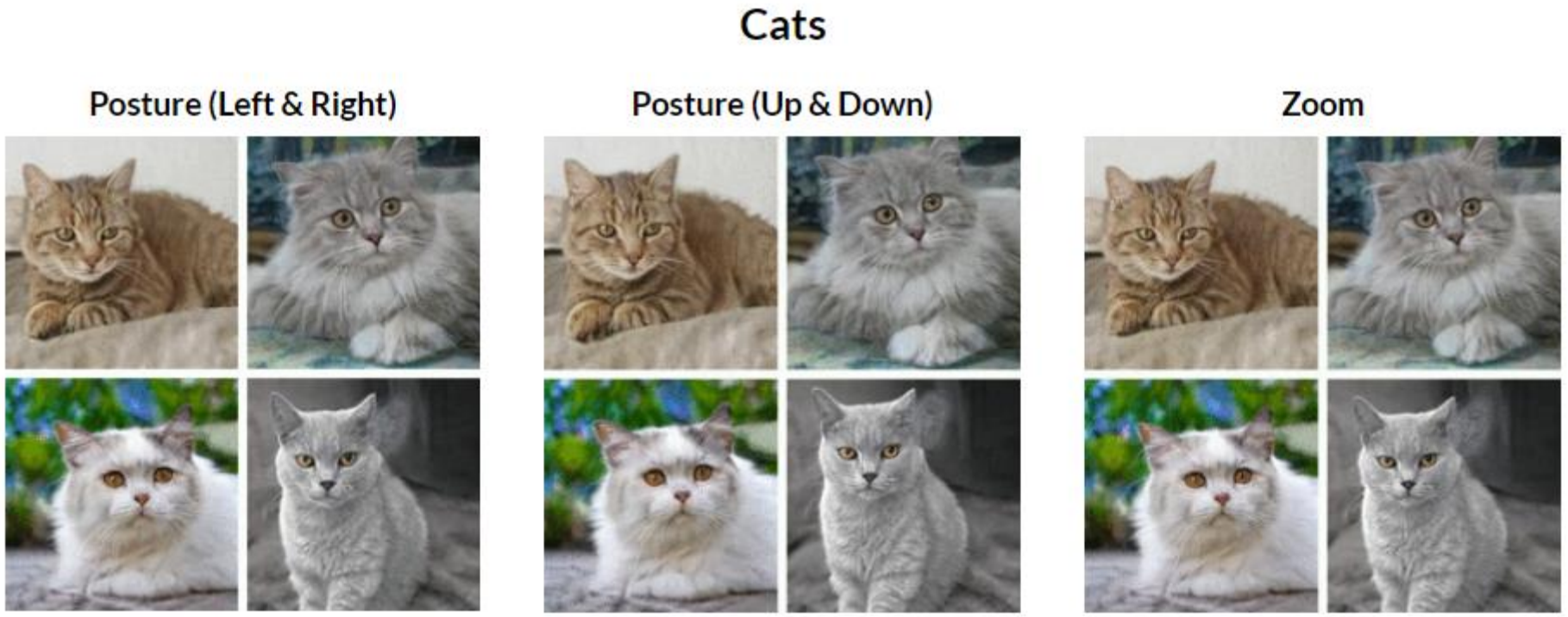
CLOSED-FORM FACTORIZATION OF LATENT SEMANTICS IN GANS

Yujun Shen, Bolei Zhou

2020, ArXiv

Presenter: Jeonghoon Park

Introduction



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

Existing Methods

[Supervised learning-based]

sampling latent codes -> synthesizing images -> annotating them with pre-defined labels (pre-trained semantic predictors , simple statistical information) -> learning a separation boundary with labeled samples

*Cons:
Limited by available semantic predictors
Unstable: may lead different boundary search by sampling
Rare attributes: time-consuming(sampling) and biased

Existing Methods

[Unsupervised learning-based]

-Voynov and Babenko introduces the regularizer proposed by InfoGAN into the semantic search process but requires a pre-defined number of semantics.

-Härkönen et al. proposes to skip the labelling process and perform PCA on the sampled data to find primary directions in the latent space.

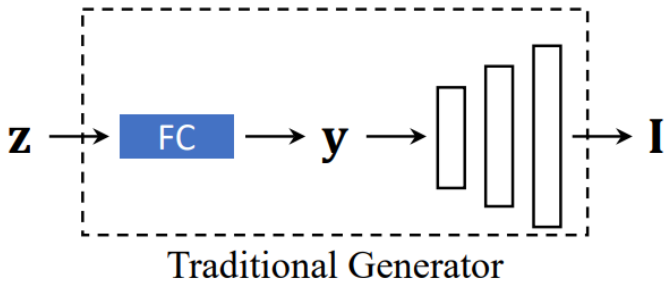
*Cons: still sampling data

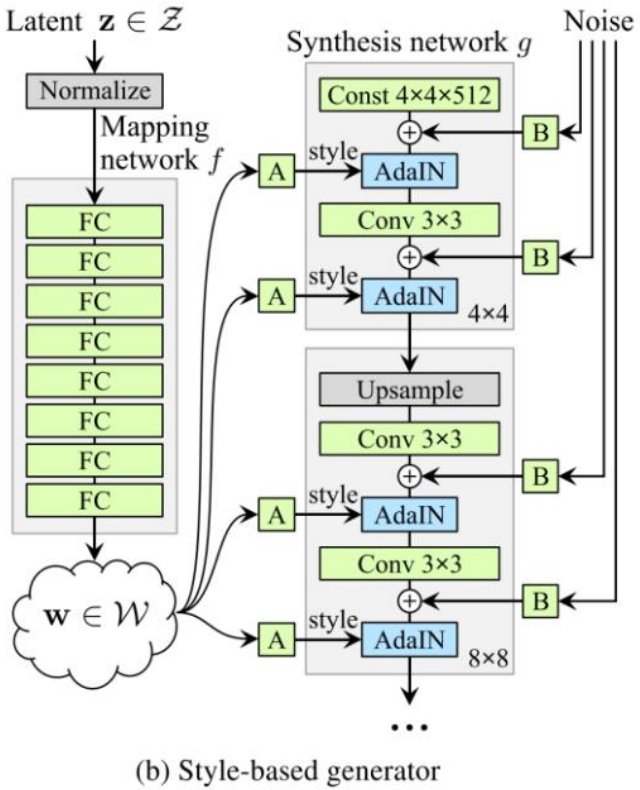
[Proposed Method: SeFa, latent Semantic Factorization in GANs]

Differently we propose a closed-form factorization method for latent semantic interpretation
Instead of ustilizing the synthesized samples, directly look into the generation mechanishm of GANs

- ✓ Sampling X
- ✓ Semantic Predictors/Labels X
- ✓ Unsupervised

Methodology: Problem Statement



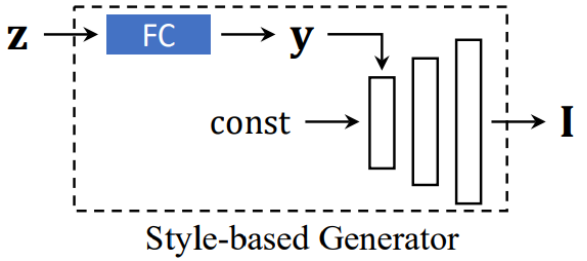


(b) Style-based generator

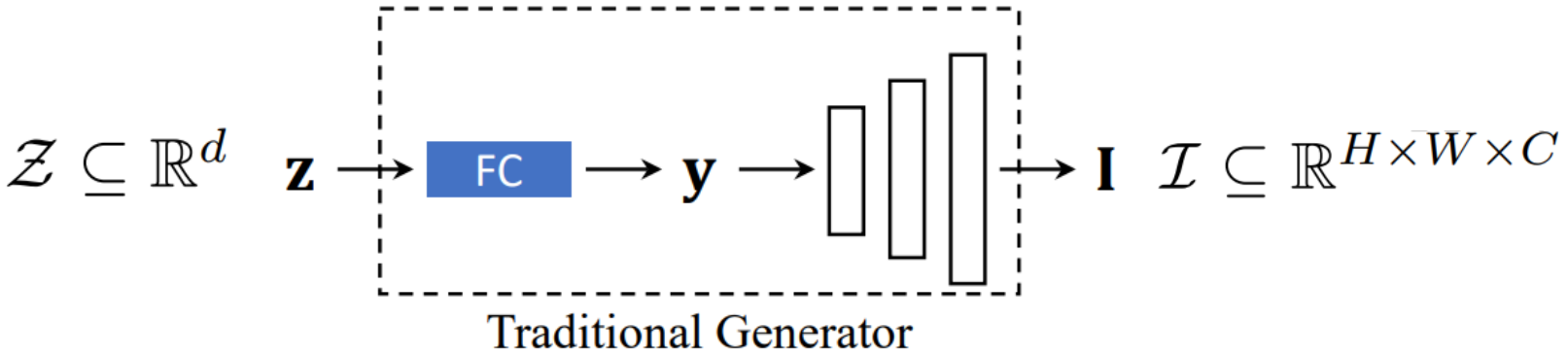
- ✓ Traditional generator: latent code -> initial feature map for convolutional layers
- ✓ Style-based generator: latent code -> layer-wise style codes (each layer has its own FC)

https://blog.lunit.io/2019/02/25/a-style-based-generator-architecture-for-generative-adversarial-networks/

Methodology: Problem Statement



$$G(\cdot)$$



$$\mathbf{I} = G'(FC(\mathbf{z})) \triangleq G'(\mathbf{y})$$

remaining part of the generator except the FC at the beginning

Methodology: SeFa

we would like to find n, instead of z, that can cause the shift of I to the most extent.

Assumption: large change of y will lead to a large content variation of I.

$\Rightarrow$ Goal: finding the directions n that can cause the significant change of y.

$$\mathbf{z}' = \mathbf{z} + \alpha\mathbf{n},\ \mathbf{n} \in \mathbb{R}^d$$

$$\mathbf{I} = G'(FC(\mathbf{z})) \triangleq G'(\mathbf{y}),$$

$$\mathbf{y} = FC(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$$

$$\Delta\mathbf{y} = FC(\mathbf{z}') - FC(\mathbf{z}) = (\mathbf{A}(\mathbf{z} + \alpha\mathbf{n}) + \mathbf{b}) - (\mathbf{A}\mathbf{z} + \mathbf{b}) = \alpha\mathbf{A}\mathbf{n}.$$

Methodology: SeFa

transformation in the fully-connected layer:"semantic selector"
For whatever semantic **n** encoded in the latent space, it has to go through this selector to be reflected in the final synthesis I.

$$\mathbf{n}^* = \underset{\{\mathbf{n}\in\mathbb{R}^d:\ \mathbf{n}^T\mathbf{n}=1\}}{\arg\max} ||\mathbf{An}||_2^2,$$

$$\mathbf{N}^* = \underset{\{\mathbf{N}\in\mathbb{R}^{d\times k}:\ \mathbf{n}_i^T\mathbf{n}_i=1\ \forall i=1,\cdots,k\}}{\arg\max} \sum_{i=1}^{k} ||\mathbf{An}_i||_2^2,$$

Lagrange Multiplier

$$\mathbf{N}^* = \underset{\mathbf{N}\in\mathbb{R}^{d\times k}}{\arg\max} \sum_{i=1}^{k} ||\mathbf{An}_i||_2^2 - \sum_{i=1}^{k} \lambda_i(\mathbf{n}_i^T\mathbf{n}_i - 1)$$

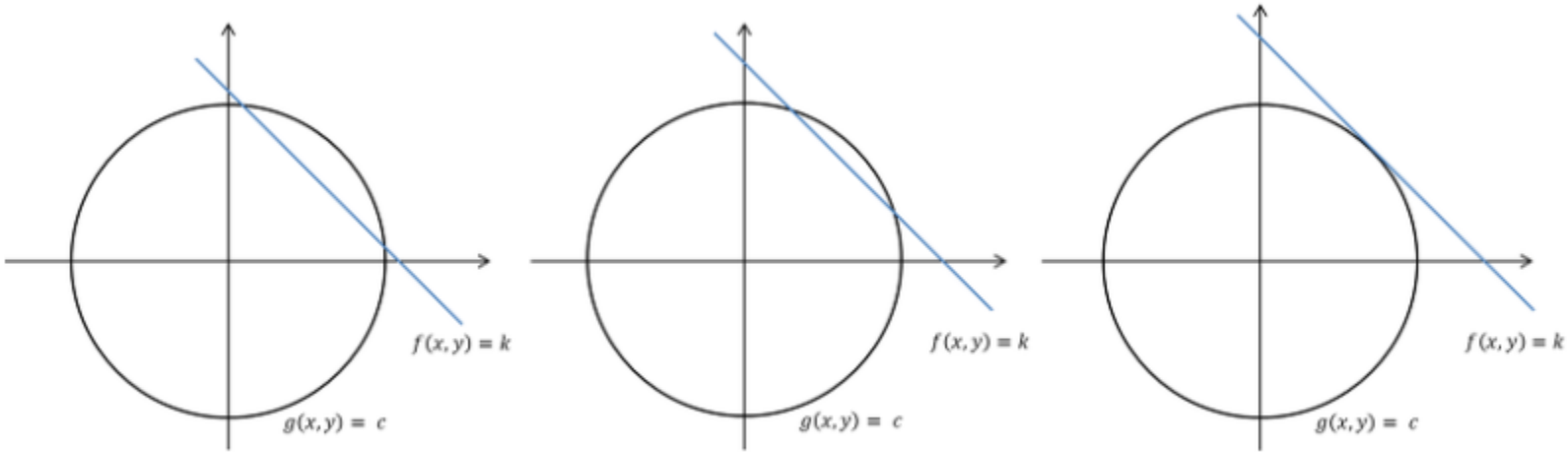
$$= \underset{\mathbf{N}\in\mathbb{R}^{d\times k}}{\arg\max} \sum_{i=1}^{k} (\mathbf{n}_i^T\mathbf{A}^T\mathbf{An}_i - \lambda_i\mathbf{n}_i^T\mathbf{n}_i + \lambda_i).$$

$$2\mathbf{A}^T\mathbf{An}_i - 2\lambda_i\mathbf{n}_i = 0.$$

All possible solutions should be the eigenvectors of the matrix A^TA.

https://untitledtblog.tistory.com/96

Methodology: SeFa



[그림 1] 제약 조건 $g(x, y) = c$를 만족하는 $f(x, y)$의 최댓값 문제에 대한 기하학적 표현

$$\nabla f = \lambda \nabla g$$

$$L(x, y, \lambda) = f(x, y) - \lambda(g(x, y) - c)$$

Methodology: SeFa

[Property of the Discovered Semantics]

all semantic directions are the eigenvectors of matrix A^TA -> **eigen decomposition**

$$\mathbf{A}^T\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$$

Λ is a diagonal matrix, indicating the eigenvalues, while Q is an orthogonal matrix, containing all the eigenvectors.

Obviously, each $\mathbf{n}_i$ is a column of Q

$$\mathbf{N}^T\mathbf{N} = \mathbf{I}_k$$

all semantic directions found by our algorithm are orthogonal to each other in the latent space

$$\Delta\mathbf{y}_i^T\Delta\mathbf{y}_j = \mathbf{n}_i^T\mathbf{A}^T\mathbf{A}\mathbf{n}_j = \mathbf{n}_i^T(\lambda_j\mathbf{n}_j) = 0 \ \ \forall i \neq j,$$

variations of the outputs of FC derived by different directions are also orthogonal to each other.

Experiments

[Models / Datasets]

- ✓ PGGAN, StyleGAN, BigGAN, and StyleGAN2
- ✓ diverse datasets, including human faces (CelebA-HQ and FF-HQ), anime faces, scenes and objects (LSUN), streetscapes, and ImageNet.

[Implementation Details]

PGGAN: the very first FC layer
StyleGAN/StyleGAN2(style-based generator), we choose the style mapping layers of each convolutional block.

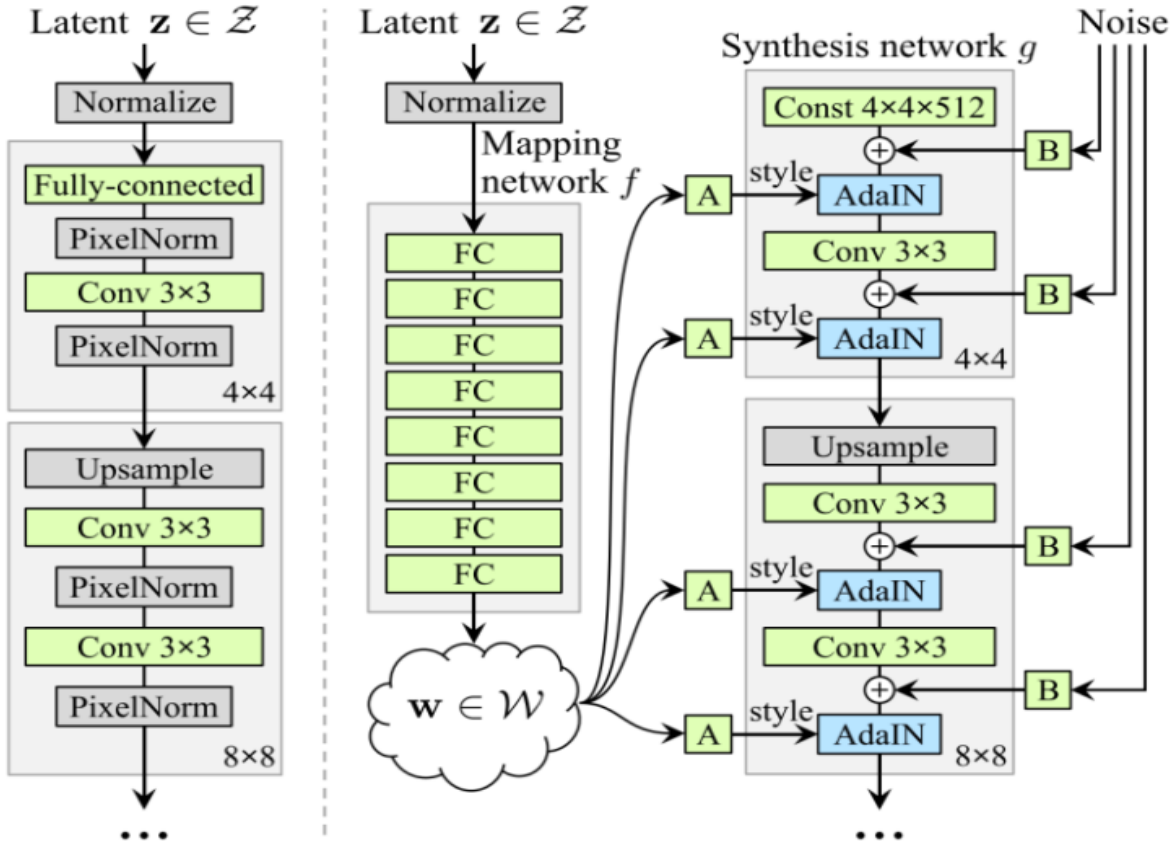
Concretely, **we concatenate the weight matrices from all target layers along the first axis, resulting in a larger matrix.**

Note that, besides the latent space Z, StyleGAN introduces a more disentangled space W. We do experiments on W space for StyleGAN and StyleGAN2 since $w \in W$, instead of $z \in Z$, is the code fed into the generator.

-normalize each row of the matrix: focus on the direction

Experiments

[Implementation Details]



(a) Traditional (b) Style-based generator

Experiments

[Comparison with Unsupervised Baselines]
Post-Annotation of the Discovered Semantics

- ✓ align them with human perception by assigning them with interpretable meanings.
- ✓ use the semantic directions found by existing supervised method, i.e., InterFaceGAN as the "ground-truth".
- ✓ compare all eigen directions from our algorithm with the "ground-truth" direction and choose the one with the **smallest cosine distance**

Table 1: Comparison between the semantics identified by different methods, including performing PCA on a collection of sampled latent codes [10] and our *closed-form* solution. Semantics learned by InterFaceGAN [23] are used as the "ground-truth". "Dist." indicates the cosine distance (smaller is better), while "No." denotes the 0-based index of the eigenvector (smaller means primary direction).

		Pose		Gender		Age		Eyeglasses		Smile	
		Dist.	No.	Dist.	No.	Dist.	No.	Dist.	No.	Dist.	No.
PGGAN $\mathcal{Z}$ Space	PCA	0.90	23	0.83	1	0.88	1	0.86	1	0.86	10
	Ours	**0.04**	2	**0.16**	1	**0.45**	4	**0.52**	1	**0.69**	20
StyleGAN $\mathcal{W}$ Space	PCA	0.28	9	0.71	2	**0.73**	5	0.72	5	0.73	25
	Ours	**0.13**	1	**0.64**	3	0.80	6	**0.65**	6	**0.65**	37

Experiments

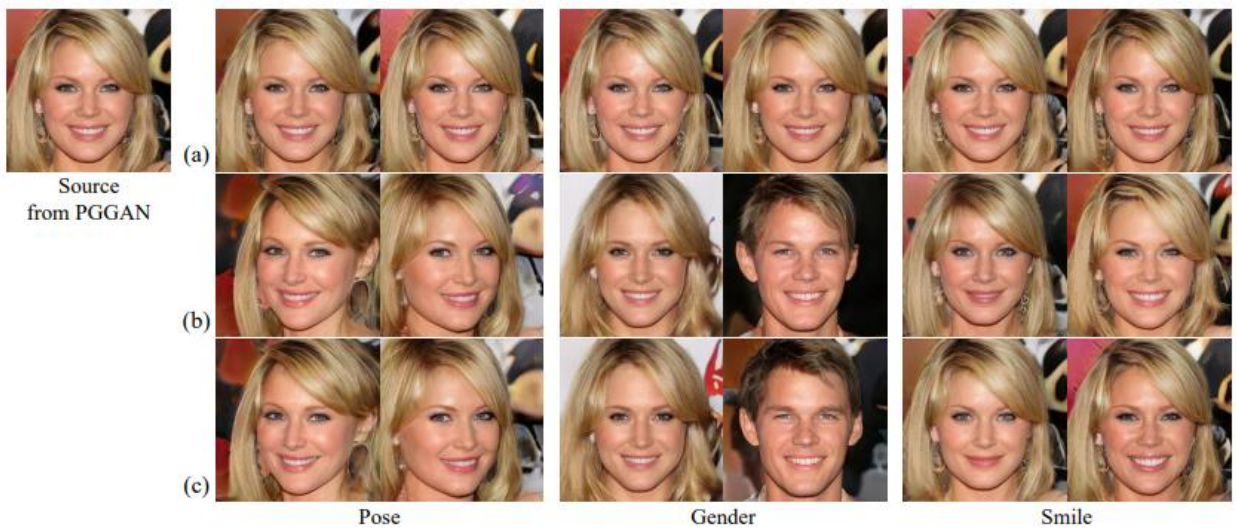
[Comparison with Unsupervised Baselines]
Comparison with Sampling-based Unsupervised Baseline
sample a collection of latent codes (500K) and then perform PCA on these samples to find principle directions

		Pose		Gender		Age		Eyeglasses		Smile	
		Dist.	No.	Dist.	No.	Dist.	No.	Dist.	No.	Dist.	No.
PGGAN $\mathcal{Z}$ Space	PCA	0.90	23	0.83	1	0.88	1	0.86	1	0.86	10
	Ours	**0.04**	2	**0.16**	1	**0.45**	4	**0.52**	1	**0.69**	20
StyleGAN $\mathcal{W}$ Space	PCA	0.28	9	0.71	2	**0.73**	5	0.72	5	0.73	25
	Ours	**0.13**	1	**0.64**	3	0.80	6	**0.65**	6	**0.65**	37

- ✓ directions from our method show a smaller distance to the "ground-truth"
  fail on the traditional generator since the latent codes are simply subject to a normal distribution.
  better results on the W space of StyleGAN
- ✓ SeFa: not rely on the sampled data and hence is more stable

Experiments

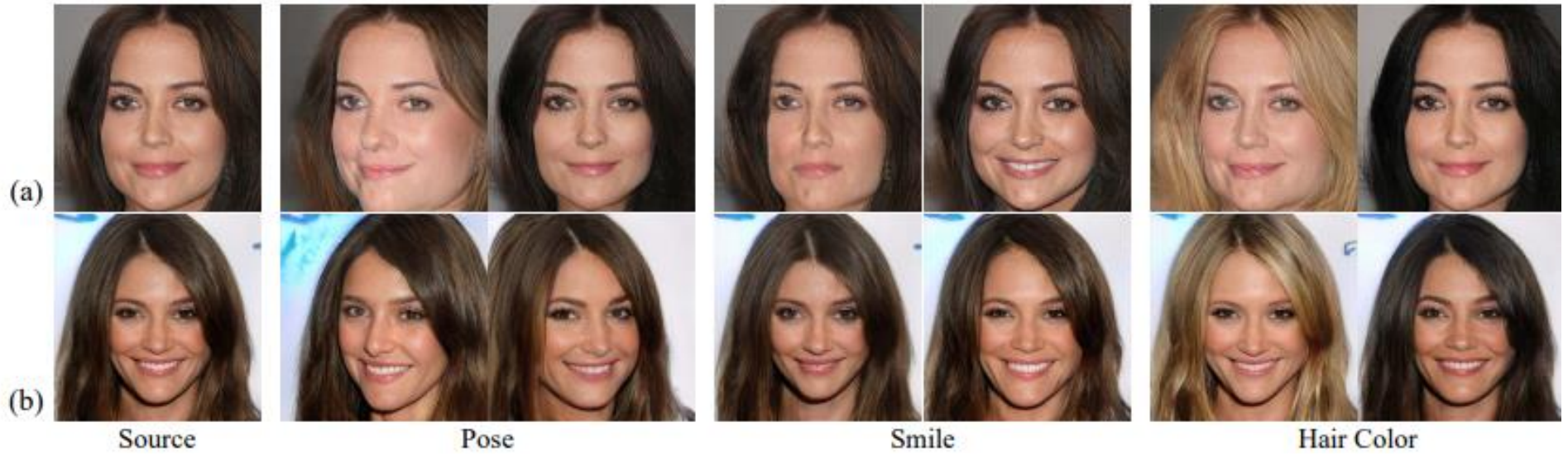
[Comparison with Unsupervised Baselines]
Comparison with Sampling-based Unsupervised Baseline



(a) Sampling-based unsupervised approach
(b) SeFa
(c) Supervised method(InterFaceGAN)

Experiments

[Comparison with Unsupervised Baselines]
Comparison with Learning-based Unsupervised Baseline



InfoGAN proposed to explicitly learn a factorized representation in an unsupervised manner.

(a) Info-PGGAN: add the regularizer to maximize the mutual information
requires knowing the number of factors before the training

(b) SeFa
identifies the semantics automatically learned by GANs
more accurate than the semantics learned with the additional regularizer

Experiments

[Comparison with Supervised Approach & Semantic Property Analysis]
Comparison with Learning-based Unsupervised Baseline

InterFaceGAN : state-of-the-art supervised method for latent semantic discovery

(a) how different semantics are disentangled from each other

(b) how diverse the identified semantics are.

Experiments

[Comparison with Supervised Approach & Semantic Property Analysis]
Disentanglement Comparison

re-scoring analysis to quantitatively evaluate the disentanglement between different semantics

randomly sample 2,000 images, then manipulate them along a certain semantic direction, and finally use pre-trained attribute predictors to check how the scores corresponding to different attributes vary in the manipulation process

Table 2: Re-scoring analysis on different methods by evaluating how the semantic scores change via modulating the latent codes. Each row shows the results by moving the latent codes towards a certain direction.

(a) InterFaceGAN [23], which is supervised.

	Pose	Gender	Age	Glasses	Smile
Pose	0.53	-0.06	-0.09	-0.01	0.05
Gender	-0.02	0.59	0.20	0.08	-0.07
Age	-0.03	0.35	0.50	0.08	-0.03
Glasses	-0.01	0.37	0.19	0.24	0.00
Smile	-0.01	-0.07	0.03	-0.01	0.60

(b) Our *closed-form* solution, which is unsupervised.

	Pose	Gender	Age	Glasses	Smile
Pose	0.51	-0.11	-0.07	0.02	0.06
Gender	0.02	0.55	0.46	0.09	-0.13
Age	-0.07	-0.25	0.34	0.10	0.10
Glasses	0.02	0.55	0.46	0.09	-0.13
Smile	0.03	-0.03	0.15	-0.16	0.42

✓ Similar disentanglement property as the supervised method (ex. Pose and smile / gender, age, glasses)

Experiments

[Comparison with Supervised Approach & Semantic Property Analysis]
Diversity Comparison

Figure 5: (a) Diverse semantics, which can *not* be identified by InterFaceGAN [23] due to the lack of semantic predictors. (b) Diverse hair styles, which can *not* be described as a binary attribute.

(a) more diverse semantics (not be identified by InterFaceGAN)
(b) binary attributes (InterFaceGAN) vs. more complex attributes (SeFa) (multiple eigen directions control the same attributes)

Conclusion

SeFa (Semantic Factorization)

- ✓ Transformation (FC layer) actually filters out some negligible directions in the latent space and highlights the directions that are critical for image synthesis.
- ✓ Fast and efficient(i.e., less than 1 second)
- ✓ identify interpretable dimensions more accurately and in a wider range (existing supervised methods and some very recent unsupervised baselines)
- ✓ edit the semantics of the synthesized image