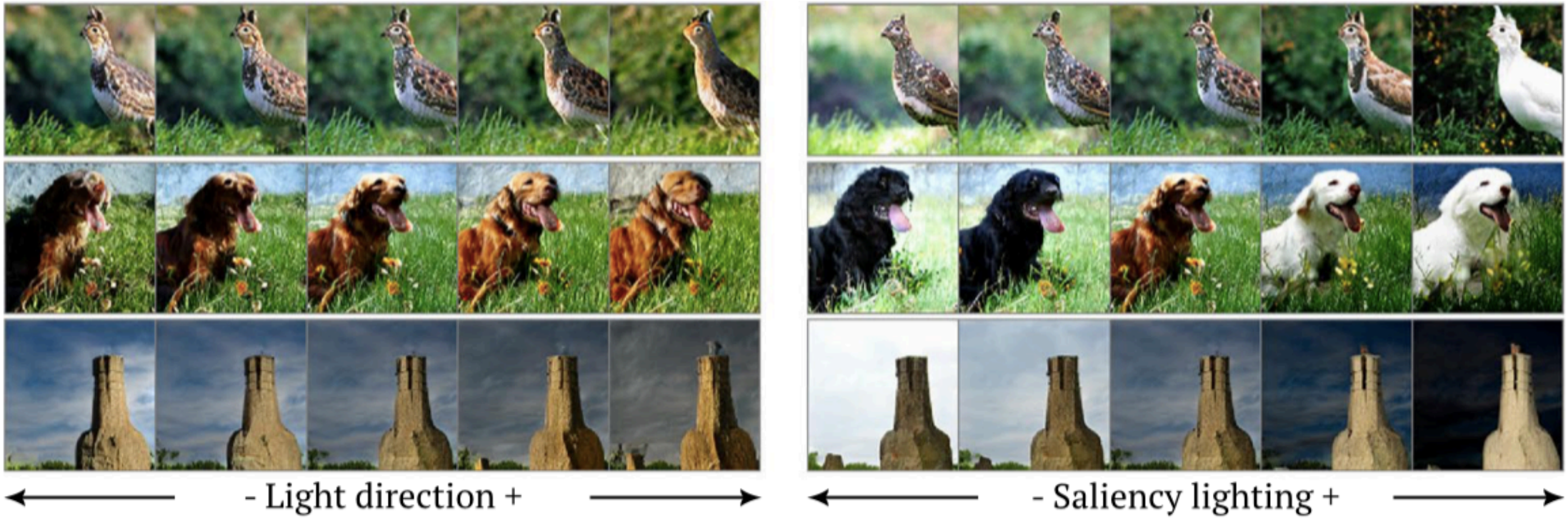
GAN Latent Discovery



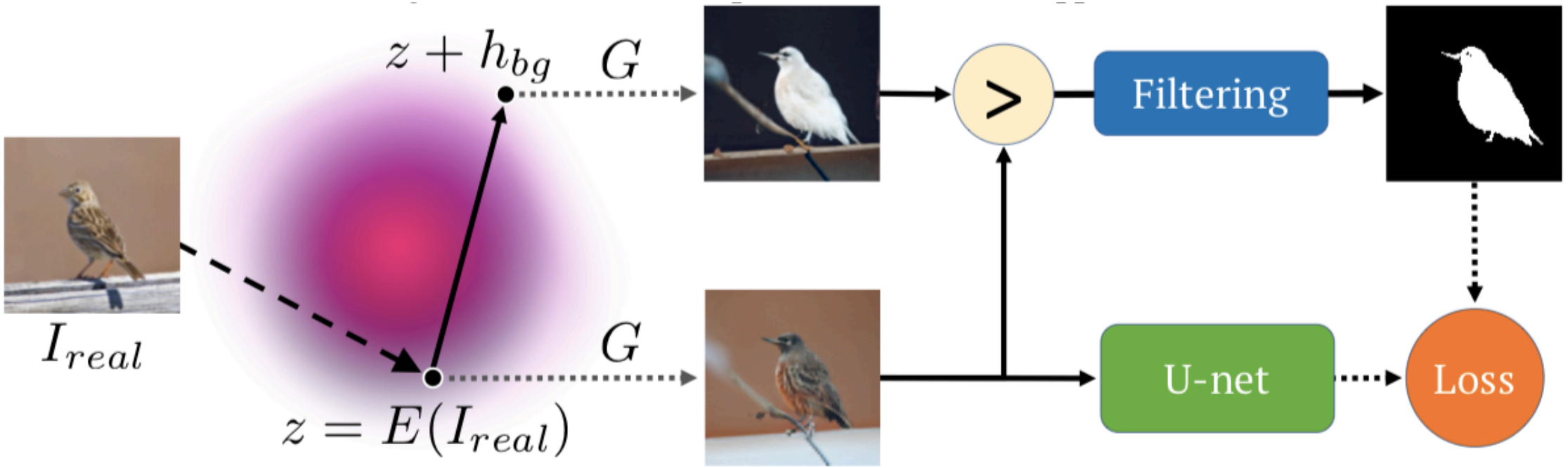
GAN Latent Discovery



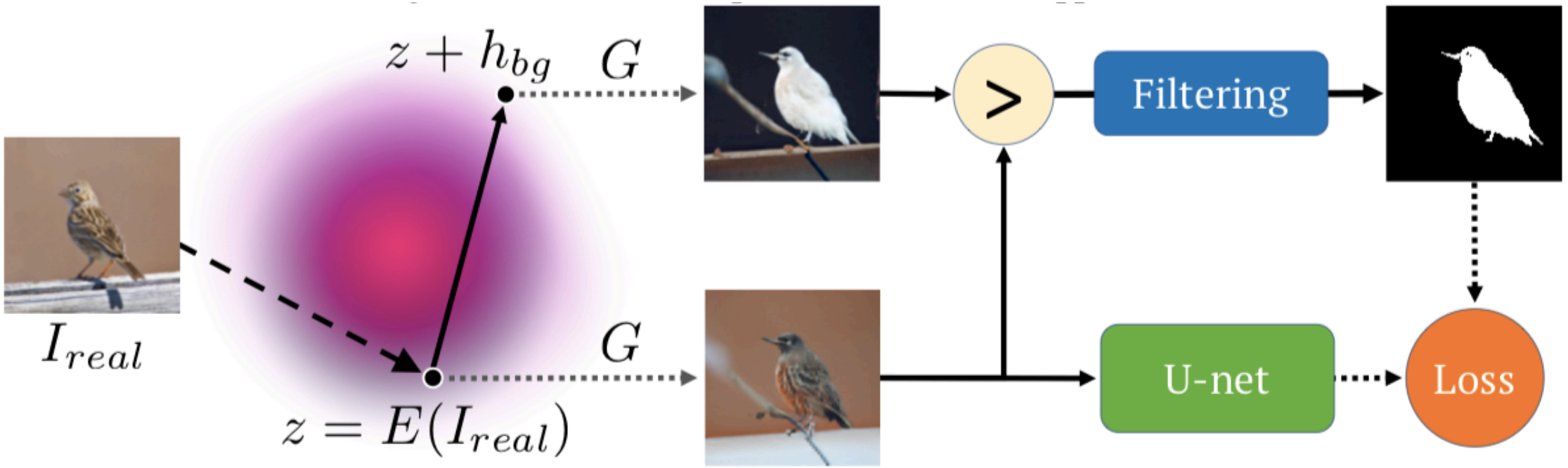
Big GANs Are Watching



Big GANs Are Watching



Big GANs Are Watching



$$M = [G(z+h_{bg}) > G(z)] \qquad ||h_{bg}|| = 5$$

Big GANs Are Watching

Figure 5: *Top:* Images from the DUTS-test dataset. *Middle:* Groundtruth masks. *Bottom:* Masks produced by the E-BigBiGAN method.

Bibliography

Oct 2017 – ProGAN (arxiv.org/abs/1710.10196)	NVIDIA
Dec 2018 – StyleGAN (arxiv.org/abs/1812.04948)	NVIDIA
Dec 2019 – StyleGAN2 (arxiv.org/abs/1912.04958)	NVIDIA
Jul 2019 – Steerability of GANs (arxiv.org/abs/1907.07171)	MIT
Jul 2019 – Interpreting GANs (arxiv.org/abs/1907.10786)	Hong Kong
Apr 2020 – GAN Inversion (arxiv.org/abs/2004.00049)	Hong Kong
Dec 2020 – GAN Latent Discovery (arxiv.org/abs/2002.03754)	Yandex
Jun 2020 – Big GANs Are Watching (arxiv.org/abs/2006.04988)	Yandex
Nov 2020 – Navigating GAN Space (arxiv.org/abs/2011.13786)	Yandex