# Closed-Form Factorization of Latent Semantics in GANs

(Shen, Y., & Zhou, B., CVPR, 2021)

딥러닝논문읽기모임

**Date: 2021.08.01**

**Member: 고형권, 허다운**

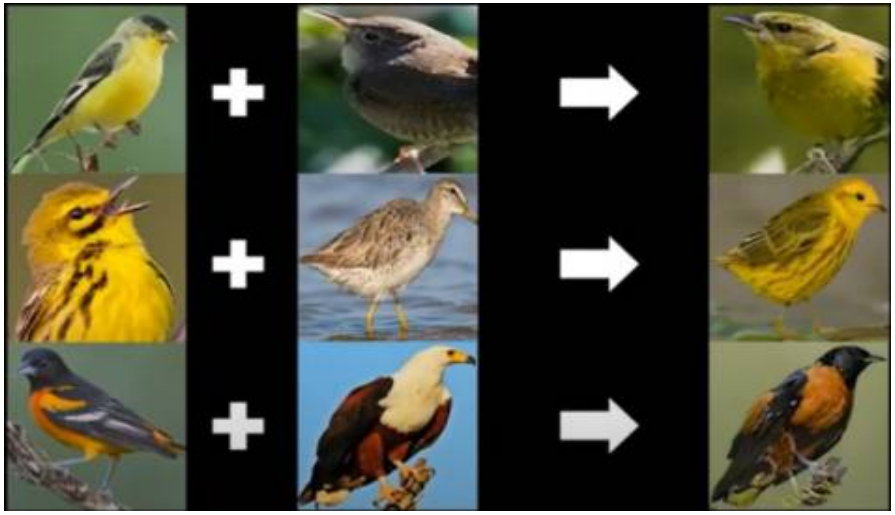
**Presenter: 김상현**

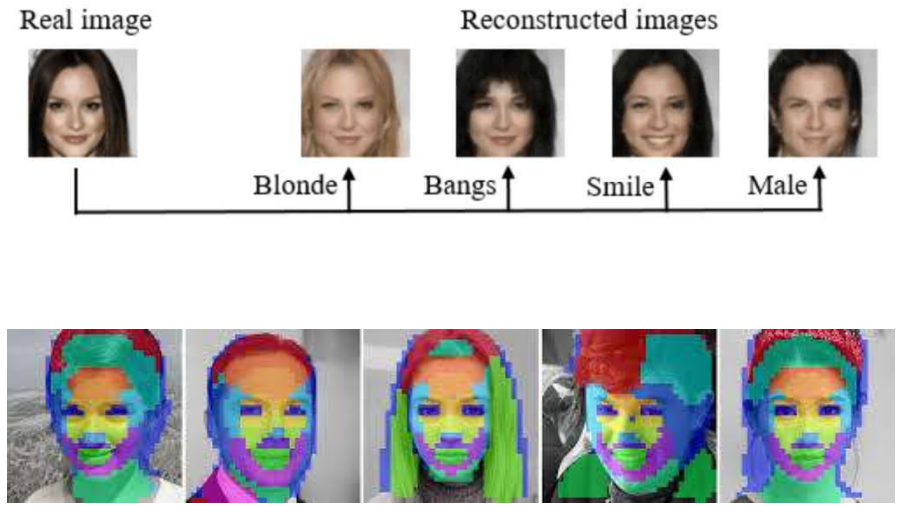
# Contents

# Introduction

- **GAN's trend**
  - **Image Synthesis → Image Editing.**
  - **But, Existing GAN requires manual annotation definition for Image Synthesis, so practical application is limited.**

**→ SeFa (Semantic Factorization) algorithm (Unsupervised Learning.)**





# Method

- **Preliminaries**
  - **Equation (1): the first projection step.**

$$G_1(\mathbf{z}) \triangleq \mathbf{y} = \mathbf{A}\mathbf{z} + \mathbf{b}, \qquad (1)$$

    - **G: Generator**
    - **z: Latent Code**

  - **Equation (2): Image Editing**

$$\texttt{edit}(G(\mathbf{z})) = G(\mathbf{z}') = G(\mathbf{z} + \alpha\mathbf{n}), \qquad (2)$$

    - **n: 방향**
    - **$\alpha$: 강도**

# Method

- **Unsupervised Semantic Factorization**
  - **Image editing has nothing to do with latent code z.**

$$\begin{aligned}\mathbf{y}' \triangleq G_1(\mathbf{z}') &= G_1(\mathbf{z}+\alpha\mathbf{n}) \\ &= \mathbf{A}\mathbf{z}+\mathbf{b}+\alpha\mathbf{A}\mathbf{n} = \mathbf{y}+\alpha\mathbf{A}\mathbf{n}. \end{aligned} \quad (3)$$

  - **Dependent on $\alpha$An.**

$$\mathbf{n}^* = \underset{\{\mathbf{n}\in\mathbb{R}^d:\ \mathbf{n}^T\mathbf{n}=1\}}{\arg\max} ||\mathbf{A}\mathbf{n}||_2^2, \quad (4)$$

$$\mathbf{N}^* = \underset{\{\mathbf{N}\in\mathbb{R}^{d\times k}:\ \mathbf{n}_i^T\mathbf{n}_i=1\ \forall i=1,\cdots,k\}}{\arg\max} \sum_{i=1}^{k} ||\mathbf{A}\mathbf{n}_i||_2^2, \quad (5)$$

  - **Lagrange multipliers Method.**

$$\begin{aligned}\mathbf{N}^* &= \underset{\mathbf{N}\in\mathbb{R}^{d\times k}}{\arg\max} \sum_{i=1}^{k} ||\mathbf{A}\mathbf{n}_i||_2^2 - \sum_{i=1}^{k}\lambda_i(\mathbf{n}_i^T\mathbf{n}_i-1) \\ &= \underset{\mathbf{N}\in\mathbb{R}^{d\times k}}{\arg\max} \sum_{i=1}^{k}(\mathbf{n}_i^T\mathbf{A}^T\mathbf{A}\mathbf{n}_i - \lambda_i\mathbf{n}_i^T\mathbf{n}_i+\lambda_i). \end{aligned} \quad (6)$$

# Q & A

# Experiments

- **Results on Diverse Models and Datasets**
  - **Interactive Editing by Tuning Interpretable Directions.**
    - **Models**
      - **StyleGAN**
      - **BigGAN**
      - **StyleGAN2**
    - **Datasets**
      - **Human Face(FF-HQ)**
      - **Anime faces**
      - **Scenes**
      - **Objects**
      - **Streetscapes**
      - **ImageNet**
  - **Interactive Interface**


Figure 4. **Interface** for interactive editing.

# Experiments

- **Results on Diverse Models and Datasets**
  - **Results on StyleGAN**
    - **User Study.**
    - **( ): layer.**
    - **Numerator: how many directions result in obvious content change.**
    - **Denominator: how many directions are semantically meaningful.**

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 [28] | 14/15 | 21/28 | 47/50 |
| LSUN Car [28] | 10/10 | 16/22 | 22/34 |
| LSUN Church [28] | 15/15 | 18/26 | 48/50 |
| Streetscape [20] | 9/9 | 12/18 | 15/36 |

# Experiments

- **Results on Diverse Models and Datasets**
  - **Results on BigGAN**
    - **The semantics found by SeFa can be applied to manipulating images from different categories.**
    - **This verifies the generalization ability of SeFa.**


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.

# Experiments

- **Comparison with Supervised Approach**
  - **Qualitative Results (SeFa vs InterFaceGAN)**
    - **SeFa achieves similar performance as InterFaceGAN.**
    - **SeFa is more efficient and generalizable.**

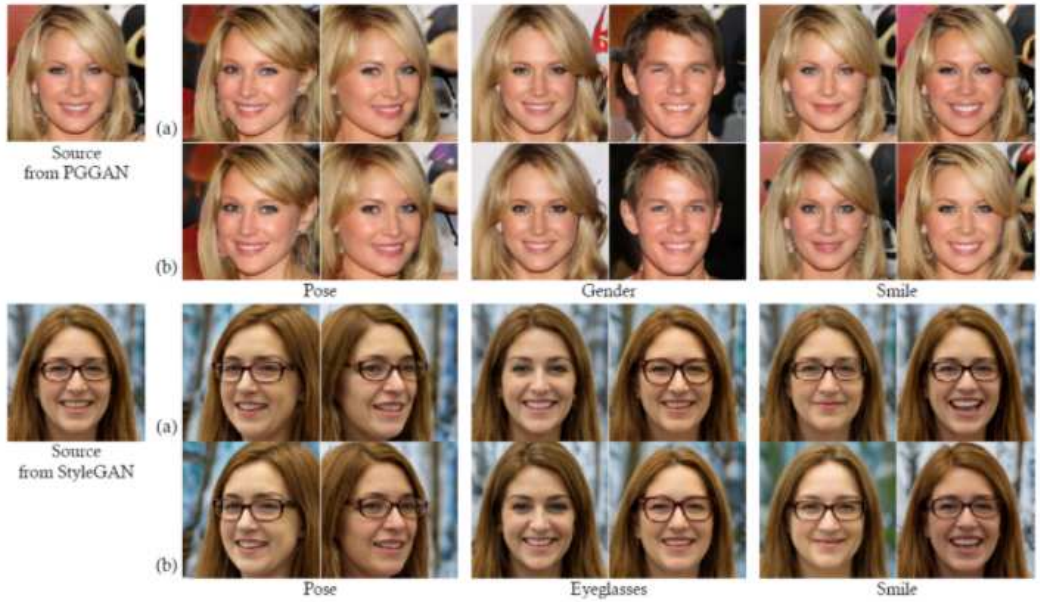



Figure 5. Qualitative comparison of the latent semantics found by (a) the supervised method, InterFaceGAN [24] and (b) our *closed-form* solution, SeFa, where SeFa achieves similar performance to InterFaceGAN. PGGAN trained on CelebA-HQ [16] and StyleGAN trained on FF-HQ [17] are used as the target models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

# Experiments

- **Comparison with Supervised Approach**
  - **Re-scoring Analysis (SeFa vs InterFaceGAN)**
    - **They train an attribute predictor on CelebA dataset with ResNet-50.**
    - **3 Observation.**
      - **SeFa can adequately control some attribute similar to InterFaceGAN.**
      - **InterFaceGAN shows stronger robustness to other attributes, benefiting from its supervised training manner.**
      - **SeFa fails to discover the direction corresponding to eyeglasses.**

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 |

# Experiments

- **Comparison with Supervised Approach**
  - **Diversity Comparison**
    - **Supervised learning: Depends on the available attribute predictors.**
    - **InterFaceGAN (Supervised learning): Binary attribute**
    - **SeFa (Unsupervised learning): Diverse attribute**


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.

# Experiments

- **Comparison with Unsupervised Baselines**
  - **Comparison with Sampling-based Baseline**
    - **GANSpace(a): To perform PCA on a collection of sampled data to find principal directions in the latent space.**
    - **The semantics found by SeFa lead to a more precise control.**

Table 3. Quantitative comparison with GANSpace [10].

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

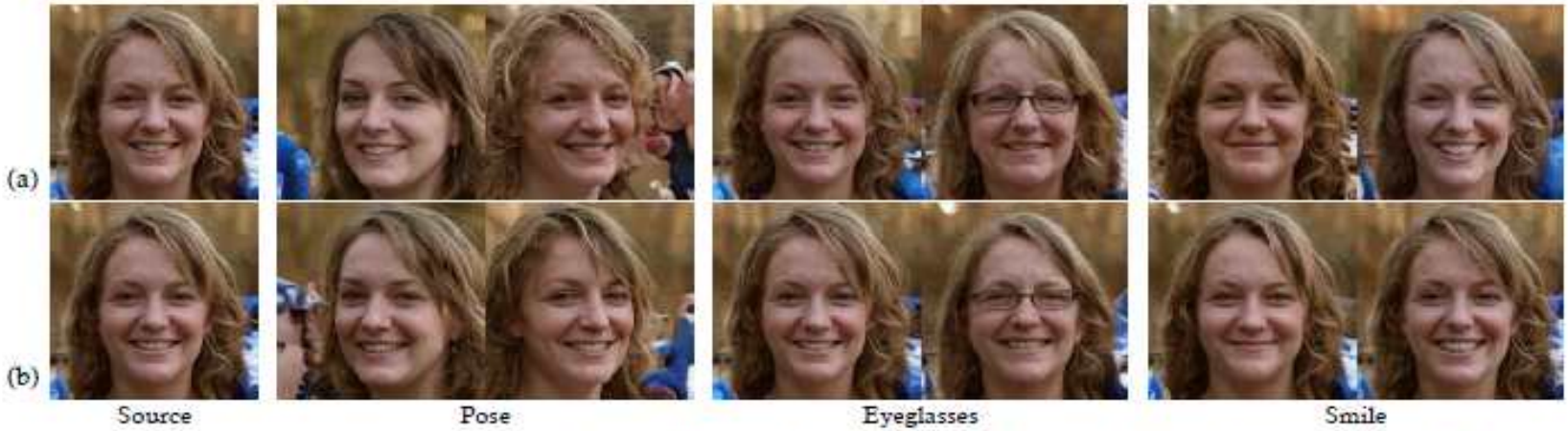



Figure 7. Qualitative comparison between (a) GANSpace [10] and (b) SeFa. The StyleGAN model trained on FF-HQ dataset [17] is used.

# Experiments

- **Comparison with Unsupervised Baselines**
  - **Comparison with Learning-based Baseline ( SeFa vs Info-GAN )**
    - **the semantics identified by SeFa are more accurate than those learned from Info-PGGAN.**
    - **ex) Taking pose manipulation as an example, the hair color varies when using Info-PGGAN for editing.**


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.

# Experiments

- **Real Image Editing**

  - **The generator lacks the inference ability to take a real image as the input.**

    - **GAN Inversion.**

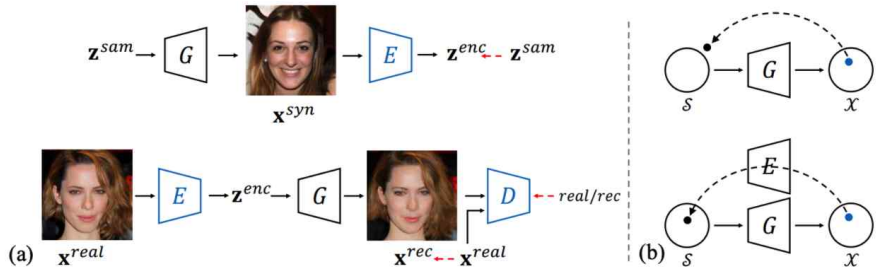



**Fig. 2.** (a) The comparison between the training of conventional encoder and *domain-guided* encoder for GAN inversion. Model blocks in **blue** are trainable and **red** dashed arrows indicate the supervisions. Instead of being trained with synthesized data to recover the latent code, our *domain-guided* encoder is trained with the objective to recover the real images. The *fixed* generator is involved to make sure the codes produced by the encoder lie in the native latent space of the generator and stay semantically meaningful. (b) The comparison between the conventional optimization and our *domain-regularized* optimization. The well-trained *domain-guided* encoder is included as a regularizer to land the latent code in the semantic domain during the optimization process.


Figure 9. **Real image editing** with respect to various facial attributes. All semantics are found with the proposed SeFa. GAN inversion [29] is used to project the target real image back to the latent space of StyleGAN [17].

# Conclusion

- **Unsupervised learning을 사용.**

- **Weight 자체가 가지고 있는 learned characteristics를 직접적으로 사용.**

# Q & A