Computer Vision Seminar (MA-INF 2206, SS2020)

UNIVERSITÄT BONN

Paper: Interpreting the Latent Space of GANs for Semantic Face Editing

Student: Ani Karapetyan
Professor: Prof. Dr. Jürgen Gall

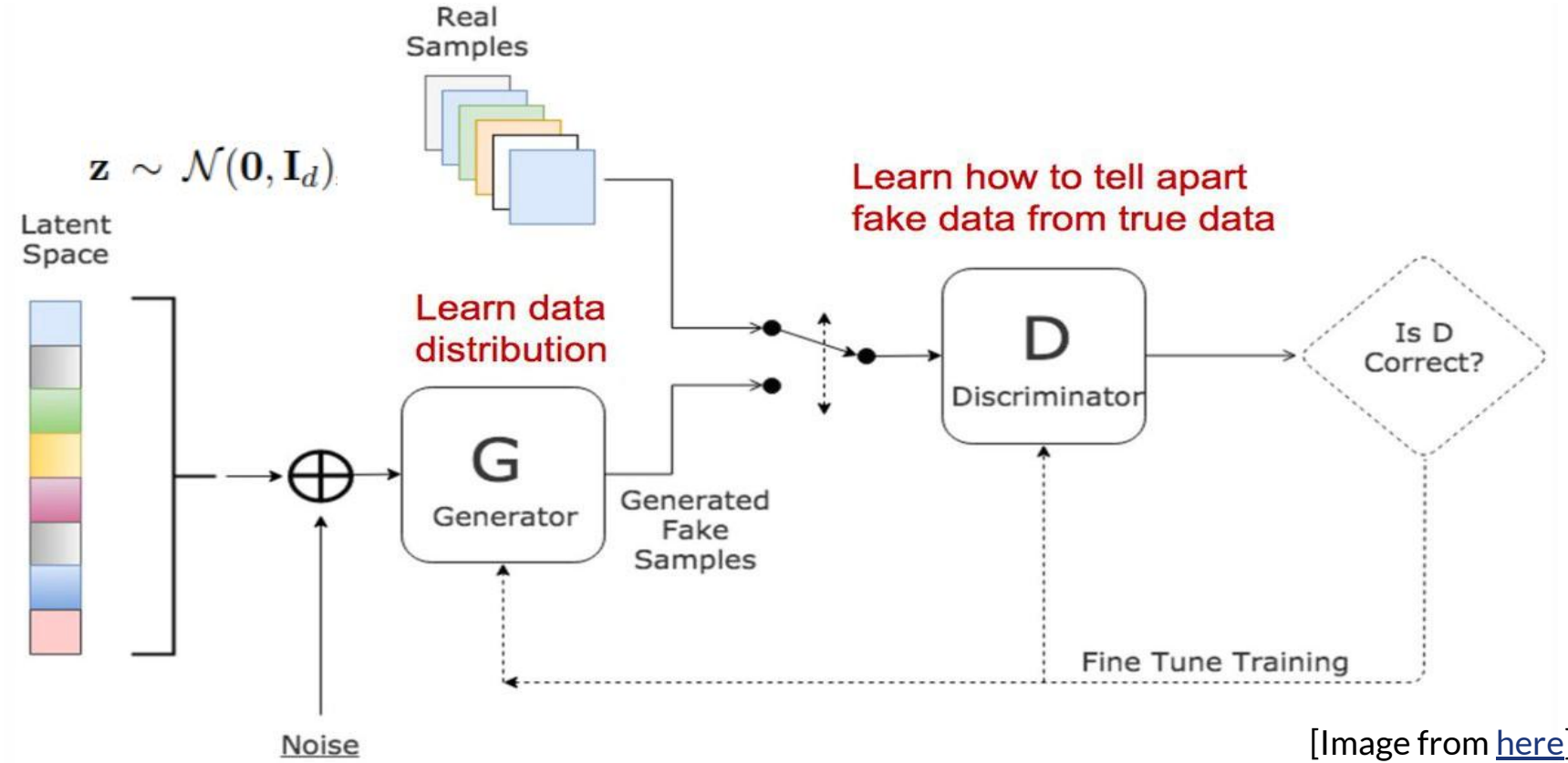
https://openaccess.thecvf.com/content_CVPR_2020/papers/Shen_Interpreting_the_Latent_Space_of_GANs_for_Semantic_Face_Editing_CVPR_2020_paper.pdf
https://arxiv.org/pdf/2005.09635.pdf

Outline

- General GAN architecture
- Motivation of the paper
- Related work
- Framework of InterFaceGAN
- Implementation details
- Experiments

GAN Architecture [16]

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



[Image from here]

What Do GANs Learn?

UNIVERSITÄT BONN

Visualizing and understanding GANs

(Bau et al: Visualizing and understanding generative adversarial networks, 2019 [1])

- The type of information explicitly represented changes from layer to layer, e.g. some units of intermediate layers are specialized to synthesize certain semantic objects (sofas, TV, etc).
- Units can match instances of an object class with diverse visual appearance, e.g. emerging dining table regions with different colors, geometry and materials.
- Units that emerge match object classes corresponding to the training dataset scene type, e.g. for kitchen scenes we find units that match stoves, cabinets, stool legs, etc.

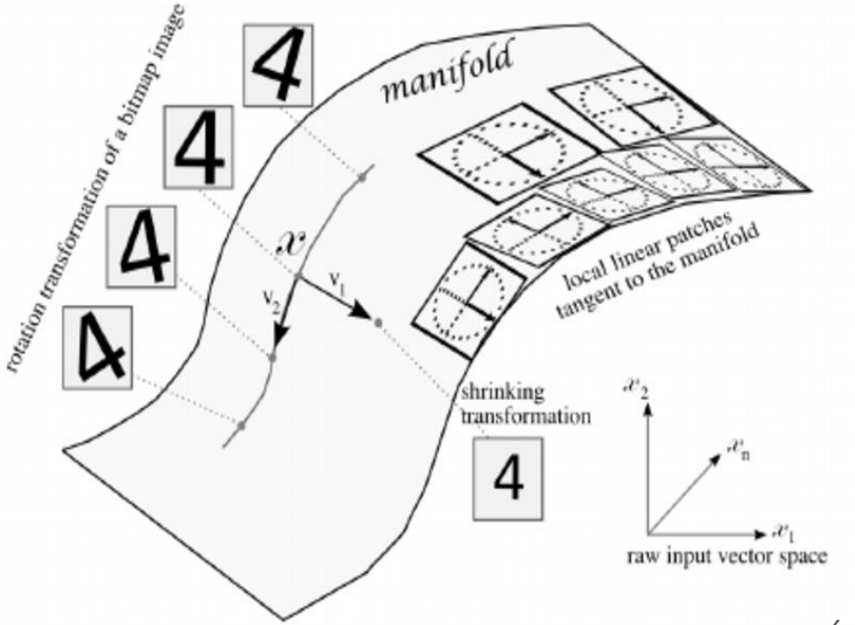
Motivation

Lack of enough understanding of:

- How various semantics are encoded in the latent space learnt by a well-trained GAN.
- How to manipulate the latent space in order to get photo-realistic image editing.
- How to decouple some entangled semantics in latent space to achieve conditional attribute manipulation.

Related Work/Study on Latent Space

- Latent space of GANs is generally treated as a **Riemannian manifold** ([2]).
- *Manifold hypothesis:* data presented in high dimensional space is expected to concentrate in the vicinity of a manifold of much lower dimensionality.



Related Work/Study on Latent Space

UNIVERSITÄT BONN

Interpolation in Latent Space

(Shao et al.: The Riemannian Geometry of Deep Generative Models, 2018 [3])

- ***Geodesic interpolation:***
  Smooth interpolation in image space achieved by interpolating between latent codes along geodesic path on the low-dimensional manifold.
- Minimize *path energy* in image space:

$$E_{z_i} = \frac{1}{2} \sum_{i=0}^{T} \frac{1}{\delta t} \| g(z_{i+1}) - g(z_i) \|^2$$

- Manifolds learnt by GANs are surprisingly close to zero curvature.
- Linear paths in the latent space closely approximate geodesics on the generated manifold.

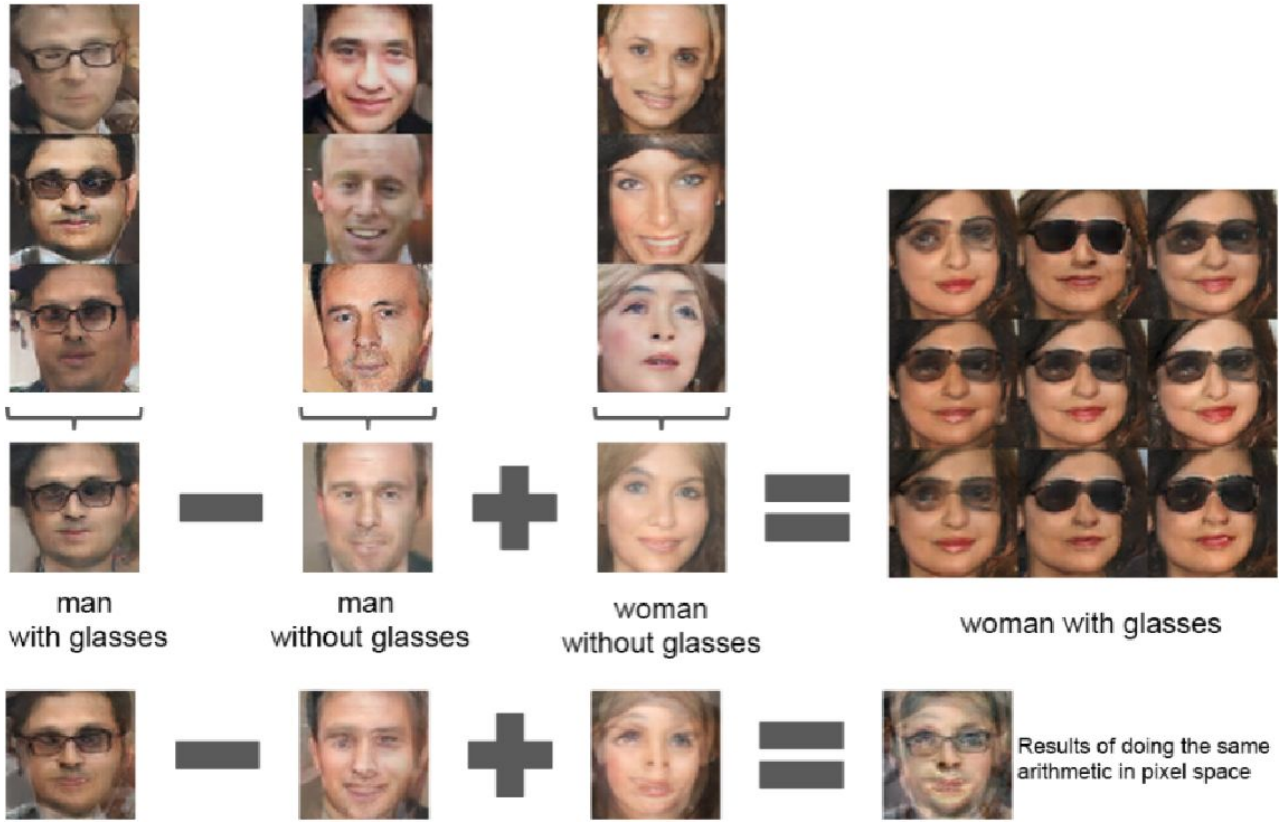
Interpolation in Latent Space

(Wu et al.: Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling, 2016)

Related Work/Study on Latent Space

Vector Arithmetic Property in Latent Space

(Radford et al: Unsupervised representation learning with DCGAN, 2016 [4])



Related Work/Study on Latent Space

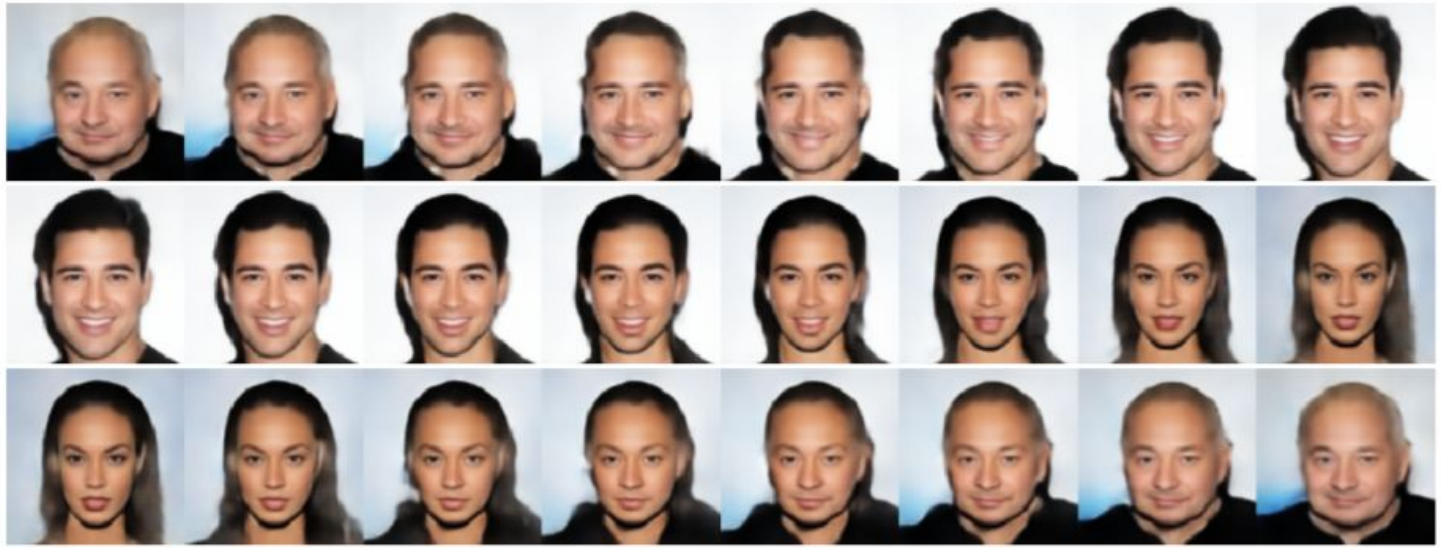
UNIVERSITÄT BONN

Generative Latent Optimization (GLO)

(Bojanowski et al: Optimizing the Latent Space of Generative Networks, 2018 [5])

- "Encoderless" Autoencoder or "Discriminator-less" GAN.
- Simple reconstruction loss:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} [\min_{z_i \in Z} L(g_\theta(z_i), x_i)]$$

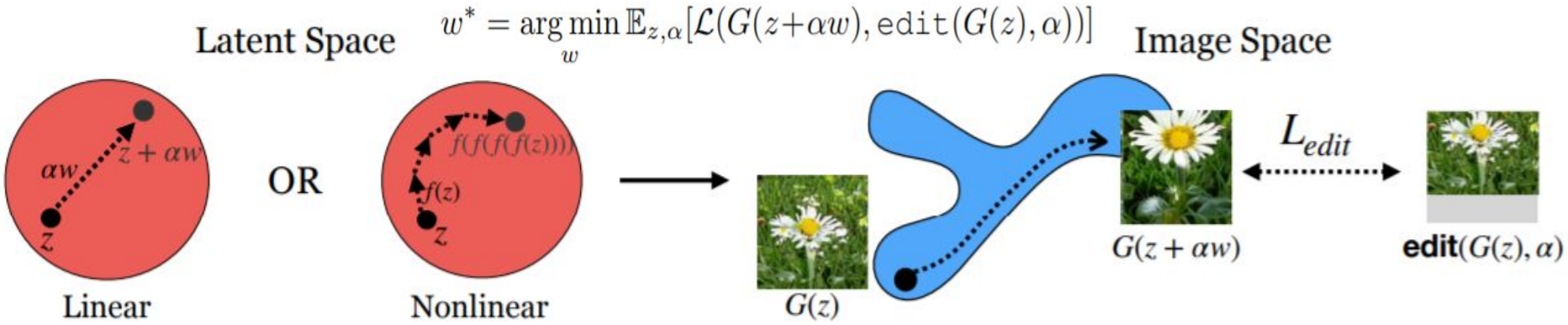


(interpolation results between the 3 images with GLO)

Related Work/Study on Latent Space

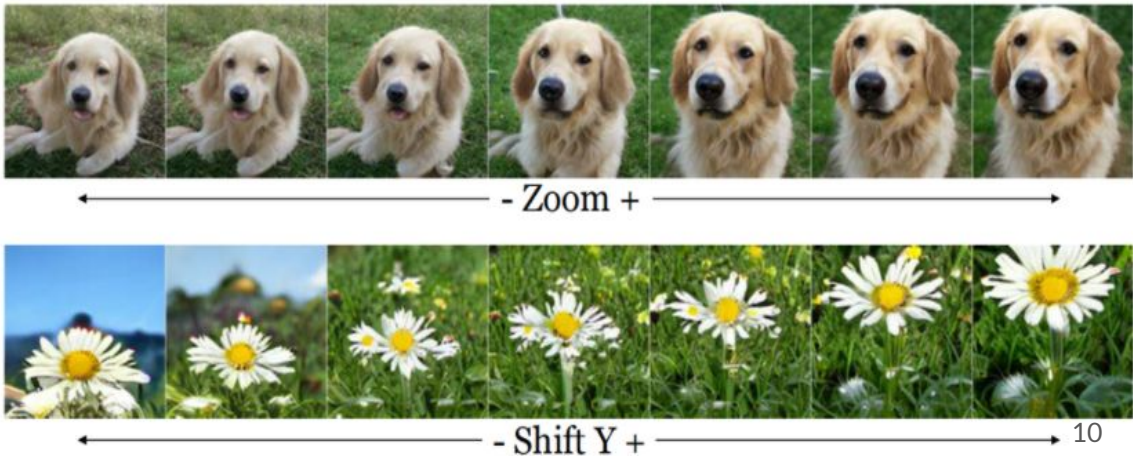
On Steerability of GANs

(Jahanian et al.: On the "steerability" of generative adversarial networks, 2020 [6])



Learn a linear walk w in the latent space that matches a predefined transformation in image space, e.g.:

- *Changing brightness*
- *Rotating*
- *Zooming in/out*
- *Shifting*

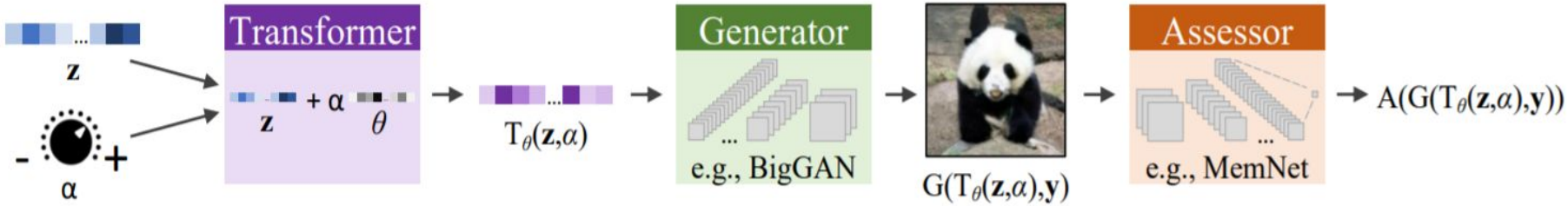


Related Work/Study on Latent Space

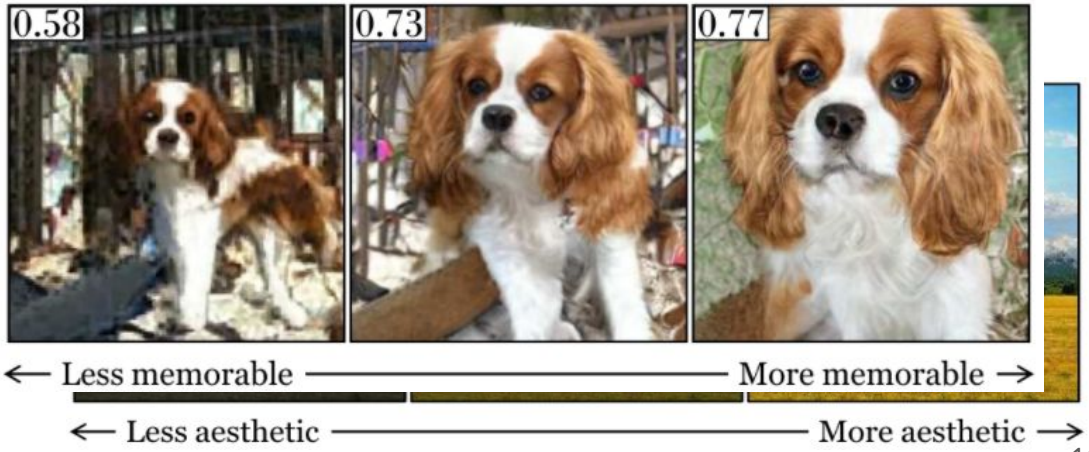
GANalyze

(Goetschalckx et al.: Ganalyze: Toward visual definitions of cognitive image properties, 2019 [7])

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z},\mathbf{y},\alpha}[(A(G(T_\theta(\mathbf{z},\alpha),\mathbf{y})) - (A(G(\mathbf{z},\mathbf{y})) + \alpha))^2]$$



Find a transformation (direction θ) in latent space, to manipulate with given cognitive property of interest of output image (e.g. memorability or aesthetics), by simply moving latent code z along that direction by score α.



Related Work/Semantic Face Editing with GANs

UNIVERSITÄT BONN

InfoGAN

(Chen et al.: Infogan: Interpretable representation learning by information maximizing GANs, 2016 [8])

- Decomposes latent vector into 2 parts: unstructured source of noise z (to control the identity) and several codes c (to control the salient semantic features).
- *Inform.-theoretic regularization term* encourages high mutual information between code c and the generator distribution to maximize the non-trivial dependency of G on c.

$$\min_G \max_D \; [V(D,G) - \lambda I(c; G(z,c))]$$

$$where\ I(X,\ Y) = H(X) - H(X|Y)$$

(d) Emotion

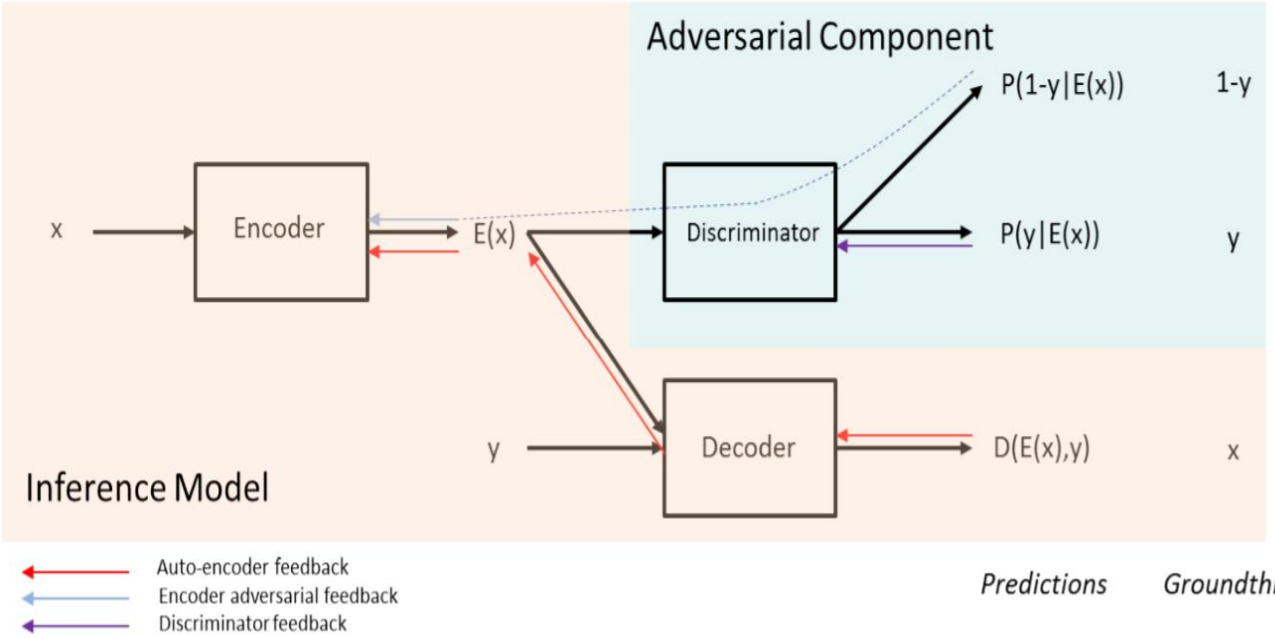
Related Work/Semantic Face Editing with GANs

UNIVERSITÄT BONN

Fader Networks

(Lample et al.: Fader networks: Manipulating images by sliding attributes, 2017 [9])

Encoder-decoder loss: $$\mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dec}} | \theta_{\text{dis}}) = \frac{1}{m} \sum_{(x,y)\in\mathcal{D}} \left\| D_{\theta_{\text{dec}}}\left(E_{\theta_{\text{enc}}}(x), y\right) - x \right\|_2^2 - \lambda_E \log P_{\theta_{\text{dis}}}(1 - y | E_{\theta_{\text{enc}}}(x))$$

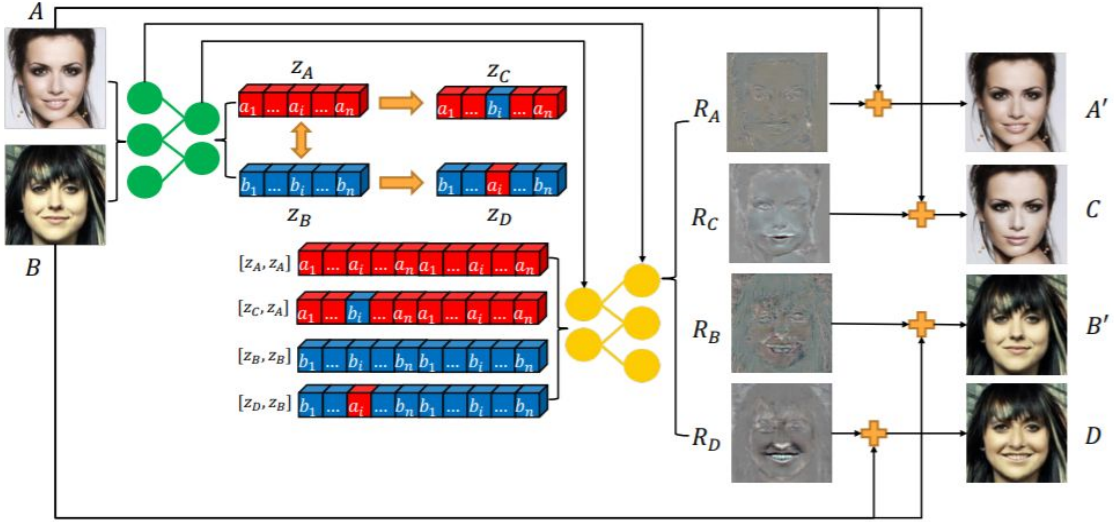


Related Work/Semantic Face Editing with GANs

UNIVERSITÄT BONN

Elegant

(Xiao et al.: Elegant: Exchanging Latent Encodings with GAN for Transferring Multiple Face Attributes, 2018 [10])



$$L_{\mathrm{D}_1} = -\mathbb{E}(\log(\mathrm{D}_1(A|Y^A))) - \mathbb{E}(\log(1-\mathrm{D}_1(C|Y^C))) - \mathbb{E}(\log(\mathrm{D}_1(B|Y^B))) - \mathbb{E}(\log(1-\mathrm{D}_1(D|Y^D)))$$

$$L_{\mathrm{D}_2} = -\mathbb{E}(\log(\mathrm{D}_2(A|Y^A))) - \mathbb{E}(\log(1-\mathrm{D}_2(C|Y^C))) - \mathbb{E}(\log(\mathrm{D}_2(B|Y^B))) - \mathbb{E}(\log(1-\mathrm{D}_2(D|Y^D)))$$

$$L_{\mathrm{D}} = L_{\mathrm{D}_1} + L_{\mathrm{D}_2}$$

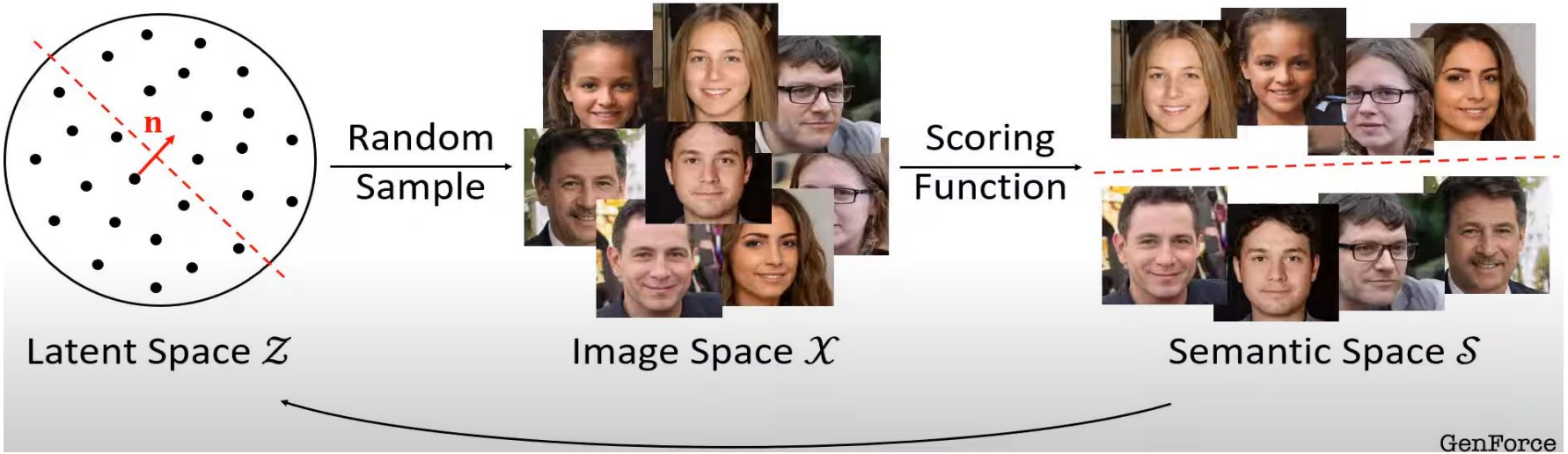
$$L_{reconstruction} = ||A - A'|| + ||B - B'||$$

$$L_{adv} = -\mathbb{E}(\log(\mathrm{D}_1(C|Y^C))) - \mathbb{E}(\log(\mathrm{D}_1(D|Y^D))) - \mathbb{E}(\log(\mathrm{D}_2(C|Y^C))) - \mathbb{E}(\log(\mathrm{D}_2(D|Y^D)))$$

$$L_G = L_{reconstruction} + L_{adv}$$

InterFaceGAN

1. Synthesize large number of images with randomly sampled latent codes.
2. Predict semantic scores using pre-trained classifiers.
3. Learn latent boundary (linear SVM) for each semantic.
4. Vary the latent code along boundary's normal direction to manipulate corresponding semantic.



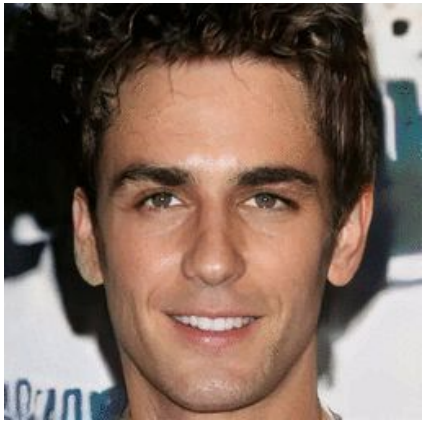
(Images from here)

InterFaceGAN/Semantic Manipulation Results

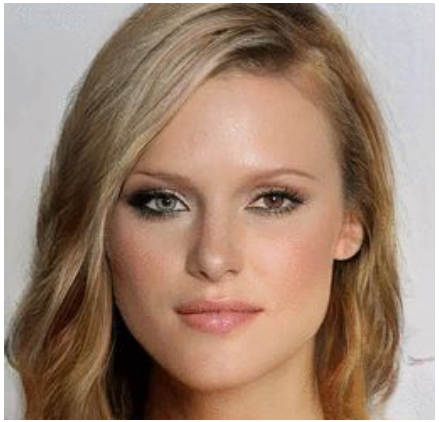
POSE

AGE

SMILE



EYEGLASSES



GENDER

ARTIFACTS

(GIF from here)

Semantics in Latent Space

UNIVERSITÄT BONN

Definitions

- $g : Z \rightarrow X$ – *generator as a deterministic function*
  - $Z \subseteq R^d$ – *d – dimensional latent space*
  - $z \sim N(0, I_d)$ – *z drawn from a Gaussian distribution*
  - X – *image space*
- $f_S : X \rightarrow S$ – *semantic scoring function*
  - $S \subseteq R^m$ – *semantic space with m semantics*
- $s = f_S(g(z))$ – *relation between latent and semantic spaces*
- $d(n, z) = n^T z$ – *signed distance from a sample z to the hyperplane defined by its normal vector n*

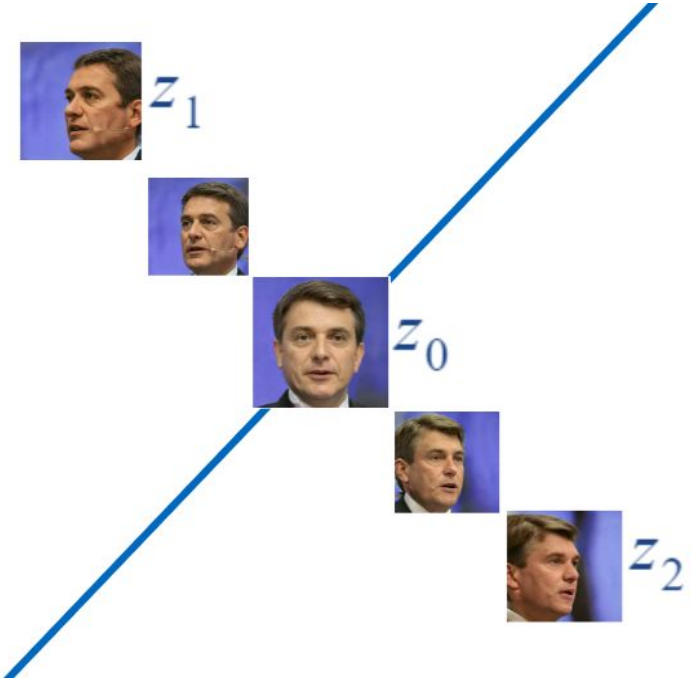
Single Semantic

Assumed $f_S(g(z_1)) * f_S(g(z_2)) < 0$.

Interpolation property in latent space => $\exists z_0, f_S(g(z_0)) = 0$.

=> Natural to expect that:

- Linear interpolation between z_1 and z_2 defines a hyperplane $n^T z = 0$, for $n \in R^d$.
- Linear relation near the boundary: $f(g(z)) = \lambda d(n, z)$.



Property 1 *Given* $\mathbf{n} \in \mathbb{R}^d$ *with* $\mathbf{n} \neq \mathbf{0}$, *the set* $\{\mathbf{z} \in \mathbb{R}^d : \mathbf{n}^T\mathbf{z} = 0\}$ *defines a hyperplane in* $\mathbb{R}^d$, *and* $\mathbf{n}$ *is called the normal vector. All vectors* $\mathbf{z} \in \mathbb{R}^d$ *satisfying* $\mathbf{n}^T\mathbf{z} > 0$ *locate from the same side of the hyperplane.*

Single Semantic

UNIVERSITÄT BONN

Property 2 => For d=512, $P(|n^T z| > 5.0) < 1e^{-6}$

=> Random samples $z \sim N(0, I_d)$ are most likely to be located close enough to the given hyperplane (within 5 unit-length).

=> ***Assumption: For any binary semantic there exists a hyperplane in the latent space serving as a decision boundary.***

Property 2 *Given* $\mathbf{n} \in \mathbb{R}^d$ *with* $\mathbf{n}^T\mathbf{n} = 1$, *which defines a hyperplane, and a multivariate random variable* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, *we have* $\mathrm{P}(|\mathbf{n}^T\mathbf{z}| \leq 2\alpha\sqrt{\frac{d}{d-2}}) \geq (1 - 3e^{-cd})(1 - \frac{2}{\alpha}e^{-\alpha^2/2})$ *for any* $\alpha \geq 1$ *and* $d \geq 4$. *Here* $\mathrm{P}(\cdot)$ *stands for probability and* c *is a fixed positive constant.*

Multiple Semantics

Single semantic case: $f_S(g(z)) = \lambda n^T z$

$$Def.: \; s \equiv f_S(g(z)) = \Lambda N^T z$$

$$s = [s_1, ..., s_m]^T \; - \; semantic\ scores$$

$$\Lambda = diag(\lambda_1, ..., \lambda_m) \; - \; diagonal\ matrix\ of\ linear\ coefficients$$

$$N = [n_1, ..., n_m] \; - \; separation\ boundaries$$

$$i.e.\ s = [\lambda_1 {n_1}^T z, ..., \lambda_m {n_m}^T z]^T$$

$$As \;\; z \sim N(0, I_d) \;\Rightarrow\; \mu_s = E(\Lambda N^T z) = \Lambda N^T E(z) = 0$$

$$and\ \Sigma_s = E((s - \mu_s)(s - \mu_s)^T) = E(\Lambda N^T z z^T N \Lambda^T) = \Lambda N^T E(zz^T) N \Lambda^T = \boxed{\Lambda N^T N \Lambda^T}.$$

$$s_i\ and\ s_j\ are\ disentangled \;\Leftrightarrow\; {n_i}^T n_j = 0$$

Manipulation in Latent Space

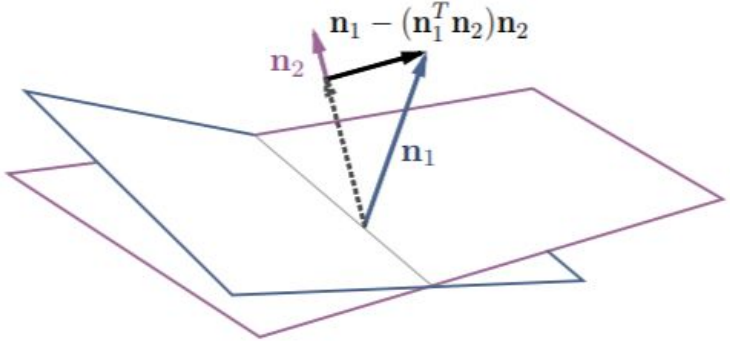
UNIVERSITÄT BONN

Single Attribute Manipulation

s_i and s_j are disentangled <=> $n_i^T n_j = 0$

$$z_{edit} = z + \alpha n \Rightarrow f(g(z_{edit})) = \lambda d(n,\ z + \alpha n) = \lambda d(n, z) + \lambda\alpha = f(g(z)) + \lambda\alpha$$

- $\alpha > 0$ => enhanced semantic
- $\alpha < 0$ => decreased semantic



Conditional Manipulation

- One condition => $n_{new} = n_1 - (n_1^T n_2) n_2$.
- More than one attribute to be conditioned on => subtract the projection from the primal direction onto the plane constructed by all conditioned directions (normals).

$$E.g.\ 2-cond.\ case:\ n_{new} = p - \alpha c_1 - \beta c_2,\ with\ \alpha = \frac{(pc_1^T - pc_2^T * c_1 c_2^T)}{(1-(c_1 c_2^T)^2)}\ and\ \beta = \frac{(pc_2^T - pc_1^T * c_1 c_2^T)}{(1-(c_1 c_2^T)^2)}$$

(p – primary direction, c_1/c_2 – conditional directions).

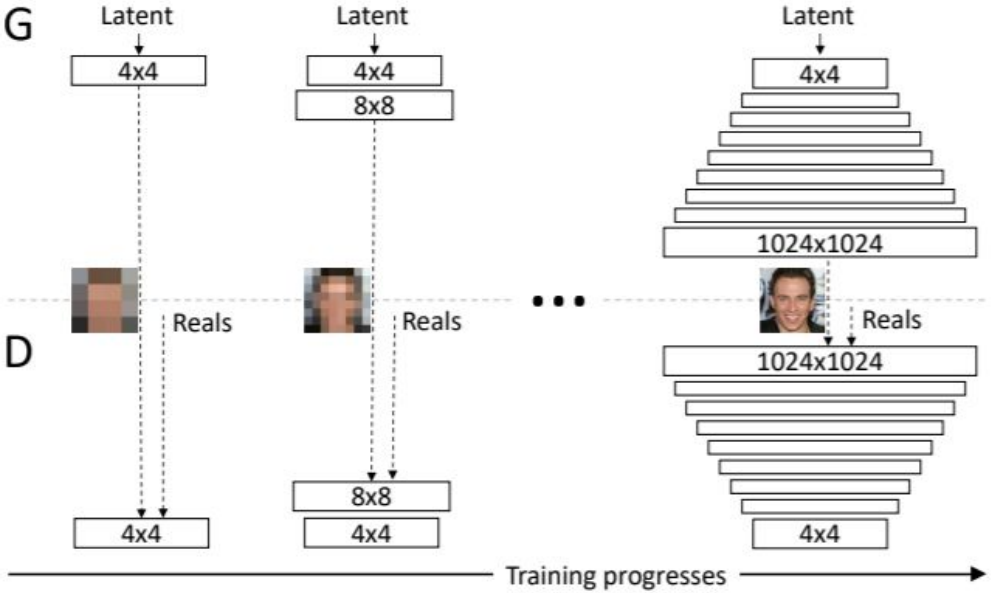
Implementation Details

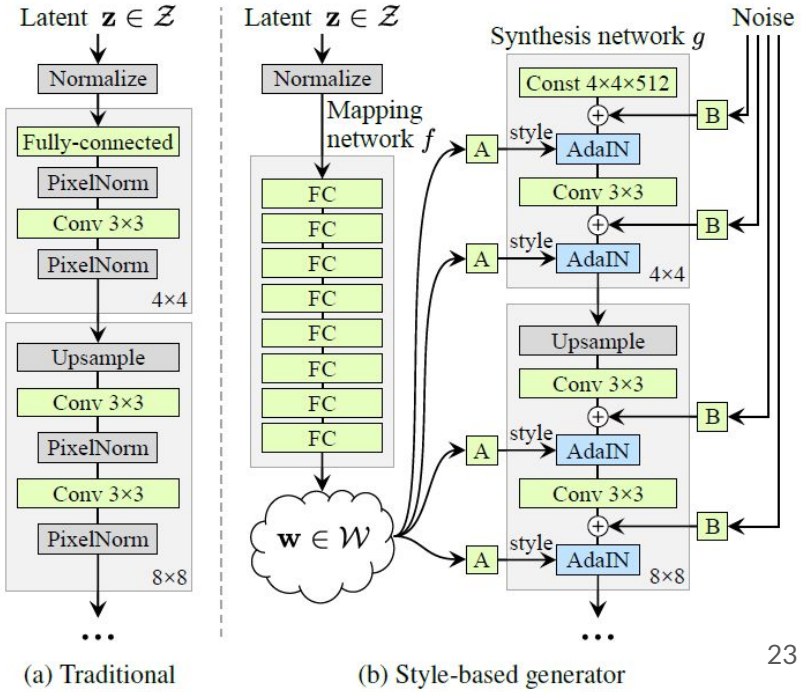
1. Synthesize ***500K*** images with randomly sampled latent codes.
2. Assign semantic labels to synthesized images with a pretrained attribute prediction model (***ResNet-50*** network trained on ***CelebA*** dataset, using multi-task loss).
3. Train 5 linear classifiers (***SVM***) in 512*d* latent space for each facial attribute (***gender, age, pose, eyeglasses, smile***) by using assigned semantic labels as ground truth.

- *For each attribute, use 10K samples with highest scores and 10K with lowest scores as candidates (to eliminate ambiguous cases).*
- *Use 70% of the candidates as training data (7K+7K) and 30% for validation (3K+3K) as well as the remaining entire dataset for testing (480K).*

Experiments

PGGAN [11] and StyleGAN [12]



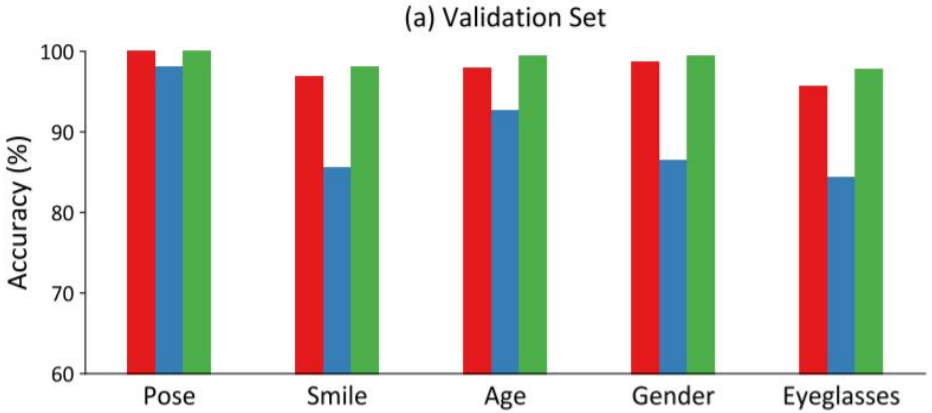


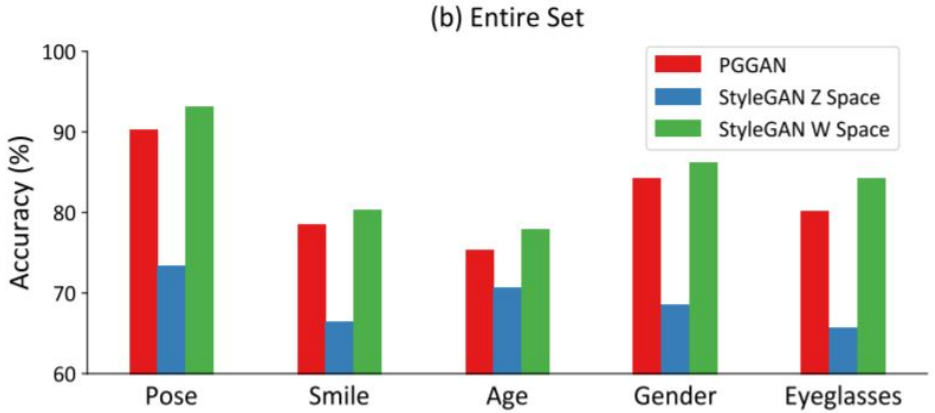
Experiments

UNIVERSITÄT BONN

Separability of Latent Space

- All learnt linear boundaries of both **PGGAN** *Z* space and **StyleGAN** *W* space show over **95%** (**75%**) accuracy on validation set (entire set).
- **StyleGAN** *W* space shows much stronger separability than *Z* space of both **StyleGAN** and **PGGAN**.

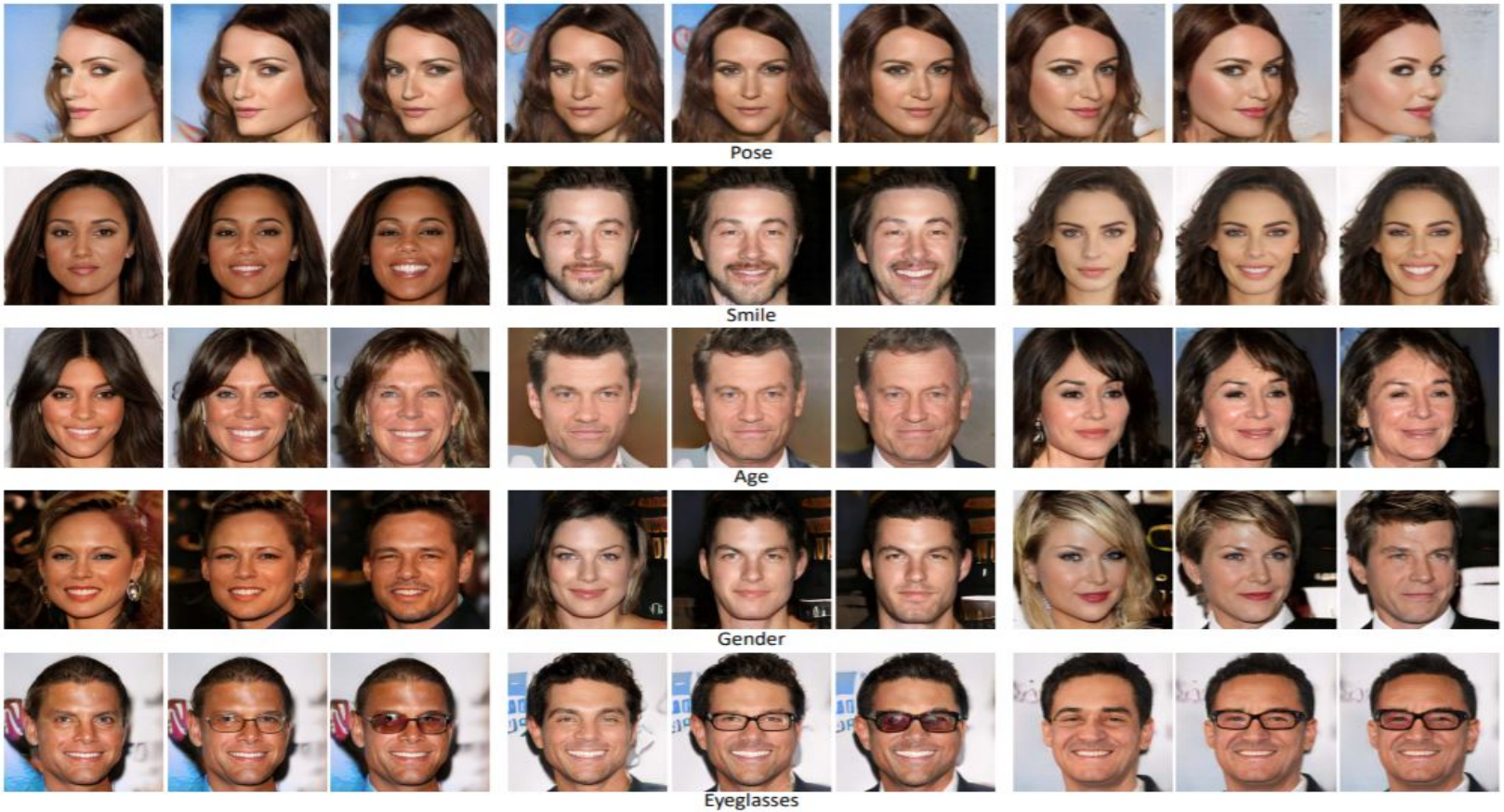




Experiments

UNIVERSITÄT BONN

Single Attribute Manipulation (PGGAN)

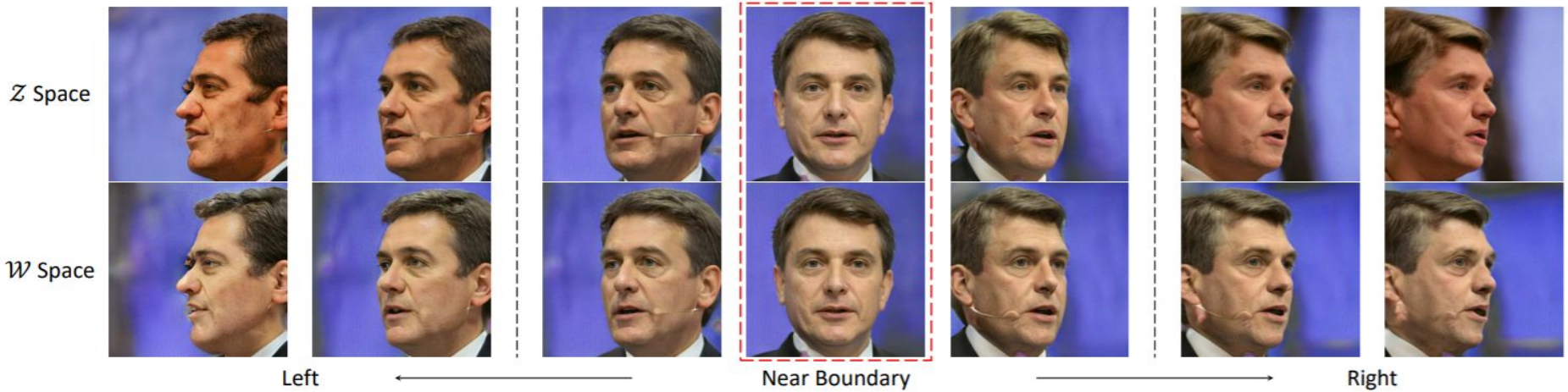


Experiments

UNIVERSITÄT BONN

Single Attribute Manipulation (StyleGAN)

- *StyleGAN* learns from more diverse dataset *FF-HQ.*
- *StyleGAN* can even ***generate children*** when making people younger.
- *StyleGAN* is even able to produce faces with ***extreme poses***.



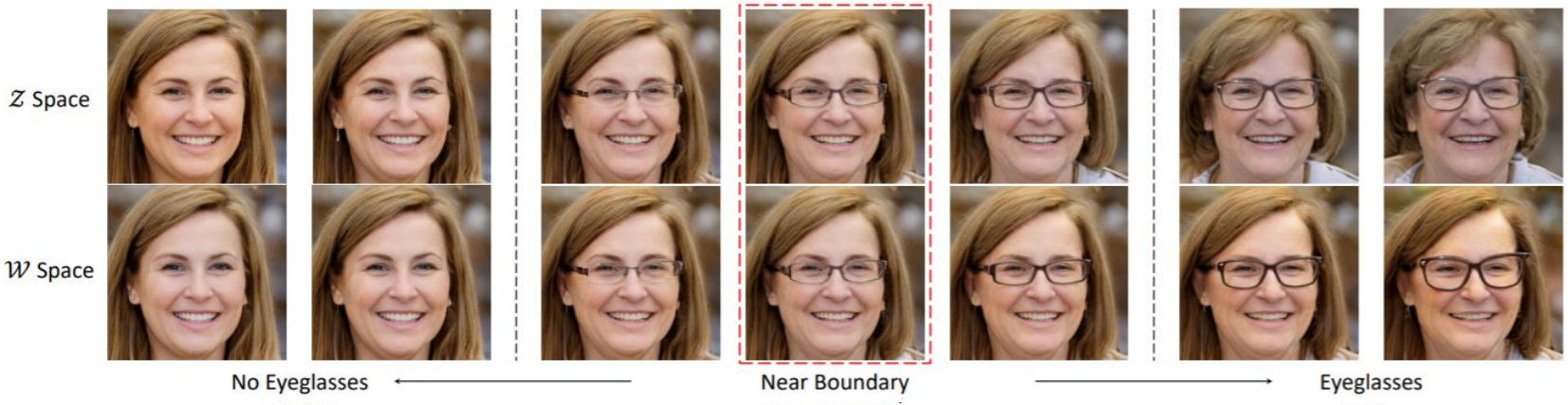
Experiments

UNIVERSITÄT BONN

Single Attribute Manipulation (StyleGAN)

For long-distance manipulation, *W* space outperforms *Z* space, where:

- People may ***take off glasses*** when moving along the ***gender*** direction.
- People become ***less feminine*** when moving along ***smile reduction*** direction.
- People tend to become ***older*** when we ***add glasses***.



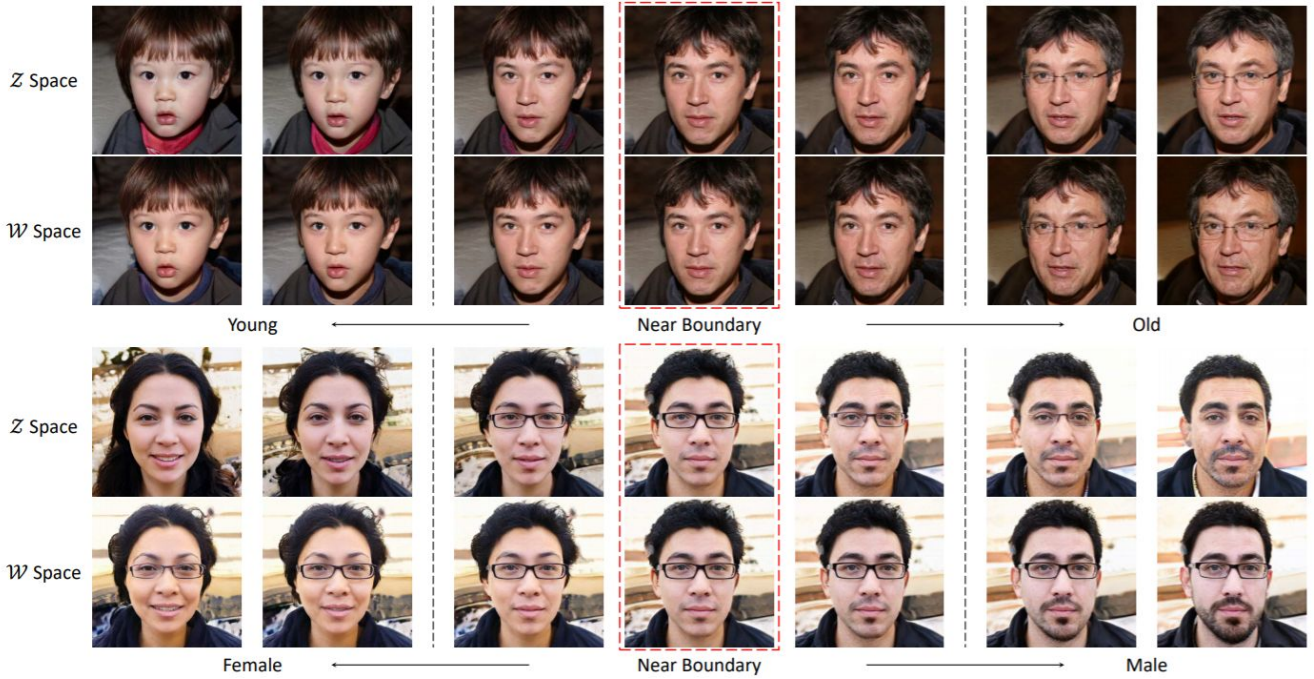
Experiments

UNIVERSITÄT BONN

Single Attribute Manipulation (StyleGAN)

Some attributes are ***correlated*** in both *Z* and *W* spaces.

- People are wearing ***eyeglasses*** when turning ***older***.
- People tend to become ***happier*** when ***feminized***.



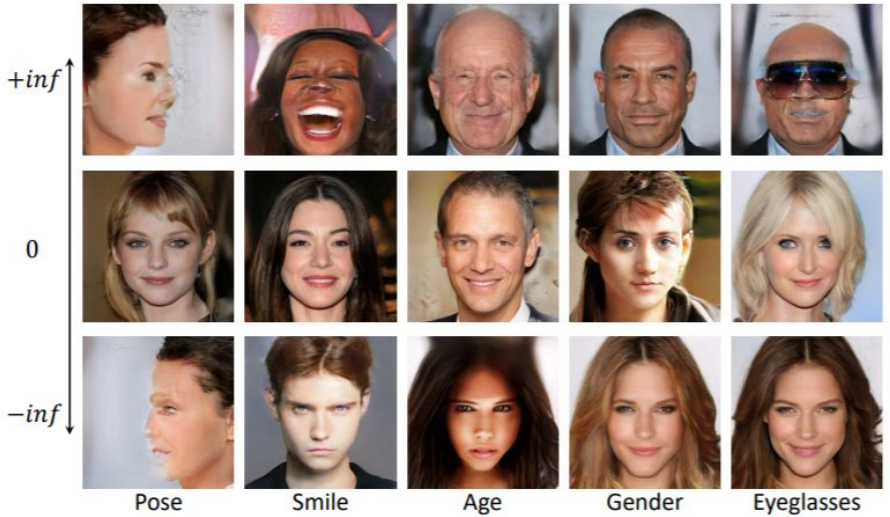
Experiments

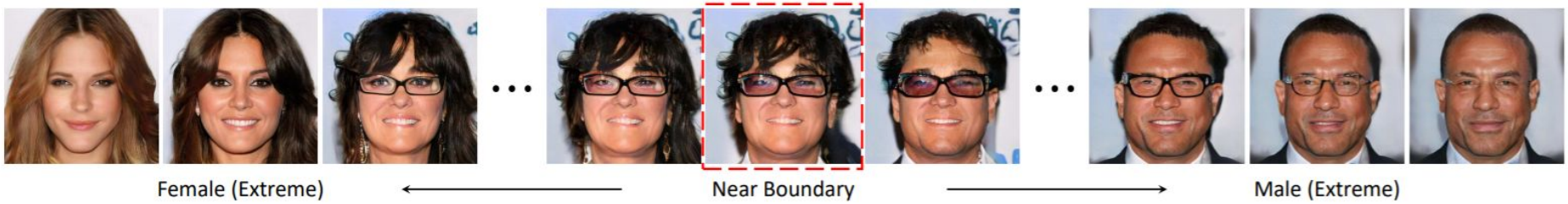
UNIVERSITÄT BONN

Distance Effect of Semantic Subspace

- Samples suffer from severe changes in appearance and finally tend to become the extreme cases when moving latent code too far from the boundary.
- Beyond a certain region (*5.0*), the results no longer look like the original person.

Property 2 => For d=512, $P(|n^T z| > 5.0) < 1e^{-6}$





Experiments

UNIVERSITÄT BONN

Artifacts Correction

- Manually labeled *4K* bad synthesis and trained a linear *SVM* for separation boundary.
- GAN also encodes such information in latent space.
- Fix the artifacts by moving the latent code towards the positive “quality” direction.

Experiments

Correlation between Attributes

- **Semantic boundary correlation**: Given 2 semantics, cosine similarity between the corresponding latent boundary normals.

$$cos(n_1, n_2) = {n_1}^T n_2$$

	Pose	Smile	Age	Gender	Glasses
Pose	1.00	-0.04	-0.06	-0.05	-0.04
Smile		1.00	0.04	-0.10	-0.05
Age			1.00	0.49	0.38
Gender				1.00	0.52
Glasses					1.00

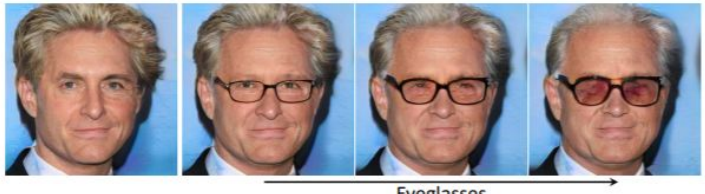
- **Attribute correlation of synthesized data**: Pearson correlation coefficient between 2 attributes *A* and *B* on *500K synthesized* data.

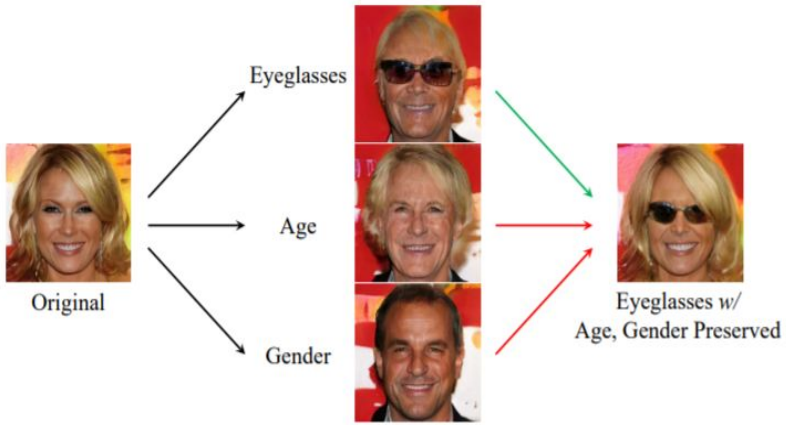
$$\rho_{A,B} = \frac{cov(A,B)}{\sigma_A \sigma_B}$$

	Pose	Smile	Age	Gender	Glasses
Pose	1.00	-0.01	-0.01	-0.02	0.00
Smile		1.00	0.02	-0.08	-0.01
Age			1.00	0.42	0.35
Gender				1.00	0.47
Glasses					1.00

Experiments

Conditional Manipulation (PGGAN)





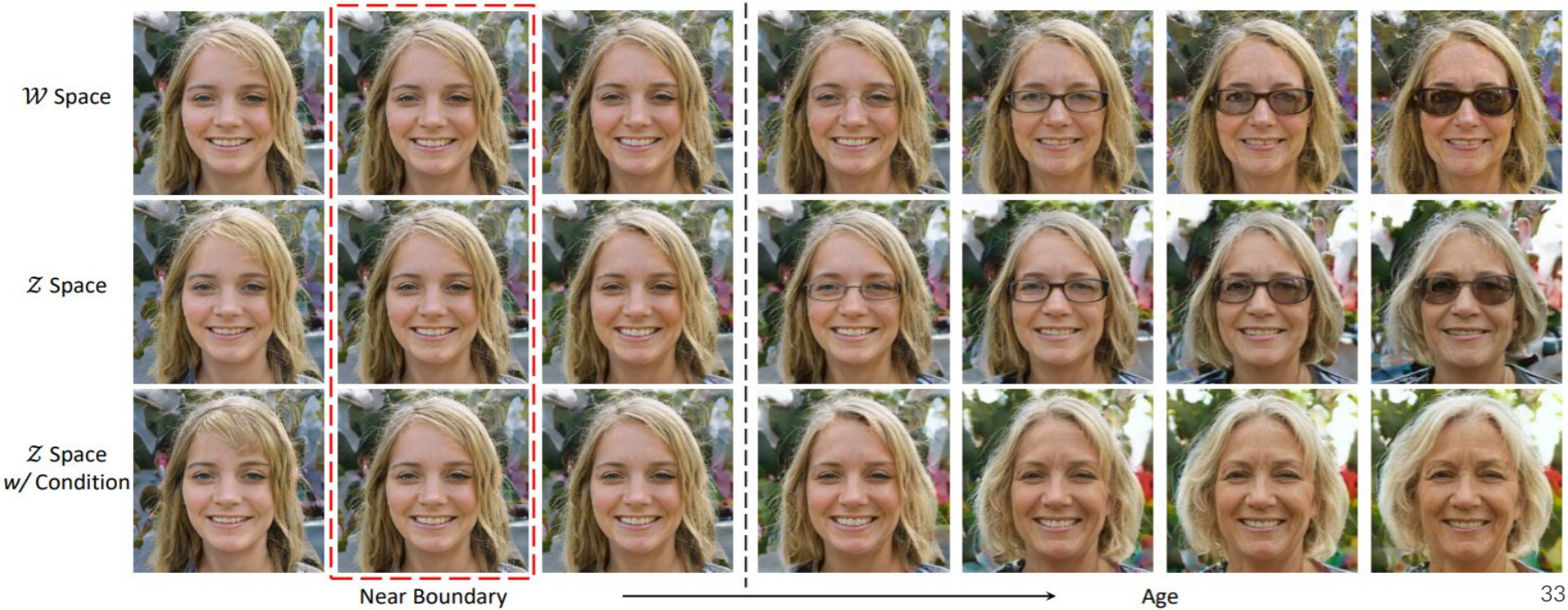
- Results tend to become male when being edited to get old => conditional manipulation to preserve "*gender*" attribute.
- "Eyeglasses" is entangled with both "age" and "gender" => conditional manipulation based on two attributes.

Experiments

UNIVERSITÄT BONN

Conditional Manipulation (StyleGAN)

"*Age*" w/ "*eyeglasses*" manipulation is not applicable in *W* space because of ***"over-disentanglement" problem***.



Real Image Manipulation

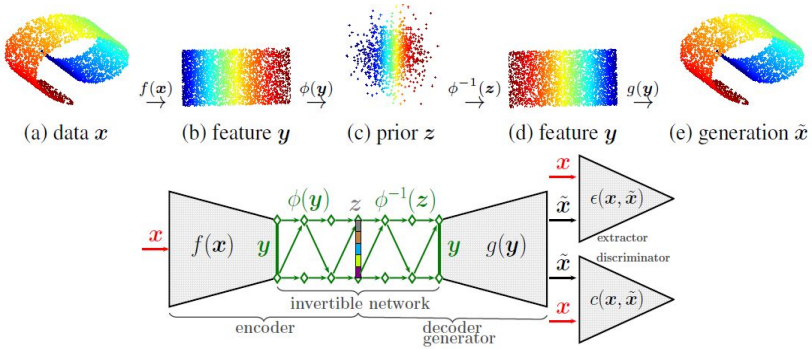
$$z^* = arg \min_{z \in Z} L(G(z), x^R)$$

$$G^{-1}(x^R) = z^* \Rightarrow z_{edit} = z^* + \alpha n \Rightarrow G(z_{edit})$$

- **Optimization - based inversion method:** Gradient descent w.r.t latent code z to minimize reconstruction loss of the given image [13].
- **Encoder - based inversion method**: Train an extra encoder for a fixed GAN to perform the inversion task [14].

$$\theta^*_P = arg \min_{\theta_P} \sum_n L(G(P(x_n^R; \theta_P)), x_n^R)$$

- **Encoder-decoder training** (use encoder as an inverter) [15].

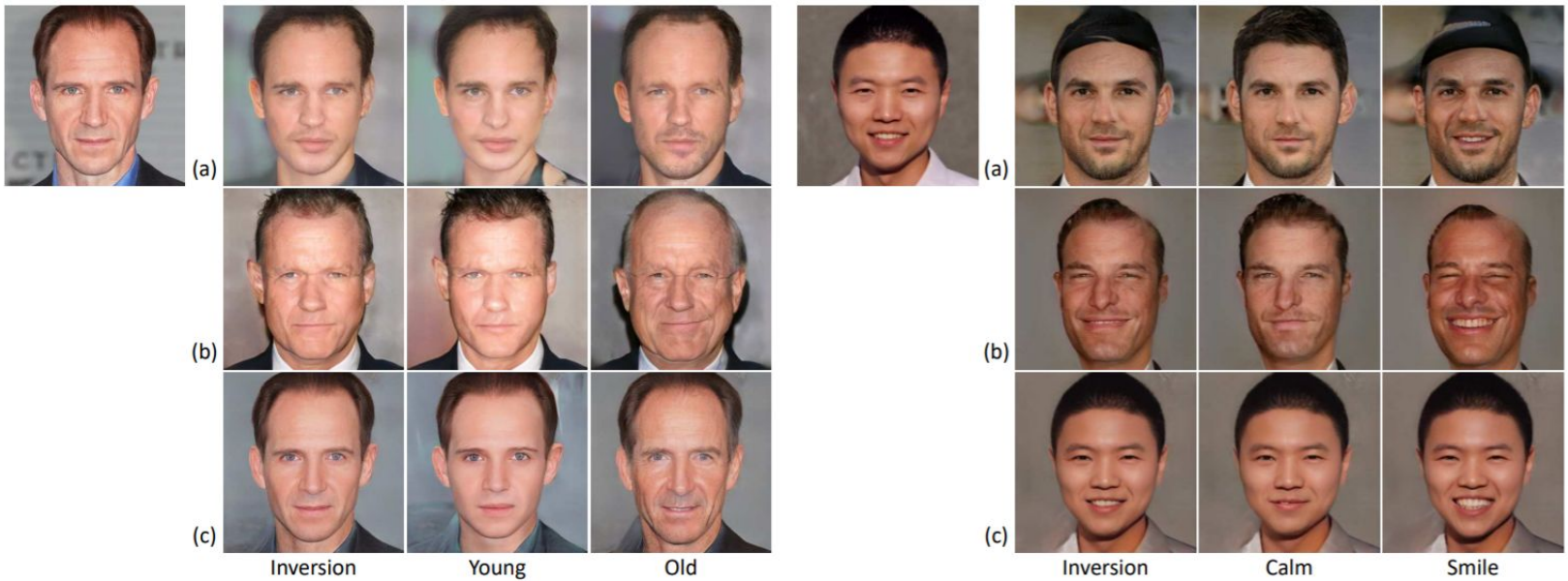


Experiments

UNIVERSITÄT BONN

Real Image Manipulation

(a) PGGAN with optimization-based inversion method
(b) PGGAN with encoder-based inversion method
(c) StyleGAN with optimization-based inversion method

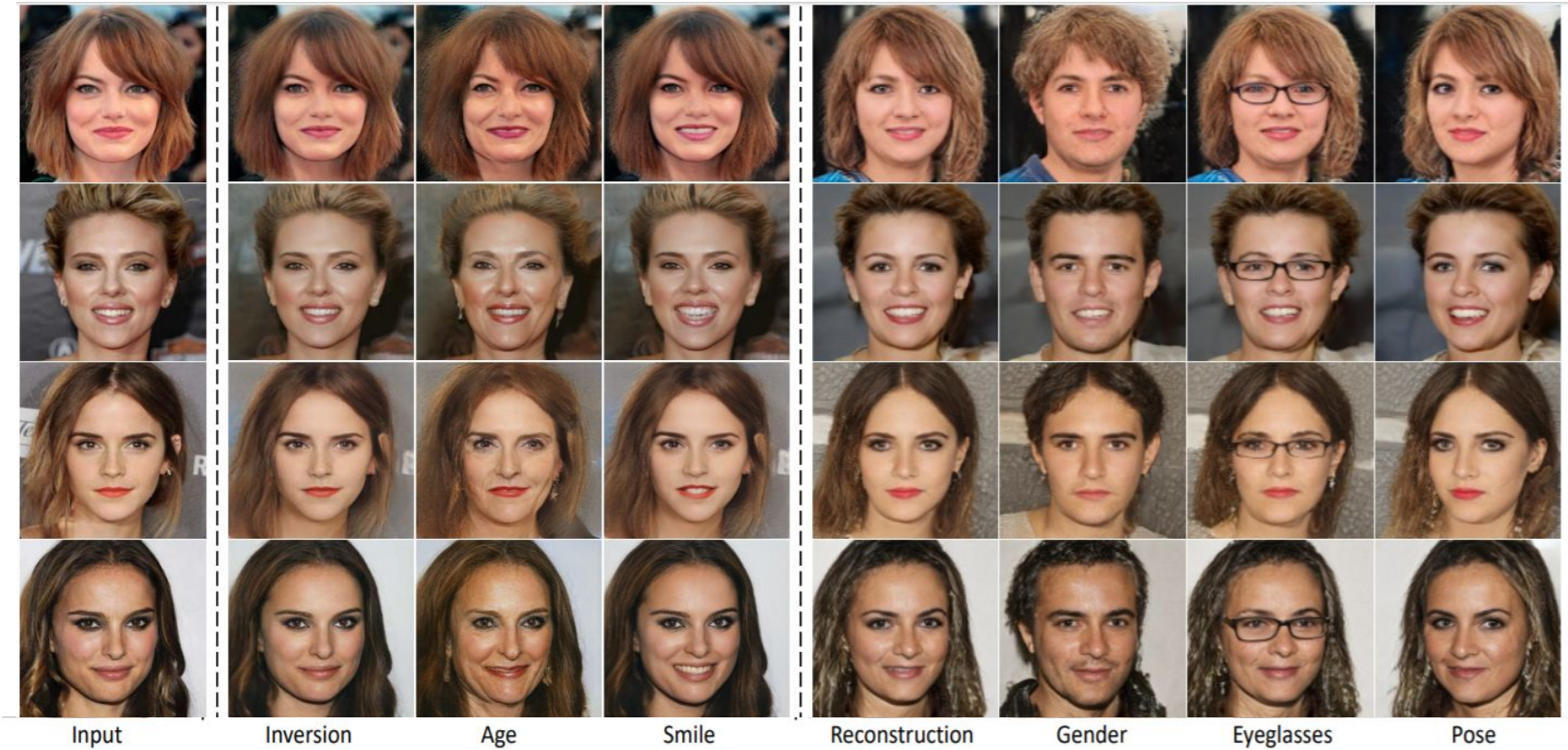


Experiments

UNIVERSITÄT BONN

Real Image Manipulation

StyleGAN with Optimization-Based Inversion Method vs LIA

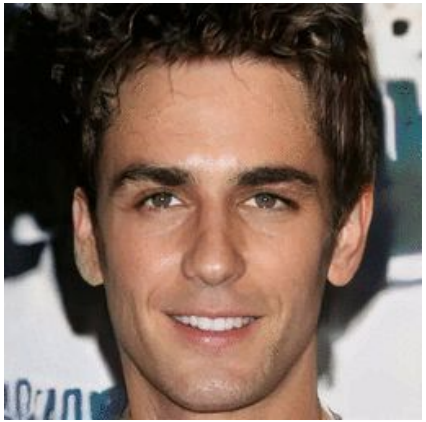


InterFaceGAN/Continuous Semantic Manipulation Results

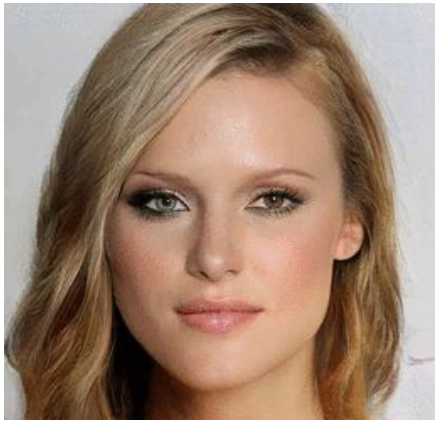
POSE

AGE

SMILE



EYEGLASSES



GENDER

ARTIFACTS

(GIF from here)

Conclusion

- InterFaceGAN successfully employs some ideas of previous works on latent space manipulation for controlling the generation process of various fixed GAN models trained on facial dataset.
- InterFaceGAN introduces conditional manipulation for attribute disentanglement.
- InterFaceGAN can be applied to real image editing by using various GAN inversion methods or encoder-envolved models.
- InterFaceGAN can even be used for fixing some generation artifacts.

References

[1] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Visualizing and understanding generative adversarial networks. In ICLR, 2019.

[2] Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick van der Smagt. Metrics for deep generative models. In AISTAT, 2018.

[3] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In CVPR Workshop, 2018.

[4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In ICLR, 2016.

[5] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. Optimizing the latent space of generative networks. In ICML, 2018.

[6] Ali Jahanian, Lucy Chai, and Phillip Isola. On the ”steerability” of generative adversarial networks. In ICLR, 2020.

References

[7] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In ICCV, 2019.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In NeurIPS, 2016.

[9] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In NeurIPS, 2017.

[10] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In ECCV, 2018.

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In ICLR, 2018.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In CVPR, 2019.

References

[13] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. In NeurIPS, 2018.

[14] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In ECCV, 2016.

[15] Jiapeng Zhu, Deli Zhao, and Bo Zhang. Lia: Latently invertible autoencoder with adversarial learning. arXiv preprint arXiv:1906.08090, 2019.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NeurIPS, 2014.

UNIVERSITÄT BONN

Thank you!