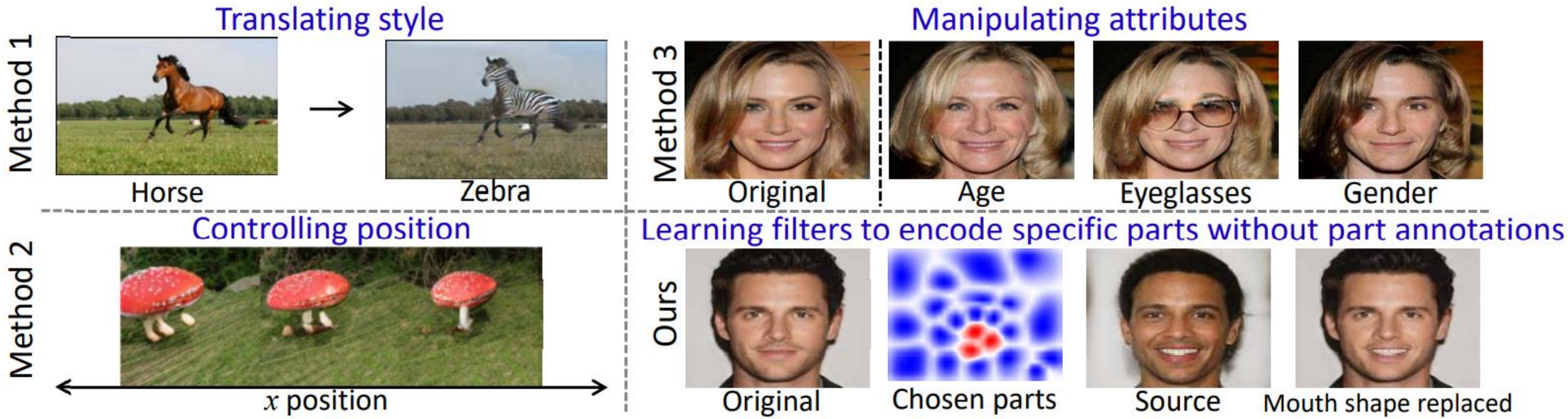
中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Interpretable Generative Adversarial Networks

Chao Li[1,3*†], Kelu Yao[1*], Jin Wang[1*], Boyu Diao[1], Yongjun Xu[1], Quanshi Zhang[2†]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]Shanghai Jiao Tong University, China
[3]Zhejiang Laboratory, Hangzhou 311100, China
*Equal contributions †Correspondence authors

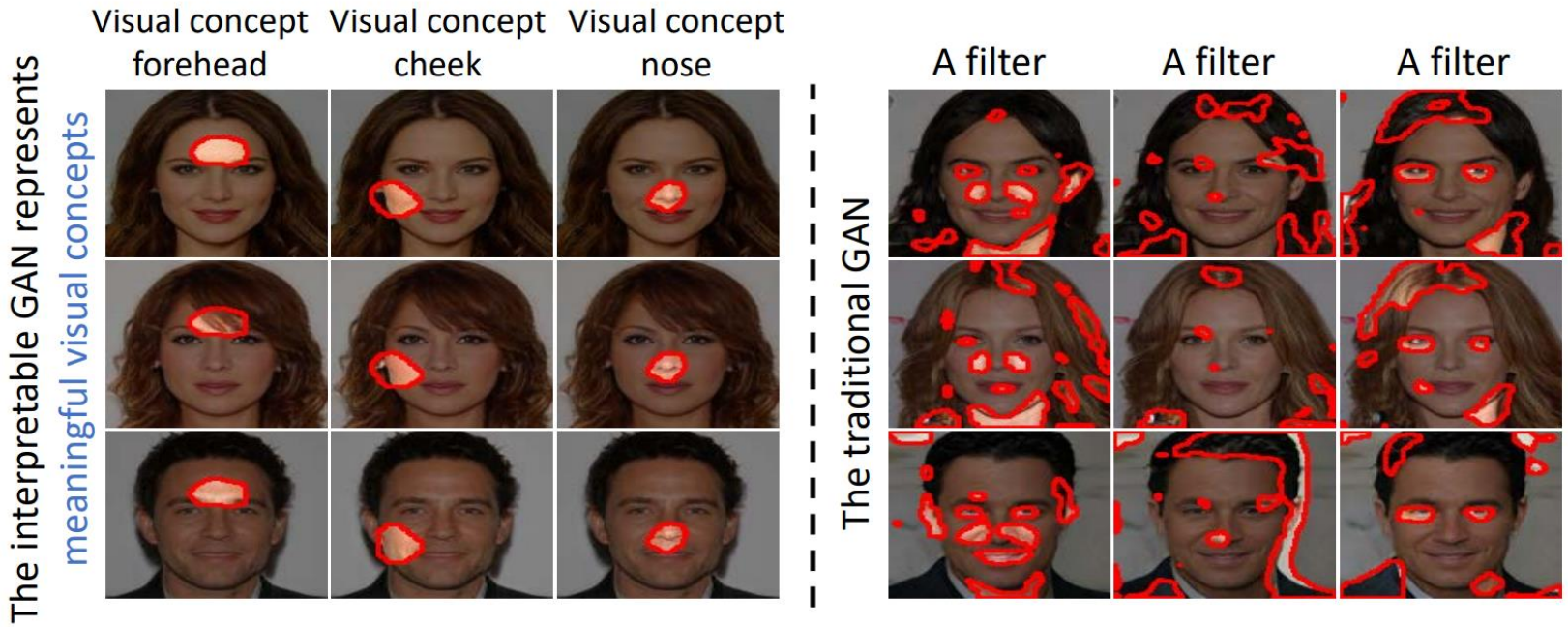
Introduction



- Method 1[1]: disentangled the structure and the style of the image.
- Method 2[2]: learned features for the localized object in the image.
- Method 3[3]: learned the disentangled features for attributes of the image.
- Our Interpretable GAN: learns each filter to encode an object part without part annotations.

[1] Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, 2223–2232.
[2] Plumerault, A.; Le Borgne, H.; and Hudelot, C. 2019. Controlling generative models with continuous factors of variations. In International Conference on Learning Representations
[3] Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2020. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9243–9252.

Introduction



- Traditional GAN: has no self-reflection of its representations.
- Our Interpretable GAN:
  - each filter consistently represents a meaningful visual concept when generating different images.
  - different filters represent different visual concepts.

Objective

- The interpretable GAN should satisfy the following two objectives.

  - **Interpretability of filters**: we expect filters in an intermediate layer of the generator to automatically learn meaningful visual concepts without manual annotations of visual concepts.
    - Each filter is supposed to consistently generate image regions corresponding to the same visual concept when generating different images.
    - Different filters are supposed to generate image regions corresponding to different visual concepts.
  - **Realism of generated images**: the generator of the interpretable GAN still generates realistic images.

Method

- To ensure the **interpretability of filters** in the target layer,
  - we use a set of filters to jointly represent a specific visual concept;
  - we use different sets of filters to represent different visual concepts.

- To ensure the **realism of generated images** in the same time,
  - we design the following loss function to revise a tradition GAN into an interpretable GAN:

$$\mathbf{L} = \mathcal{L}_{GAN}(G, D) + \lambda_0 Loss(\mathbf{Q}, G)$$

Method

- To learn the partition **Q** for the **interpretability of filters**, we use a Gaussian Mixture Model (GMM) to ensure that filters in the same group generate similar image regions.

$$\max_{\Theta} \mathcal{L}_{GMM} = \max_{\Theta} \sum_{j=1}^{M} log P_{\Theta}(F^j)$$

- Given the optimal parameters of GMM, the partition **Q** is solved as follows:

$$\mathbf{Q} = \{q_j | \arg\max_{q^j} P_{\Theta'}(q^j | F^j)\}$$

Method

- To further ensure the **realism of generated images**, we use an energy-based model to estimate the realism of the feature maps. The energy-based model is learned as follows.

$$\mathcal{L}_{real}(W, G) = -\frac{1}{N}\sum_{i=1}^{N} log P_W(f_G(z_i)|\mathbf{Q})$$

- The energy-based model is formulated as follows:

$$P_W(f_G(z)|\mathbf{Q}) = \frac{1}{Z(W)} exp\Big(g_W(f_G(z))\Big) P_0(z)$$

$$g_W(f_G(z)) = \sum_{j=1}^{M}\sum_{c=1}^{C}[W_{jc} \cdot (f^j \odot \bar{f}^c)]$$

Method

- To increase the **interpretability of the filters** in the target layer, we expect each filter in the same group to exclusively generate the same image region.

- Based on the aforementioned energy-based model, we design the following loss:

$$\mathcal{L}_{interp}(W) = -\sum_{j=1}^{M}\sum_{c=1}^{C}\sum_{k=1}^{K}\mathcal{I}(q_j = c)W_{jck} + \lambda_1 \sum_{j=1}^{M}\sum_{c=1}^{C}\sum_{k=1}^{K}\mathcal{I}(q_j \neq c)W_{jck}$$

Method

- To sum up, the added loss is designed as follows:

$$Loss(W, \mathbf{Q}, G) = \sum_{q_j \in \mathbf{Q}} P_{\Theta'}(q^j | F^j) + \lambda_2 \mathcal{L}_{real}(W, G) + \lambda_3 \mathcal{L}_{interp}(W)$$

- The overall loss is optimized as follows:

$$\min_{W,G} \max_{D,\mathbf{Q}} \mathbf{L}$$

Method

- **Learning.**
  - The gradient of $\mathcal{L}_{real}(W, G)$ w.r.t. W can not be calculated directly and has to be approximated by Markov chain Monte Carlo (MCMC), such as the Langevin dynamics[4].

$$\frac{\partial}{\partial W}\mathcal{L}_{real}(W, G)$$
$$\approx \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial W}g_W(f_{G'}(\hat{z}_i)) - \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial W}g_W(f_G(z_i))$$

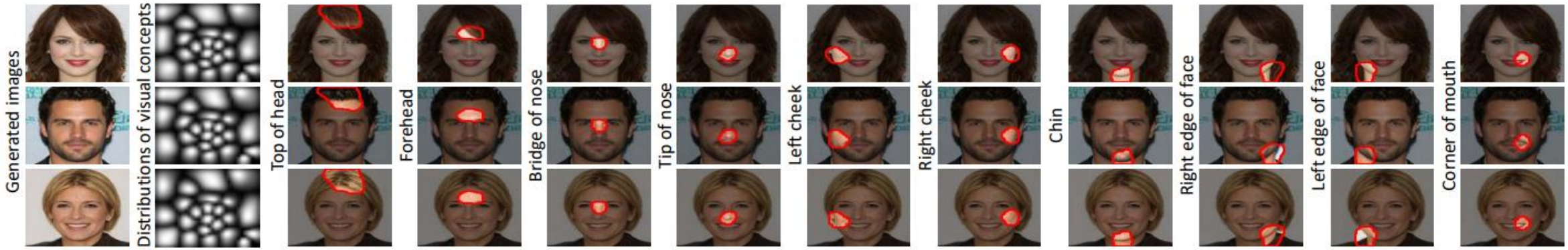
  - The iterative process of Langevin dynamics is carried out as follows:

$$z^{\tau+1} = z^{\tau} + \frac{\delta^2}{2}\frac{\partial}{\partial z}P_W(f_G(z^{\tau})|\mathbf{Q}) + \delta U^{\tau}$$

[4] Girolami, M.; and Calderhead, B. 2011. Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(2): 123–214

Qualitative evaluation

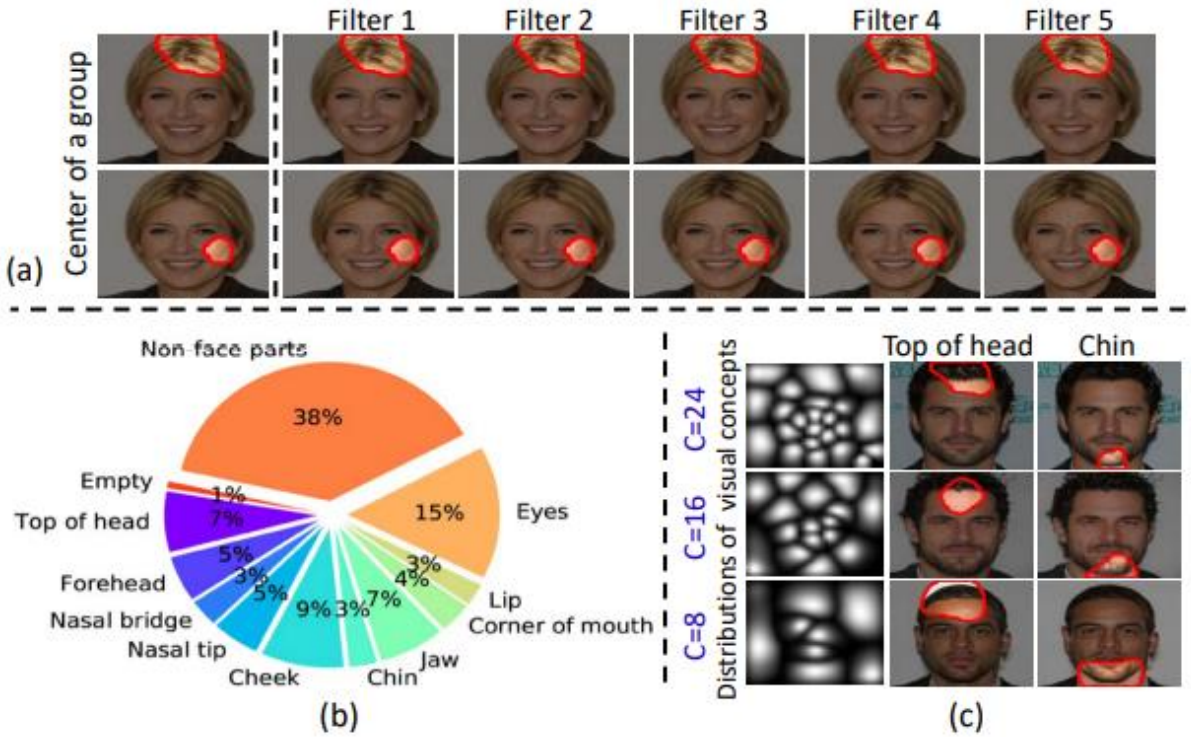
- **Visualization.**



Visualization results show that each filter in an interpretable GAN consistently generated image regions corresponding to the same visual concept. Different filters generated image regions corresponding to different visual concepts.

Qualitative evaluation

- **Visualization.**



- *We compared RFs between the group center and filters in this group, as shown in Fig (a).*
- *Fig. (b) illustrates the proportions of filters representing different visual concepts when setting C = 24.*
- *As shown in Fig (c), when setting different values of C, GANs with a larger value of C learned more detailed concepts.*

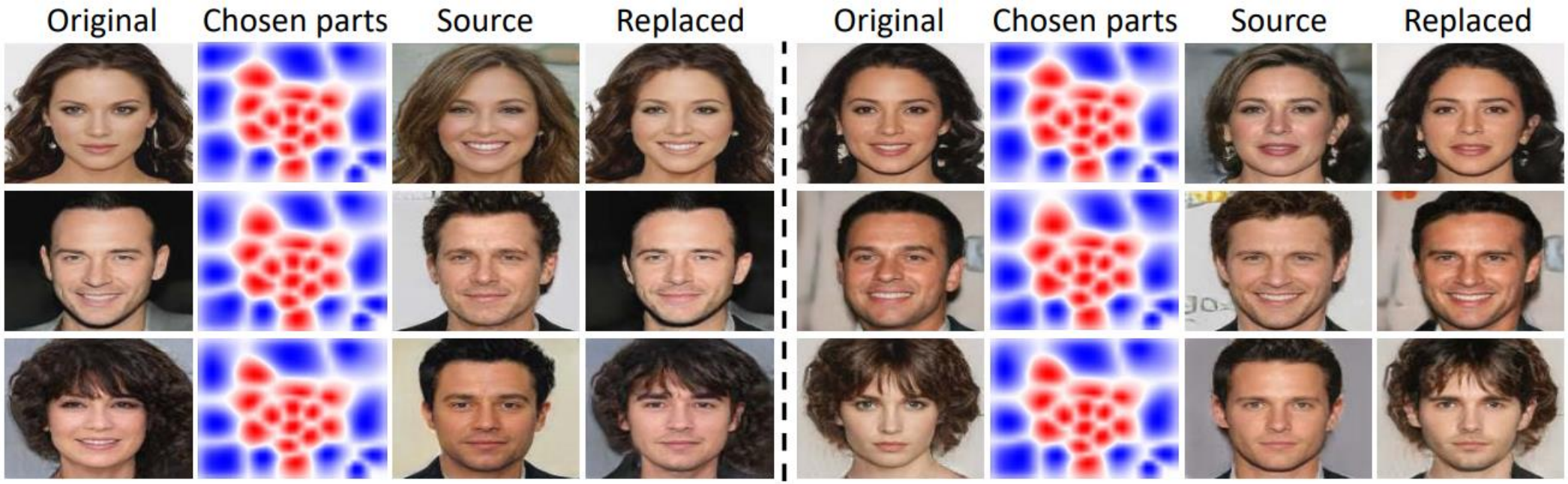
Qualitative evaluation

- **Modifying visual concepts on images.**

- *We exchanged a specific visual concept between pairs of images by exchanging the corresponding feature maps in the interpretable layer.*
- *Our method only modified a localized visual concept without changing other unrelated regions.*

Qualitative evaluation

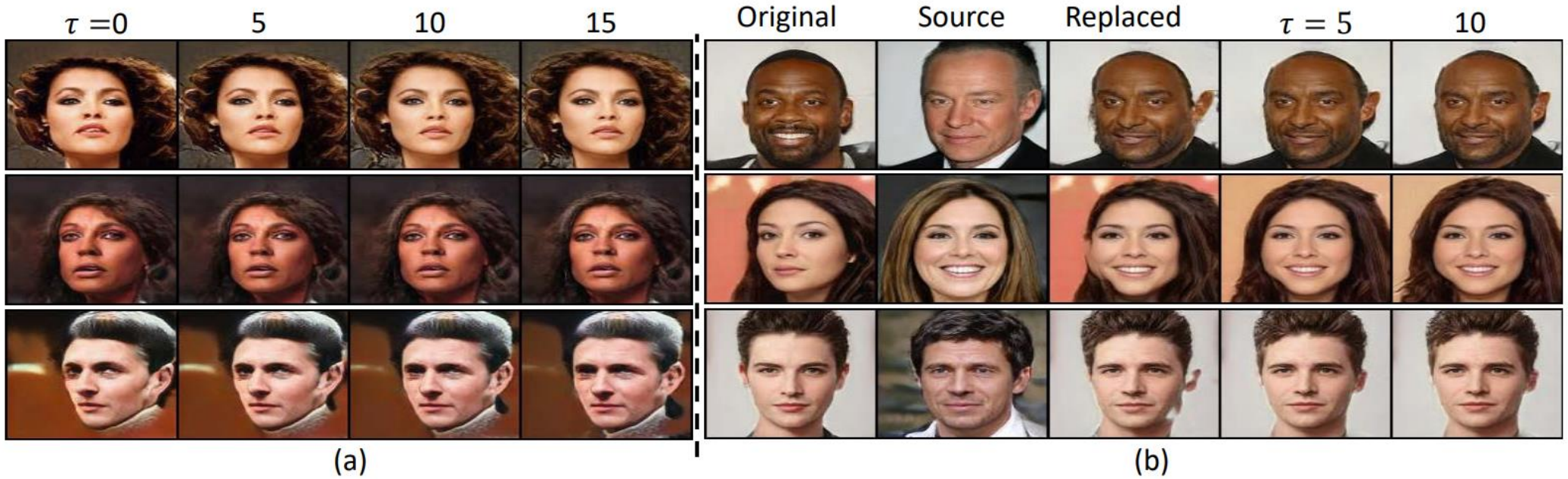
- **Modifying visual concepts on images.**



  - *We also exchanged whole faces between pairs of images.*
  - *The replaced images show that our method successfully exchanged faces between pairs of images.*

Qualitative evaluation

- **Improving the realism of images.**



- *Fig (a): Improving the realism of generated images by Langevin dynamics.*
- *Fig (b): Improving the realism of modified images by Langevin dynamics.*

Quantitative analysis

- **Human perception evaluation.**

Model	Mouth(%)	Eyes(%)	Chin(%)
Editing in Style (Collins et al. 2020)	37.90	34.60	-
Feature Collaging (Suzuki et al. 2018)	56.00	45.40	46.40
Interpretable StyleGAN	83.60	63.70	81.67
Interpretable BigGAN	89.60	82.10	92.30

  - *We conduct a user study to evaluate the results of modifying a specific visual concept on generated images.*
  - *Our method outperformed the baseline methods in the user study.*

- **Identity preserving evaluation.**

Model	Face verfication accuracy(%)
SimSwap (Chen et al. 2020)	87.40
FaceShifter [1] (Li et al. 2020)	85.45
FSGAN (Nirkin, Keller, and Hassner 2019)	89.20
Interpretable StyleGAN	90.25

  - *We performed a face verification experiment to evaluate the results of face swapping..*
  - *Our method was superior to other state-of-the-art face swapping methods for identity preserving.*

Quantitative analysis

- **Locality evaluation.**

Model	Mouth	Eyes	Chin
Editing in Style (Collins et al. 2020)	1.3649	0.9745	-
Feature Collaging (Suzuki et al. 2018)	0.1872	0.1293	0.0576
Interpretable StyleGAN	0.0606	0.0502	0.0163
Interpretable BigGAN	0.0296	0.0197	0.0311

- *To evaluate the locality of modifying a specific visual concept, we calculated the mean squared error (MSE) between the original images and the modified images.*
- *Our method had better localization, i.e. less change outside the region of a specific visual concept.*

- **Realism evaluation.**

Model	FID
StyleGAN, 128×128	12.86
Interpretable StyleGAN, 128×128	18.81
Interpretable StyleGAN†,128×128	19.42
BigGAN, 64×64	41.81
Interpretable BigGAN, 64×64	56.74
Interpretable BigGAN†, 64×64	57.72

- *We used the Fréchet Inception Distance (FID) to measure the realism of generated images.*
- *Forcing filters to encode disentangled visual concepts decreased the realism of generated images a bit.*

Contribution

We propose a generic method to modify a traditional GAN into an interpretable GAN without any annotations of visual concepts. In the interpretable GAN, each filter in an intermediate layer of the generator consistently generates the same localized visual concept when generating different images. Experiments show that our method can be applied to different types of GANs and enables people to modify a specific visual concept on generated images.

THANK YOU !