Closed-Form Factorization of Latent Semantics in GANs

Yujun Shen, Bolei Zhou

CVPR 2021 Oral

STRUCT Group Seminar
Presenter: Wenjing Wang
2020.03.14

OUTLINE

- Authorship
- Background
- Proposed Method
- Experimental Results
- Conclusion

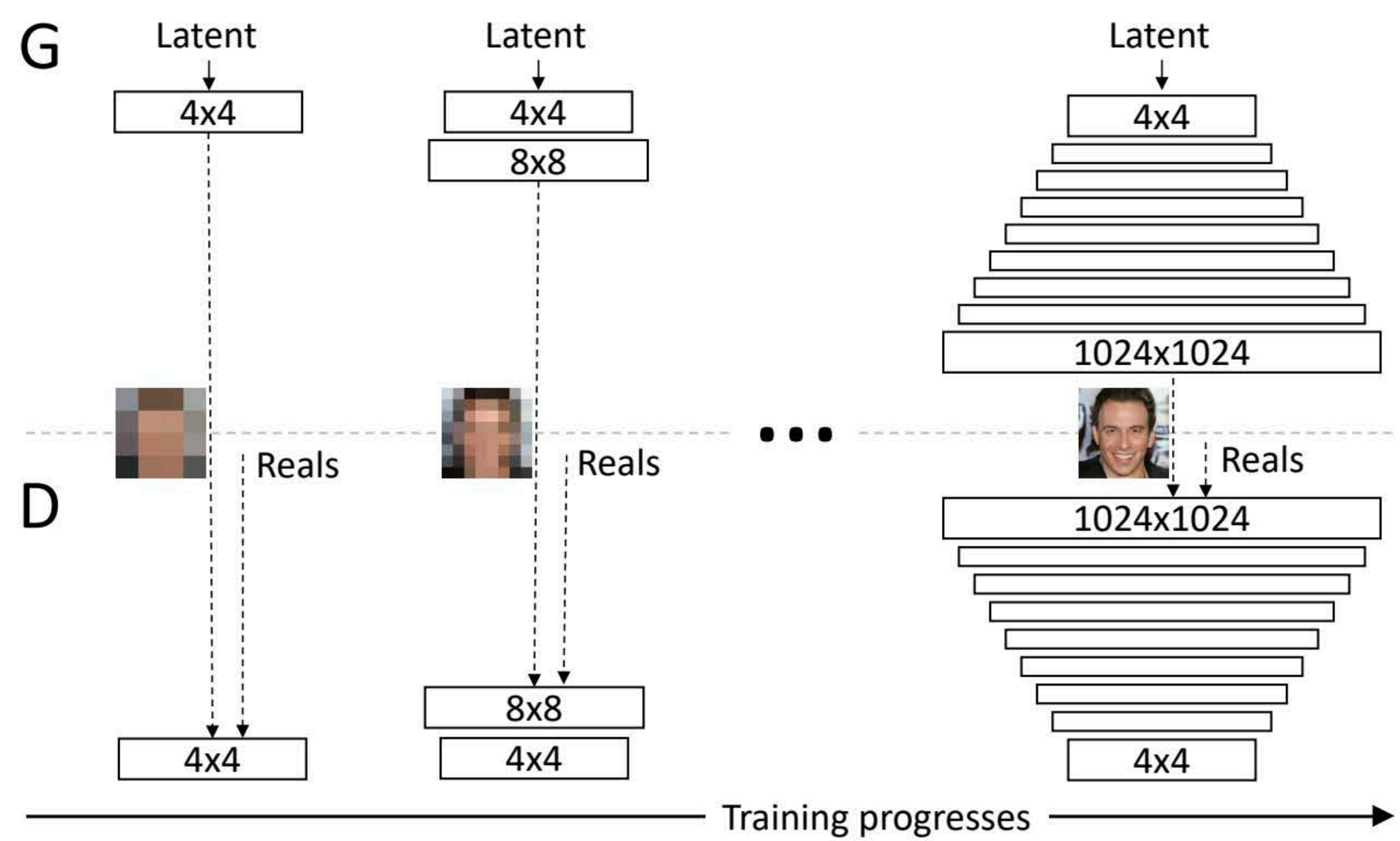
BACKGROUND

- GAN Generator $\mathbf{I} = G(\mathbf{z})$, z: d-dimensional latent; I: image
- PGGAN (ICLR-18)
- BigGAN (ICLR-19)
- StyleGAN (CVPR-19)

BACKGROUND

- GAN Generator $\mathbf{I} = G(\mathbf{z})$, z: d-dimensional latent; I: image

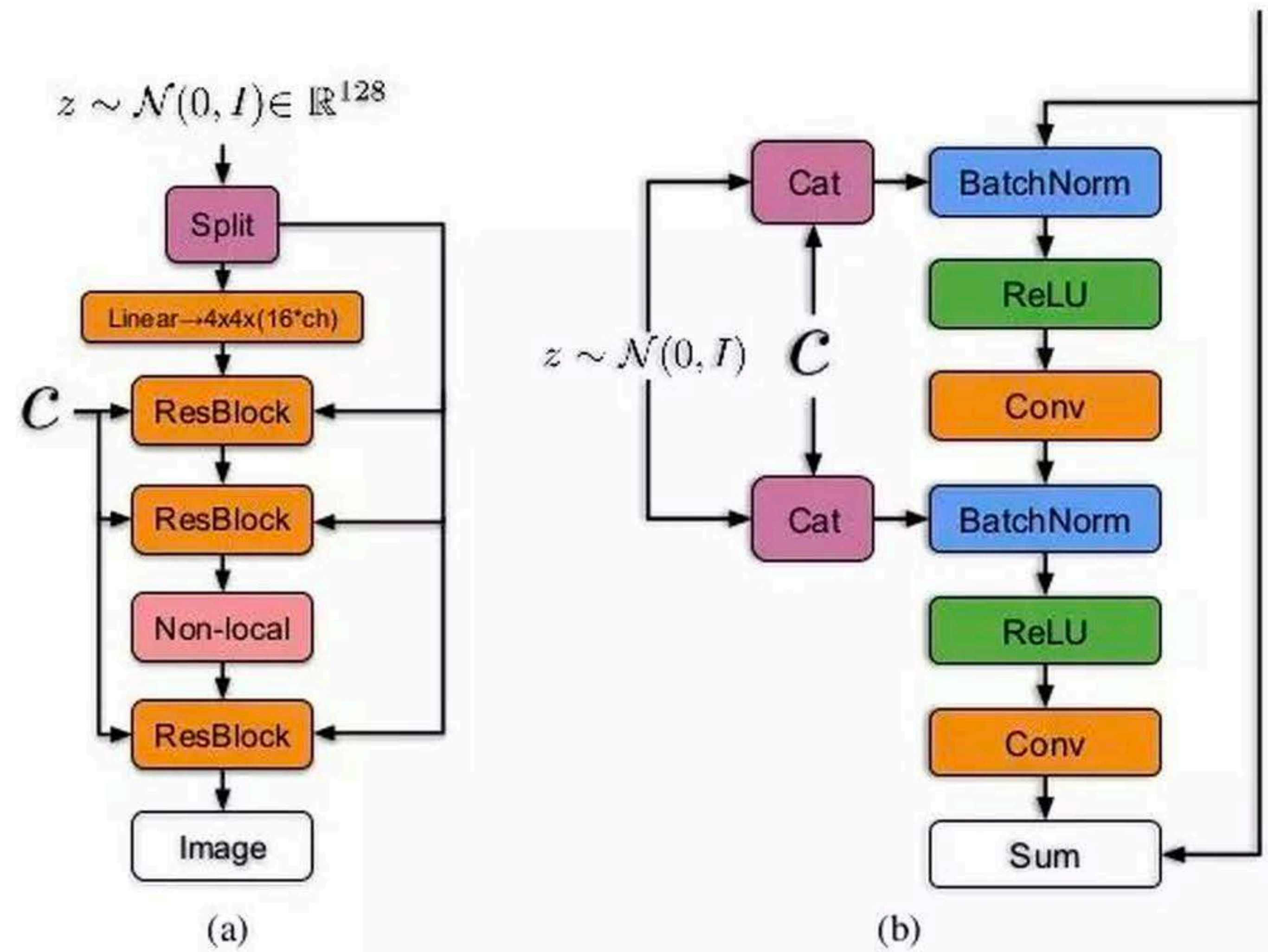
- PGGAN (ICLR-18)
- BigGAN (ICLR-19)
- StyleGAN (CVPR-19)



BACKGROUND

- GAN Generator $\mathbf{I} = G(\mathbf{z})$, z: d-dimensional latent; I: image

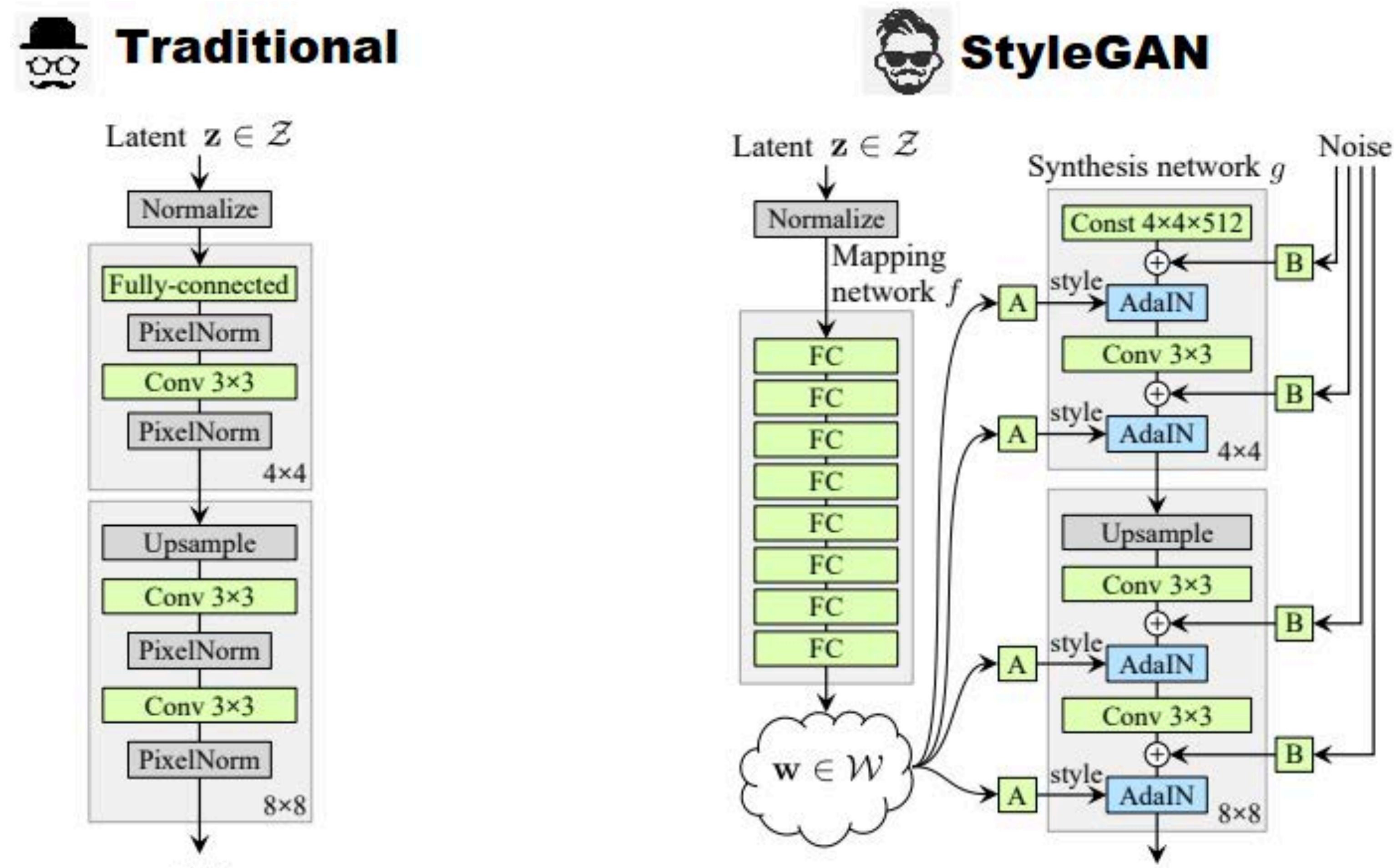
- PGGAN (ICLR-18)
- BigGAN (ICLR-19)
- StyleGAN (CVPR-19)



BACKGROUND

- GAN Generator $\mathbf{I} = G(\mathbf{z})$, z: d-dimensional latent; I: image

- PGGAN (ICLR-18)
- BigGAN (ICLR-19)
- StyleGAN (CVPR-19)



OUTLINE

- Authorship
- Background
- Proposed Method
- Experimental Results
- Conclusion

PROPOSED METHOD

➤ Preliminaries

- GAN Generator $\mathbf{I} = G(\mathbf{z})$, z: d-dimensional latent; I: image
- Focus on the first step, since it is most relevant to the latent space we would like to explore

$$G_1(\mathbf{z}) \triangleq \mathbf{y} = \mathbf{A}\mathbf{z} + \mathbf{b}$$

  - y: m-dimensional projected code
  - A: weight
  - b: bias

PROPOSED METHOD

➤ Preliminaries

- Manipulation/Editing

$$\texttt{edit}(G(\mathbf{z})) = G(\mathbf{z}') = G(\mathbf{z} + \alpha\mathbf{n})$$

  - n: a certain direction to represent a semantic concept
  - α: the manipulation intensity

PROPOSED METHOD

➤ Unsupervised Semantic Factorization

- Manipulation/Editing only consider the first projection step

$$\begin{aligned}\mathbf{y}' \triangleq G_1(\mathbf{z}') &= G_1(\mathbf{z} + \alpha\mathbf{n}) \\ &= \mathbf{A}\mathbf{z} + \mathbf{b} + \alpha\mathbf{A}\mathbf{n} = \mathbf{y} + \alpha\mathbf{A}\mathbf{n}\end{aligned}$$

- Find directions that can cause large variations

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

  - If An = 0, the editing will keep the output unchanged

PROPOSED METHOD

➤ Unsupervised Semantic Factorization

- Find the k most important directions

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

- How to solve? Lagrange multipliers

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

PROPOSED METHOD

➤ Unsupervised Semantic Factorization

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

- Take the partial derivative on each n_i

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

  - Solutions are the eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$

- The proposed method is called **SeFa** (Semantic Factorization)

OUTLINE

- Authorship
- Background
- Proposed Method
- Experimental Results
- Conclusion

EXPERIMENTAL RESULTS

- ➤ Results on Diverse Models and Datasets
- ➤ Comparison with Supervised Approach
- ➤ Comparison with Unsupervised Baselines
- ➤ Real Image Editing

EXPERIMENTAL RESULTS

- Interactive Editing by Tuning Interpretable Directions

Pose

Mouth

Eye

EXPERIMENTAL RESULTS

➤ Interactive Editing by Tuning Interpretable Directions

Posture (Left & Right)

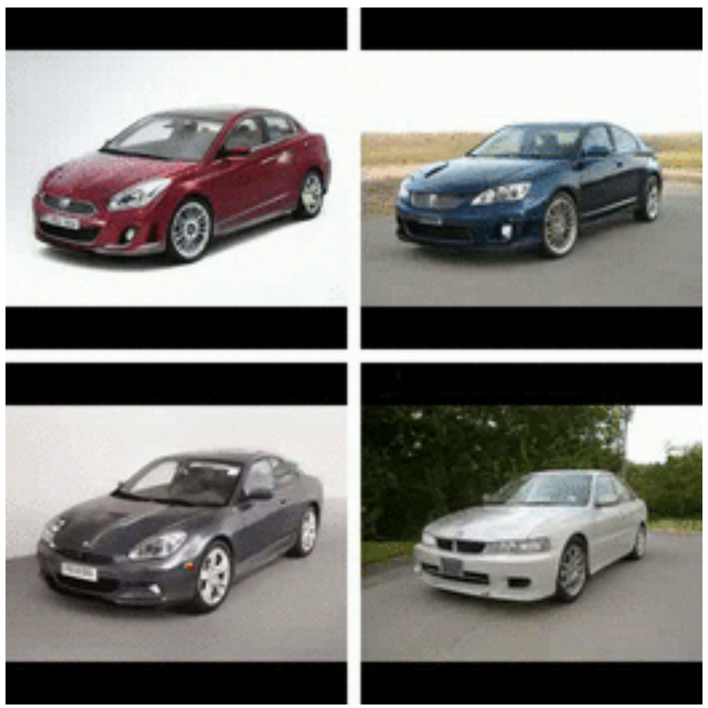
Posture (Up & Down)

Zoom

EXPERIMENTAL RESULTS

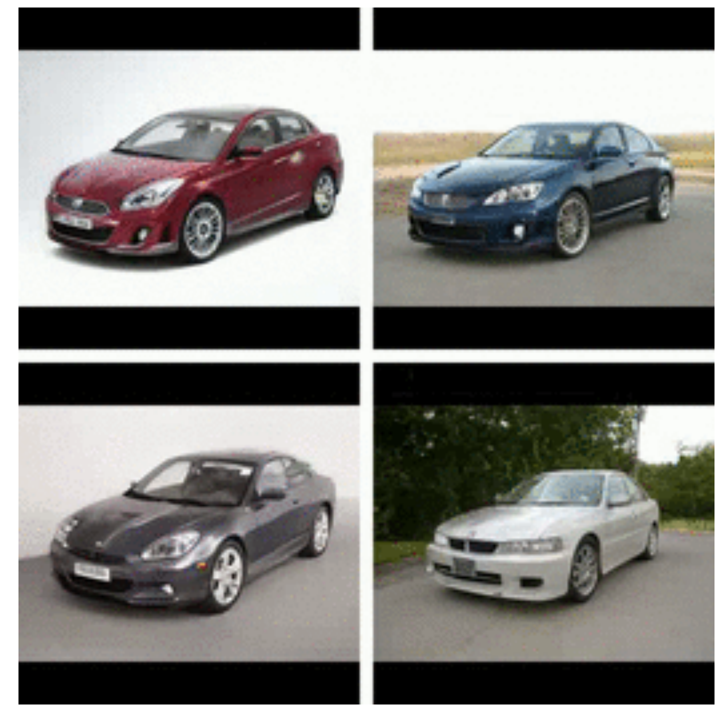
➤ Interactive Editing by Tuning Interpretable Directions

Orientation



Vertical Position

Shape



SeFa: Closed-Form Factorization of Latent Semantics in GANs

Demo Video

Yujun Shen, Bolei Zhou
The Chinese University of Hong Kong

EXPERIMENTAL RESULTS

- ➤ Results on StyleGAN
- ➤ Cars:
- Bottom layers - rotation
- Middle layers - shape
- Top layers - color

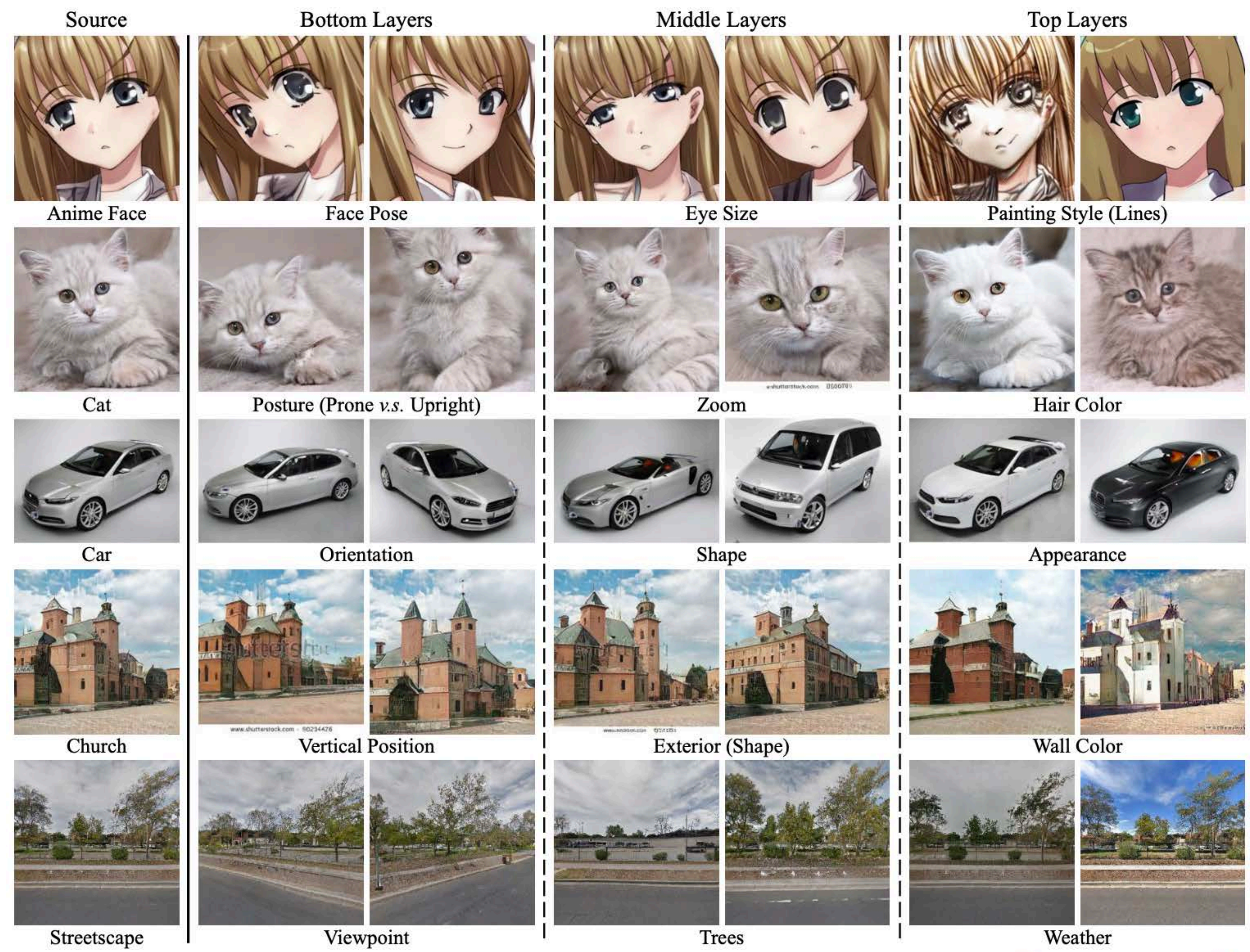


Figure 2. **Hierarchical interpretable directions** discovered in the style-based generators, *i.e.*, StyleGAN [17] and StyleGAN2 [18]. Among them, the streetscapes model is trained with StyleGAN2, while the others are using StyleGAN.

EXPERIMENTAL RESULTS

- SeFa can indeed find human-understandable concepts

Table 1. User study. We randomly generate $2K$ images for each dataset, and use the Top-50 eigen directions from each level of layers to manipulate these images. Numbers in brackets indicate the index of the layers to interpret. Users are asked how many directions result in *obvious* content change (numerator) and how many directions are semantically meaningful (denominator).

Dataset	Bottom (0-1)	Middle (2-5)	Top (6-)
Anime Face [1]	12/12	26/26	38/50
LSUN Cat [27]	14/15	21/28	47/50
LSUN Car [27]	10/10	16/22	22/34
LSUN Church [27]	15/15	18/26	48/50
Streetscape [20]	9/9	12/18	15/36

EXPERIMENTAL RESULTS

➤ Results on BigGAN

Figure 3. **Diverse interpretable directions** found in the BigGAN [4], which is conditionally trained on ImageNet [6]. These semantics are further used to manipulate images from different categories.

EXPERIMENTAL RESULTS

- Results on Diverse Models and Datasets
- Comparison with Supervised Approach
- Comparison with Unsupervised Baselines
- Real Image Editing

EXPERIMENTAL RESULTS

- Comparison with Supervised Approach
- InterFaceGAN (CVPR-20) with well defined facial attributes
  - Requires sampling numerous data and pre-training attribute predictors

EXPERIMENTAL RESULTS

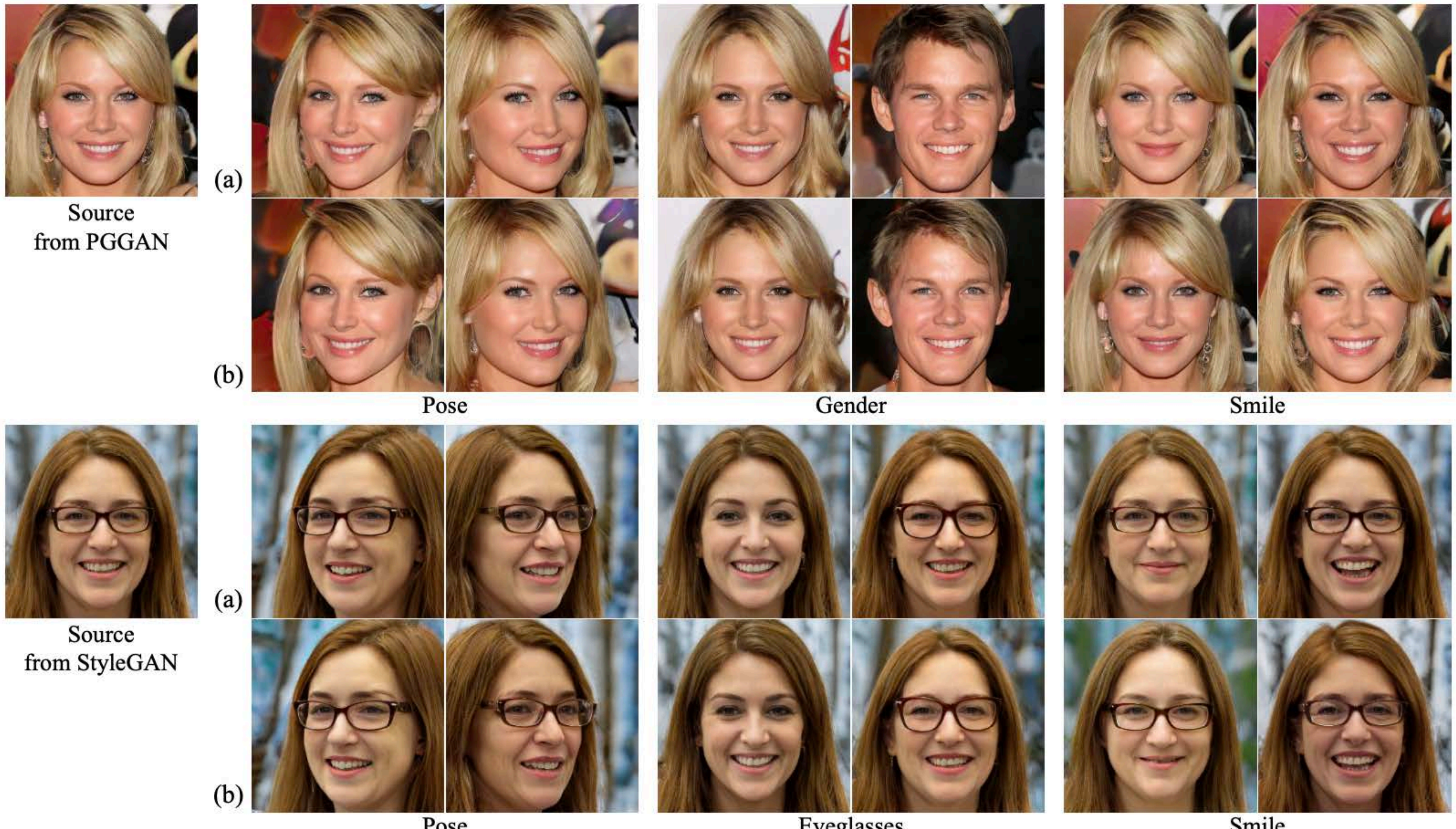


Figure 5. Qualitative comparison of the latent semantics found by (a) the supervised method, InterFaceGAN [24] and (b) our *closed-form* solution, SeFa, where SeFa achieves similar performance to InterFaceGAN. PGGAN trained on CelebA-HQ [16] and StyleGAN trained on FF-HQ [17] are used as the target models to interpret.

EXPERIMENTAL RESULTS

- Train an attribute predictor on CelebA with ResNet50
- Quantitatively evaluate whether the identified directions can properly represent the corresponding attributes

Table 2. **Re-scoring analysis** of the semantics identified by InterFaceGAN [24] and SeFa from the PGGAN model trained on CelebA-HQ dataset [16]. Each row evaluates how the semantic scores change after moving the latent code along a certain direction.

(a) InterFaceGAN [24], 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) SeFa, 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

- SeFa can adequately control some attribute similar to InterFaceGAN.

EXPERIMENTAL RESULTS

- ➤ Train an attribute predictor on CelebA with ResNet50
- ➤ Quantitatively evaluate whether the identified directions can properly represent the corresponding attributes

Table 2. **Re-scoring analysis** of the semantics identified by InterFaceGAN [24] and SeFa from the PGGAN model trained on CelebA-HQ dataset [16]. Each row evaluates how the semantic scores change after moving the latent code along a certain direction.

(a) InterFaceGAN [24], 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) SeFa, 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

- When altering one semantic, InterFaceGAN shows stronger robustness to other attributes, benefiting from its supervised training manner.

EXPERIMENTAL RESULTS

- ➤ Train an attribute predictor on CelebA with ResNet50
- ➤ Quantitatively evaluate whether the identified directions can properly represent the corresponding attributes

Table 2. **Re-scoring analysis** of the semantics identified by InterFaceGAN [24] and SeFa from the PGGAN model trained on CelebA-HQ dataset [16]. Each row evaluates how the semantic scores change after moving the latent code along a certain direction.

(a) InterFaceGAN [24], 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) SeFa, 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

- SeFa fails to discover the direction corresponding to eyeglasses.
- The presence of eyeglasses is not a large variation.

EXPERIMENTAL RESULTS

➤ SeFa can find more diverse semantics in the latent space

- Hair color, hair style, and brightness (not easy to acquire)
- More complex attributes

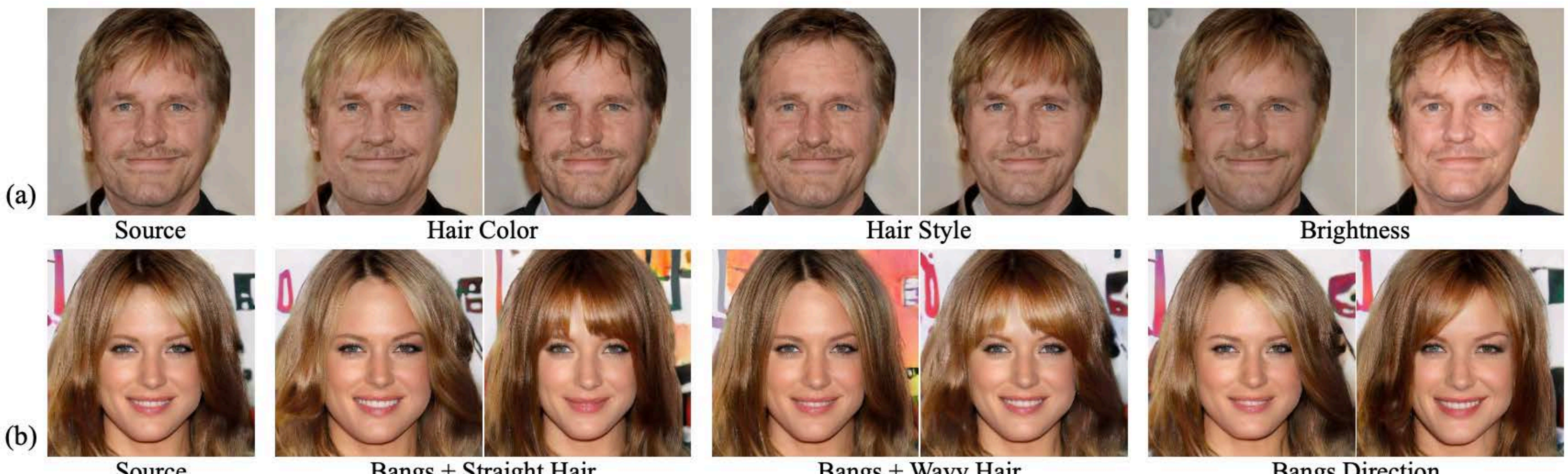


Figure 6. (a) Diverse semantics, which can *not* be identified by InterFaceGAN [24] due to the lack of semantic predictors. (b) Diverse hair styles, which can *not* be described as a binary attribute. The PGGAN model trained on CelebA-HQ dataset [16] is used.

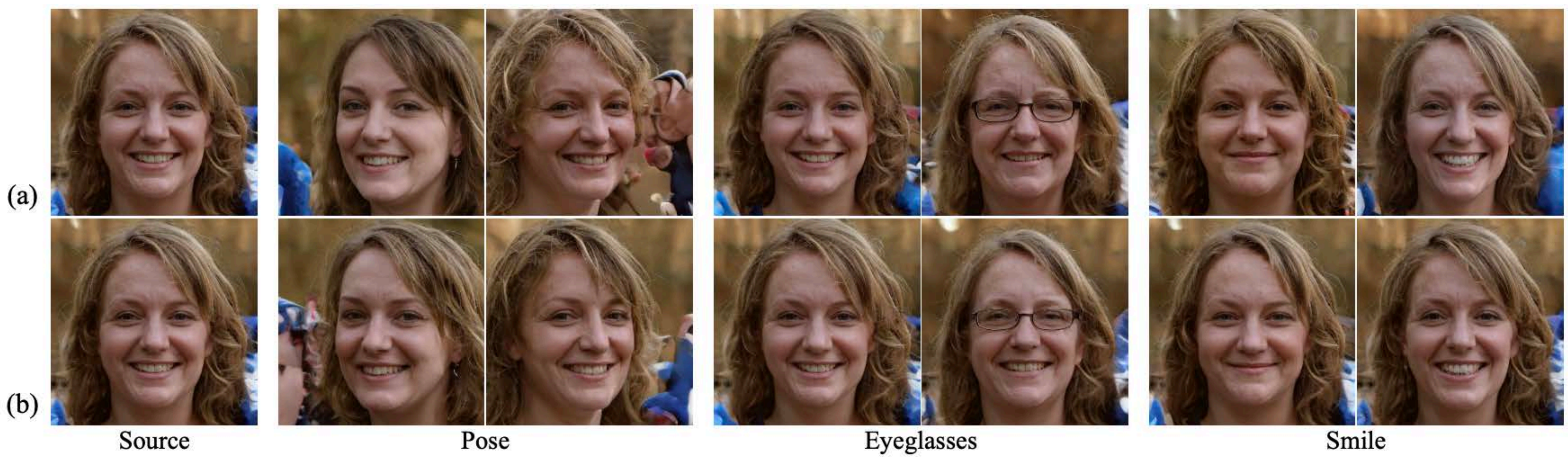
EXPERIMENTAL RESULTS

- Results on Diverse Models and Datasets
- Comparison with Supervised Approach
- Comparison with Unsupervised Baselines
- Real Image Editing

EXPERIMENTAL RESULTS

➤ Comparison with Unsupervised Baselines

- Sampling-based Baseline
- GANSpace (NeurIPS-20): PCA on a collection of sampled data



The semantics found by SeFa lead to a more precise control

EXPERIMENTAL RESULTS

➤ Comparison with Unsupervised Baselines

- Sampling-based Baseline
- GANSpace (NeurIPS-20): PCA on a collection of sampled data

	FID	Re-scoring	User Study
GANSpace [10]	7.43	0.33	41%
SeFa (Ours)	**7.36**	**0.38**	**59%**

EXPERIMENTAL RESULTS

➤ Comparison with Unsupervised Baselines

- Learning-based Baseline
- InfoGAN (NeurIPS-16): use a regularizer to maximize the mutual information between the output image and the input latent code

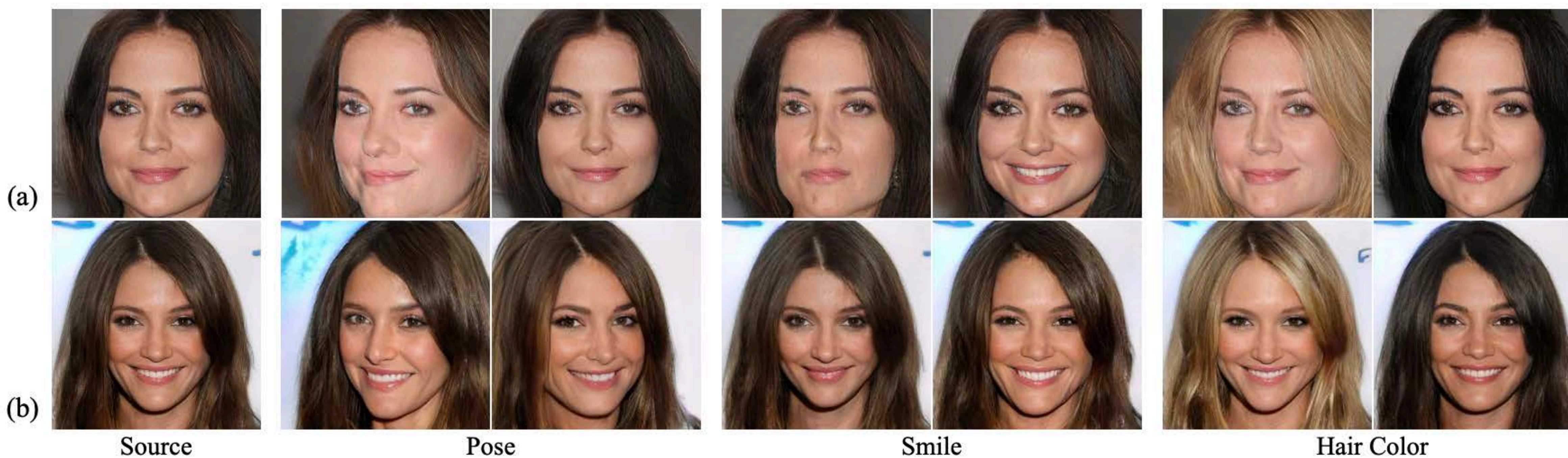


Figure 8. Qualitative comparison between (a) Info-PGGAN [21, 5] and (b) SeFa. The result of the Info-PGGAN model is extracted directly from [21], and the official PGGAN model trained on CelebA-HQ dataset [16] is used for SeFa.

EXPERIMENTAL RESULTS

- ➤ Results on Diverse Models and Datasets
- ➤ Comparison with Supervised Approach
- ➤ Comparison with Unsupervised Baselines
- ➤ Real Image Editing

EXPERIMENTAL RESULTS

- ➤ Real Image Editing
  - Given a target image to edit, first project it back to the latent space, then use the variation factor found by SeFa to modulate

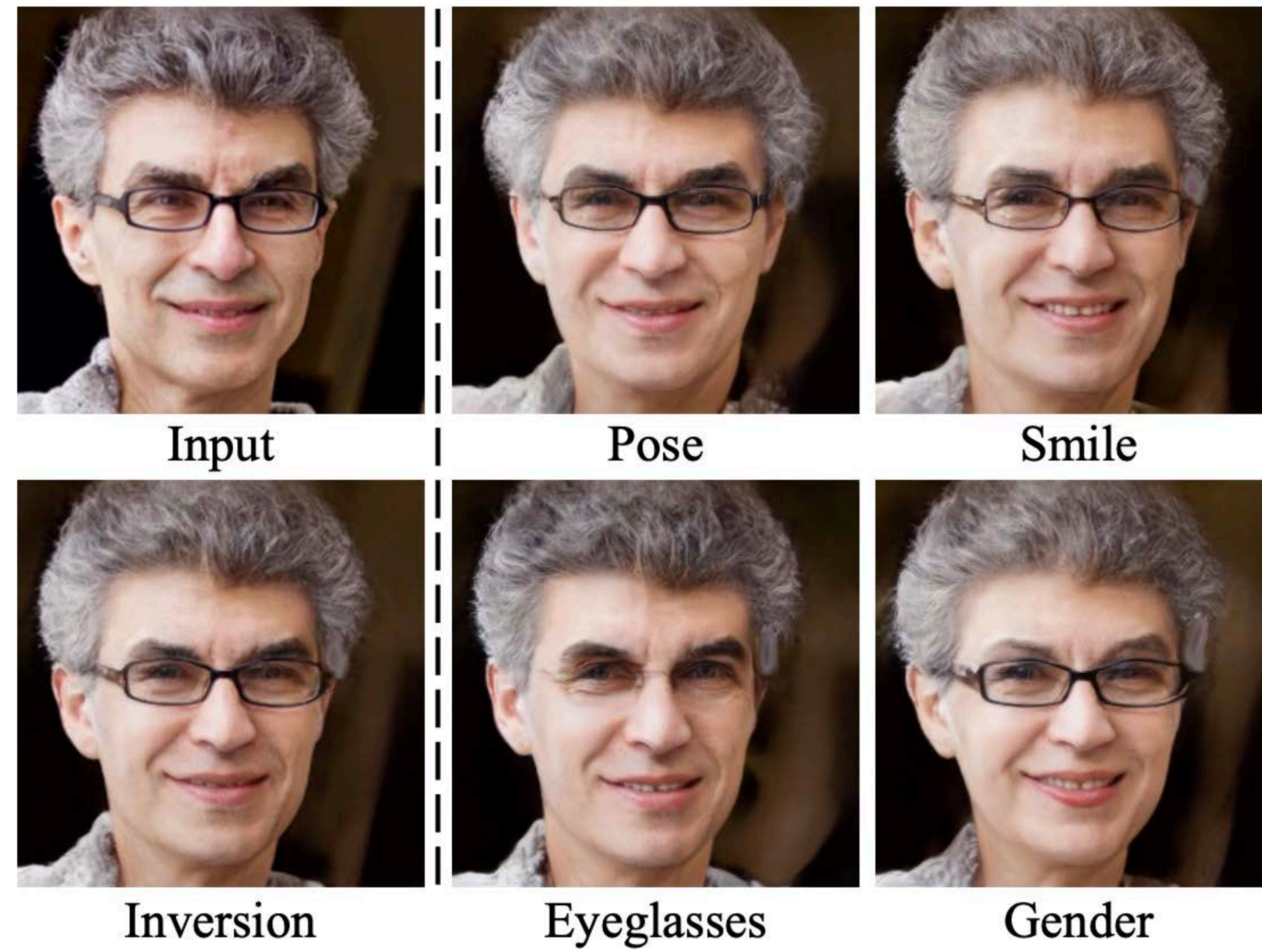


Figure 9. **Real image editing** with respect to various facial attributes. All semantics are found with the proposed SeFa. GAN inversion [28] is used to project the target real image back to the latent space of StyleGAN [17].

OUTLINE

- Authorship
- Background
- Proposed Method
- Experimental Results
- Conclusion

CONCLUSION

- Factorizing the latent semantics learned by GANs
- Identifying versatile semantics from different types of GAN models in an unsupervised manner