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
{lichao, yaokelu, wangjin20g, diaoboyu2012, xyj}@ict.ac.cn

Abstract

Learning a disentangled representation is still a challenge in the field of the interpretability of generative adversarial networks (GANs). This paper proposes a generic method to modify a traditional GAN into an interpretable GAN, which ensures that filters in an intermediate layer of the generator encode disentangled localized visual concepts. Each filter in the layer is supposed to consistently generate image regions corresponding to the same visual concept when generating different images. The interpretable GAN learns to automatically discover meaningful visual concepts without any annotations of visual concepts. The interpretable GAN enables people to modify a specific visual concept on generated images by manipulating feature maps of the corresponding filters in the layer. Our method can be broadly applied to different types of GANs. Experiments have demonstrated the effectiveness of our method.

Introduction

Recently, generative adversarial networks (GANs) have achieved huge success in generating high-resolution and realistic images (Brock, Donahue, and Simonyan 2018; Karras, Laine, and Aila 2019). In addition, the interpretability of GANs has attracted increasing attention in recent years. In this field, learning a disentangled representation is still a challenge to start-of-the-art algorithms. The disentangled representation of a GAN means that each component of the representation only affects a distinct aspect of a generated image. Previous studies on the disentanglement of GANs mainly focused on two perspectives. Some studies (Radford, Metz, and Chintala 2016; Chen et al. 2016) disentangled the attributes of images, such as the expression and eyeglasses of the generated human face images. Other studies (Zhu et al. 2017; Huang et al. 2018) disentangled the structure and texture of images. However, these works failed to provide clear and symbolic features for visual concepts in the intermediate layer of the generator.

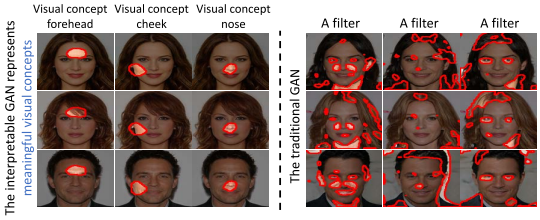

Figure 1: Compared with the traditional GAN, each filter in the interpretable GAN consistently represents a meaningful visual concept when generating different images. Different filters represent different visual concepts.

Therefore, we aim to propose a generic method to modify a traditional GAN into an interpretable GAN, which ensures that filters in an intermediate layer of the generator encode the disentangled and localized visual concepts (*e.g.* object parts like eyes, noses and mouths of human faces). Specifically, each filter in the intermediate layer is expected to consistently generate image regions corresponding to the same visual concept when generating different images. Different filters in the intermediate layer are expected to generate image regions corresponding to different visual concepts.

Learning the disentangled and localized visual concepts is of great value in both theory and practice. For example, Shen *et al.* (2020) enabled people to manipulate various facial attributes on the generated images through varying the latent codes. In contrast, this research enables people to modify a specific visual concept on generated images by manipulating feature maps of the corresponding filters, such as changing the appearance of a specific visual concept.

However, it still presents continuous challenges to ensure the learned visual concepts in the GAN have clear meanings, *i.e.* exploring the essence of meaningful visual concepts. To the best of our knowledge, there is no specific method to directly guarantee filters in an intermediate layer of the generator to encode meaningful visual concepts. In particular, we expect filters in the intermediate layer to automatically

*These authors contributed equally.
†Chao Li and Quanshi Zhang are the corresponding authors. Chao Li is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and the Zhejiang Laboratory, Hangzhou, China.

	Replacing		Attribute		Interpretable intermediate features	Moving part	Without annotations of facial semantics
	Object	Part	Global	Local			
Editing in Style (2020)	✗	✓	✗	✗	✗	✗	✓
MaskGAN (2020)	✗	✗	✓	✓	✗	✓	✗
InterFaceGAN (2020)	✗	✗	✓	✓	✗	✗	✗
MegaFS (2021)	✓	✗	✗	✗	✗	✗	✓
Label&Feature Collaging (2018)	✓	✓	✗	✓	✗	✓	✗
InfoSwap (2021a)	✓	✗	✓	✓	✗	✗	✗
RSGAN (2018)	✓	✗	✓	✓	✗	✗	✗
HifaFace (2021b)	✗	✗	✓	✓	✗	✗	✓
Age Embedding (2021)	✗	✗	✓	✗	✗	✗	✓
ELEGANT (2018)	✗	✗	✓	✓	✗	✗	✓
StyleFlow (2021)	✗	✗	✓	✓	✗	✗	✗
Facial Semantics (2021)	✗	✗	✓	✓	✗	✗	✓
NaviGAN (2021)	✗	✗	✗	✓	✗	✗	✓
Ours	✓	✓	✗	✓	✓	✓	✓

Table 1: Comparisons with other face-editing methods. The first column refers to replacing whole objects (*e.g.* faces) and object parts (*e.g.* noses of faces) on images. The second column refers to changing the global attributes (*e.g.* age) and local attributes (*e.g.* smiling) on images. The third column represents whether the method learns interpretable intermediate features. The fourth column refers to changing the location of parts on images. The fifth column represents whether the method requires annotation of facial semantics. Our method meets most of the requirements.

learn meaningful visual concepts without any manual annotations of visual concepts. It is because that such annotations usually represent human's understanding of images and can not reflect the representations inside the GAN.

In order to ensure the GAN learns meaningful visual concepts, we expect each filter in an intermediate layer of the generator to consistently represent the same visual concept across different images. We notice that a specific visual concept is usually represented by multiple filters in an intermediate layer of the generator. In this way, we divide filters in the intermediate layer into different groups and assume that different groups represent different visual concepts. Specifically, we expect filters in the same group to consistently generate image regions corresponding to the same visual concept when generating different images. Note that filters in the same group are expected to represent almost the entire visual concept rather than sub-parts of the visual concept, which ensures the clarity of the visual concept represented by each filter.

Furthermore, it is also crucial to ensure the strictness of explanation results. In other words, if a filter represents a certain visual concept, then neural activations in the feature map of this filter should exclusively correspond to this visual concept without any noise activations in other unrelated regions. To this end, we propose a probability model to measure the fitness between explanation results and the neural activations in the feature maps. Specifically, the probability model is formulated as an energy-based model. The input of the energy-based model is the feature maps in the target layer. The output is a probability of neural activations in feature maps corresponding to visual concepts. High probability represents that neural activations in each feature map of the filters correspond to a clear visual concept. Low probability represents vice versa. In this way, we expect to train the energy-based model to learn to evaluate the feature maps in the target layer. Then, this energy-based model is used to refine the representations inside the target layer of the GAN.

In this study, we evaluate our interpretable GANs both qualitatively and quantitatively. For qualitative evaluation, we visualize the feature map of each filter to evaluate the consistency of the visual concept represented by each filter through different images. For quantitative evaluation, we evaluate the results of modifying visual concepts on generated images, in order to show the correctness and locality of the modification for a specific visual concept. Besides, we also evaluate the realism of generated images both qualitatively and quantitatively.

Contributions of this paper can be summarized as follows. 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.

Related Work

Disentanglement of GANs. Previous works have mainly explored the disentanglement of GANs from two perspectives. Several works (Radford, Metz, and Chintala 2016; Chen et al. 2016; Härkönen et al. 2020; Shen and Zhou 2021; Voynov and Babenko 2020; Wu, Lischinski, and Shechtman 2021) focused on disentangling the attributes of the generated images. Shen *et al.* (2020) disentangled the gender, age and expression of the generated human faces. Jahanian *et al.* (2019) and Plumerault *et al.* (2019) disentangled simple transformations of the generated images, such as translation and zooming, to control the image generation of GANs. Other works (Zhu et al. 2017; Huang et al. 2018) focused on

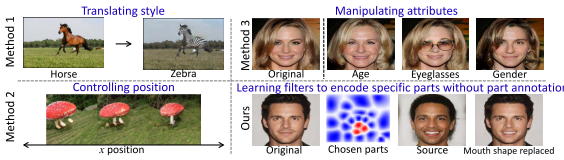

Figure 2: Visual comparisons with other methods for interpretable GANs. These methods focus on different types of interpretability. Method 1 (2017) disentangled the structure and the style of the image. Method 2 (2019) learned features for the localized object in the image. Method 3 (2020) learned the disentangled features for attributes of the image. In contrast, our method learns each filter to encode an object part without part annotations. To the best of our knowledge, no other GANs can ensure such part interpretability.

disentangling the structure and texture of the generated images. FineGAN (Singh, Ojha, and Lee 2019) disentangled the object shape, background and object appearance (color/-texture) to generate images. Ma *et al.* (2018) disentangled foreground, background and pose information to generate images of persons. However, these studies failed to provide clear and symbolic representations for visual concepts in the generation of images. Bau *et al.* (2018) identified a group of filters closely related to objects and object parts, but required supervision from a semantic segmentation model. Collins *et al.* (2020) exploited to disentangle object parts of the generated images without external supervision, but did not ensure each filter represented a clear meaning. In contrast, our method ensures each filter represents a localized visual concept without human supervision.

Face editing with GANs. Previous methods on face editing were mainly conducted in the image-to-image settings (Shen and Liu 2017; Zhang et al. 2018; Richardson et al. 2021). Fader networks (Lample et al. 2017) learned to vary the values of attributes to change the attributes of the generated images. StarGAN (Choi et al. 2018) learned to perform image-to-image translations across multiple domains using a single model to edit different attributes of human face images. However, these methods could not edit images with exemplars. To this end, ELEGANT (Xiao, Hong, and Ma 2018) learned to transfer attributes between two images by exchanging their latent codes. MaskGAN (Lee et al. 2020) exploited diverse manipulations of human face images by modifying masks of target images according to the source images. However, these methods required supervision from annotated attributes or masks. In comparison, our method modifies a specific visual concept according to other generated images without manual annotations of visual concepts.

Algorithm

Given training images without annotations of visual concepts, we aim to train an interpretable GAN in an end-to-end manner. Specifically, given a target convolutional layer of the generator, we expect each filter in this layer to represent a meaningful visual concept (*e.g.* object parts like eyes, noses and mouths of human faces). In other words, each filter in the target layer is expected to consistently generate the same visual concept when generating different images.

The key challenge is to ensure that each filter in the target layer of the GAN represents a meaningful visual concept. To this end, we notice that multiple filters usually represent a certain visual concept, when they generate similar image regions corresponding to this visual concept. This phenomenon was also discussed in (Shen et al. 2021) for filters in convolutional neural networks (CNNs). Therefore, we divide filters in the target layer into different groups, which represent different visual concepts respectively. We expect filters in the same group to represent the same visual concept. Let M denote the number of filters in the target layer. In this way, we divide M filters in the target layer into C groups. Let $q^j \in \{1, 2, \cdots, C\}$ denote the index of the group, where the j-th filter belongs across different images. $\mathbf{Q} = \{q^1, q^2, \cdots, q^M\}$ denotes the partition of filters. Let G denote the generator of the GAN. To encourage each filter in the target layer to represent a meaningful visual concept, we aim to optimize the generator G and the partition $\mathbf{Q}$ to force filters in the same group to generate the same image region on a generated image.

In addition to ensuring each filter represents a meaningful visual concept, it is also important that the generator of the GAN generates realistic images. In this way, we design the following loss function to train the interpretable GAN:

$$\mathbf{L} = \mathcal{L}_{GAN}(G, D) + \lambda_0 Loss(\mathbf{Q}, G) \quad (1)$$

where λ_0 denotes a positive weight; $\mathcal{L}_{GAN}(G, D)$ denotes the traditional GAN loss function (Goodfellow et al. 2014; Gulrajani et al. 2017), where D denotes the discriminator of the GAN; $Loss(\mathbf{Q}, G)$ is the interpretability loss to encourage each filter in the target layer to represent a meaningful visual concept, which will be introduced later.

Learning the partition Q. Given the generator G, we expect to learn the partition $\mathbf{Q}$ to ensure filters in the same group generate similar image regions. In other words, we expect feature maps in each group have similar neural activations. To this end, we use a Gaussian mixture model (GMM) to learn the partition $\mathbf{Q}$ for feature maps in the target layer. Let $\{z_i\}_{i=1}^N$ denote the set of N input latent vectors, which generate N different images through the generator G. Given the i-th latent vector $z_i \in R^d$, let $f_G(z_i) = [f_i^1, f_i^2, \cdots, f_i^j, \cdots, f_i^M]$ denote the feature map in the target convolutional layer of the generator G after the ReLU operation. Here $f_i^j \in R^K$ denotes the feature map of the j-th filter. Then, let $F^j = [f_1^j, f_2^j, \cdots, f_N^j]$ denote the feature maps of the j-th filter given the set of N input latent vectors $\{z_i\}_{i=1}^N$. The Gaussian mixture model is formulated as $P_\Theta(F^j)$, where Θ denotes the model parameters.

The key challenge is to optimize the GMM parameters Θ to learn the partition $\mathbf{Q}$. Specifically, we take the j-th filter's group index q^j as a latent variable and $P_\Theta(F^j)$ estimates the likelihood of the j-th filter's feature maps belonging to any group c, *i.e.* $P_\Theta(F^j) = \sum_c P_\Theta(F^j, q^j = c)$. In this way, we have $P_\Theta(F^j, q^j = c) = P_\Theta(q^j = c) P_\Theta(F^j | q^j = c)$. We define $P_\Theta(q^j = c) = p_c$, where p_c denotes the prior probability of the c-th group. $P_\Theta(F^j | q^j = c)$ denotes the probability of the j-th filter's feature maps having similar neural

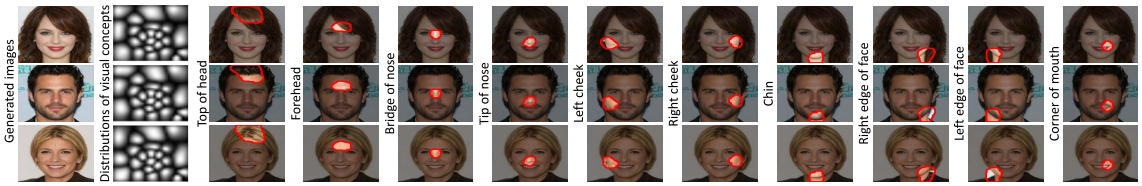

Figure 3: Visualization of feature maps in interpretable GANs based on the method in (Zhang, Wu, and Zhu 2018). The first column shows the generated images. The second column shows the visualization of the distributions of visual concepts encoded in an intermediate layer filters. Each remaining column in the figure corresponds to a certain filter.

activations with feature maps in the c-th group. To simplify the calculation of $P_\Theta(F^j|q^j = c)$, we assume that when the j-th filter belongs to the c-th group, the probabilities of the j-th filter's feature maps across different images are independent of each other, *i.e.* $P_\Theta(F^j|q^j = c) = \prod_{i=1}^{N} P_\Theta(f_i^j|q^j = c)$. We assume that $f_i^j|(q^j = c) \sim \mathcal{N}(\mu_c, \sigma_c^2 I)$. Here I denotes the identity matrix. To learn the model parameters $\{p_c, \mu_c, \sigma_c^2\} \in \Theta$, we design the following loss.

$$\max_{\Theta} \mathcal{L}_{GMM} = \max_{\Theta} \sum_{j=1}^{M} log P_\Theta(F^j) \tag{2}$$

Let Θ' denote the optimal Θ for equation (2). In this way, the optimal partition $\mathbf{Q}$ is solved as $\mathbf{Q} = \{q_j | \arg\max_{q^j} P_{\Theta'}(q^j|F^j)\}$.

Realism of generated images. Given the partition $\mathbf{Q}$ for filters in the target layer, forcing each filter in the same group to exclusively generate the same visual concept may decrease the realism of the generated images, even with the help of the discriminator D. To this end, we use an energy-based model (Gao et al. 2018; Nijkamp et al. 2019) that outputs a probability of the realism of the feature maps $f_G(z)$ in the target layer. Specifically, the energy-based model outputs a probability of feature maps, which generate realistic images. In this way, we can conclude that feature maps $f_G(z)$ with high realism can generate realistic images. The energy-based model is formulated as $P_W(f_G(z)|\mathbf{Q})$, where W denotes the model parameters. To increase the realism of the images generated from the feature maps $f_G(z)$, we use the following loss to learn the energy-based model $P_W(f_G(z)|\mathbf{Q})$ via maximizing the log-likelihood.

$$\mathcal{L}_{real}(W, G) = -\frac{1}{N} \sum_{i=1}^{N} log P_W(f_G(z_i)|\mathbf{Q}) \tag{3}$$

To measure the realism of the feature maps in the target layer, the energy-based model $P_W(f_G(z)|\mathbf{Q})$ is designed as follows.

$$P_W(f_G(z)|\mathbf{Q}) = \frac{1}{Z(W)} exp\Big(g_W(f_G(z))\Big) P_0(z) \tag{4}$$

where $Z(W) = \int exp(g_W(f_{G'}(z)))P_0(z)dz$ is used for normalization. Here we consider G' as the current generator with fixed parameters for calculating $Z(W)$, in order to learn the parameters W. $P_0(z)$ denotes the Gaussian distribution, *i.e.* $P_0(z) \sim \mathcal{N}(0, \sigma_0^2 I)$. $g_W(f_G(z))$ denotes the metric, which measures the realism of the feature maps in the target layer. Specifically, we have $g_W(f_G(z)) = \sum_{j=1}^{M} \sum_{c=1}^{C} [W_{jc} \cdot (f^j \odot \bar{f}^c)]$, where $\cdot$ denotes the inner product and $\odot$ denotes the element-wise product. $W \in R^{M \times C \times K}$ denotes the parameters of the energy-based model. $f^j \in R^K$ denotes the feature map of the j-th filter. $\bar{f}^c \in R^K$ denotes the c-th group center of feature maps, which can be computed as the mean of feature maps in the c-th group.

Interpretability of filters in the target layer. In order 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. In other words, when the j-th filter belongs to the c-th group, we expect the j-th filter's feature map f^j to be close to the group center $\bar{f}^c$. Besides, we also consider the diversity of visual concepts represented by different filters. To this end, when the j-th filter does not belong to the c-th group, we expect the j-th filter's feature map f^j to be different from the group center $\bar{f}^c$. In this way, we design the following loss.

$$\begin{aligned} \mathcal{L}_{interp}(W) = &- \sum_{j=1}^{M} \sum_{c=1}^{C} \sum_{k=1}^{K} \mathcal{I}(q_j = c) W_{jck} \\ &+ \lambda_1 \sum_{j=1}^{M} \sum_{c=1}^{C} \sum_{k=1}^{K} \mathcal{I}(q_j \neq c) W_{jck} \end{aligned} \tag{5}$$

where λ_1 denotes a positive weight; $\mathcal{I}(\cdot)$ is the indicator function. In this way, when the j-th filter belongs to the c-th group, the metric $g_W(f_G(z))$ forces the j-th filter's feature map f_j and the c-th group center $\bar{f}^c$ to have neural activations in similar positions by pushing W_{jck} to be positive. Otherwise, $g_W(f_G(z))$ forces f_j and $\bar{f}^c$ to have neural activations in different positions by pushing W_{jck} to be negative. Please see Fig. 4 for more details.

To sum up, $Loss(W, \mathbf{Q}, G)$ is designed as follows:

$$\begin{aligned} Loss(W, \mathbf{Q}, G) = &\sum_{q_j \in \mathbf{Q}} P_{\Theta'}(q^j|F^j) + \lambda_2 \mathcal{L}_{real}(W, G) \\ &+ \lambda_3 \mathcal{L}_{interp}(W) \end{aligned} \tag{6}$$

where λ_2 and λ_3 are positive weights. $\sum_j P_{\Theta'}(q^j|F^j)$ is designed to learn the partition $\mathbf{Q}$ for filters. $\mathcal{L}_{real}(W, G)$ and $\mathcal{L}_{interp}(W)$ are designed to increase the realism of the generated images and the interpretability of the filters in the tar-

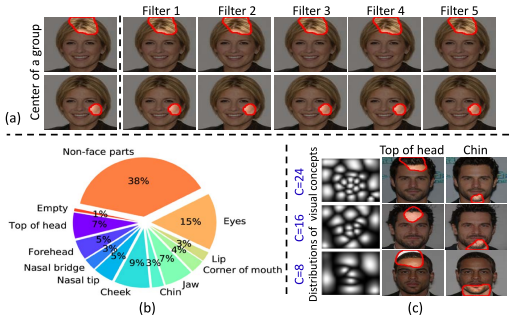

Figure 4: (a) Comparisons of receptive fields (RFs) between the center of a group and each filter in the group. (b) Proportions of filters representing different visual concepts. (c) Filters learned with different values of C.

get layer. The overall loss is optimized as follows.

$$\min_{W,G} \max_{D,\mathbf{Q}} \mathbf{L} \tag{7}$$

Learning. Given the partition of filters $\mathbf{Q}$, we optimize $Loss(W, \mathbf{Q}, G)$ *w.r.t.* W, G for once after optimizing $\mathcal{L}_{GAN}(G, D)$ *w.r.t.* G, D for T times. However, 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 (Girolami and Calderhead 2011; Zhu and Mumford 1998). Specifically, following the method in (Gao et al. 2018), the gradient of $\mathcal{L}_{real}(W, G)$ *w.r.t.* W is approximately calculated as follows.

$$\begin{aligned} &\frac{\partial}{\partial W}\mathcal{L}_{real}(W, G) \\ \approx &\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial W}g_W(f_{G'}(\hat{z}_i)) - \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial W}g_W(f_G(z_i)) \end{aligned} \tag{8}$$

where $\{\hat{z}_i\}_{i=1}^N$ denotes the revised latent vectors sampled from Langevin dynamics. The iterative process of Langevin dynamics is carried out as follows.

$$z^{\tau+1} = z^{\tau} + \frac{\delta^2}{2}\frac{\partial}{\partial z}P_W(f_G(z^{\tau})|\mathbf{Q}) + \delta U^{\tau} \tag{9}$$

where τ denotes time steps; δ denotes step size; $U^{\tau} \sim N(0, I)$ is a Gaussian noise. In this way, equation (9) and (8) are calculated alternately to update the energy-based model parameters W.

Experiments

We applied our method to two state-of-the-art GANs trained on two different datasets. For qualitative evaluation, we visualized feature maps of filters to show the consistency of the visual concept represented by each filter. We also visualized the results of modifying specific visual concepts on generated images. Besides, we demonstrated that performing Langevin dynamics could improve the realism of some bad generated images and modified images. For quantitative evaluation, we conduct a user study and a face verification

Figure 5: Exchanging a specific visual concept between the original images and the source images. The second column shows chosen parts for exchanging, which are marked in red. The fifth column shows the mean squared-error heatmaps between the original images and the modified images.

experiment to examine the correctness of exchanging a specific visual concept and faces between pairs of images. We also calculated the mean squared-error (MSE) between original images and modified images in terms of a certain visual concept, in order to evaluate the locality of our modifications. We calculated the Fréchet Inception Distance (FID) (Heusel et al. 2017) to measure the realism of generated images. Experiments show that our method successfully disentangled localized visual concepts encoded in filters of the generator.

Models and datasets. We applied our method to two different GANs, BigGAN (Brock, Donahue, and Simonyan 2018) and StyleGAN (Karras, Laine, and Aila 2019). BigGAN was trained on FFHQ dataset (Karras, Laine, and Aila 2019). StyleGAN (Karras, Laine, and Aila 2019) was trained on CelebA-HQ dataset (Karras et al. 2018).

Implementation details. We set hyperparameters as $C = 24$, $\lambda_0 = 1$ and $\lambda_1 = \frac{2}{3}$. Since $\mathcal{L}_{real}(W, G)$ was used to update two seperate models, *i.e.* the generator and the energy-based model, we set λ_2 different values to update different models. Specifically, we set $\lambda_2 = 1$ for updating the energy-based model parameters W. For updating the generators of BigGAN and StyleGAN, we set $\lambda_2 = 0.1$ and $\lambda_2 = 0.05$ respectively. To ensure the interpretability of filters, we expected that $\mathcal{L}_{interp}$ dominated the learning process in the early stage. To this end, for StyleGAN, λ_3 was set to be $3e^{-2}$ at first and exponentially decayed to $3e^{-6}$ during 1000 batches. For BigGAN, λ_3 was set the same but exponentially decayed during 500 batches. We set $T = 50$ for BigGAN and $T = 100$ for StyleGAN. We initialized each dimension of parameters W to be zero. We used the learning rate of 10,

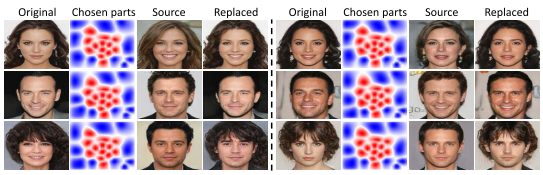

Figure 6: Swapping whole faces between the original images and the source images. The second column shows the chosen parts for swapping, which are marked in red. The fourth column shows the replaced images.

Figure 7: (a) Improving the realism of generated images by Langevin dynamics. Each column shows the generated images by doing the iterative process of Langevin dynamics τ steps. (b) Improving the realism of modified images by Langevin dynamics. The third column shows the replaced images.

SGD optimizer for parameters W.

Learning Interpretable GANs

Learning an interpretable BigGAN. We learned an interpretable GAN based on the BigGAN architecture to generate images with the size of 64×64. We first trained the BigGAN following the experiment settings in (Brock, Donahue, and Simonyan 2018). Then, we added our proposed loss $Loss(W, \mathbf{Q}, G)$ to an intermediate layer of the generator to fine-tune BigGAN, where the size of feature maps $f_G(z)$ is 32×32. To be clear, we only fine-tuned the generator of the BigGAN. The discriminator of the BigGAN was reinitialized and trained from scratch. It was because that the discriminator usually converged faster than the generator in BigGAN.

Learning an interpretable StyleGAN. We learned an interpretable GAN based on the StyleGAN architecture to generate images with the size of 128×128. We first trained the StyleGAN following the experiment settings in (Karras, Laine, and Aila 2019). We noticed that the activation functions in the generator were all leaky-ReLU (Maas et al. 2013). To this end, we added a ReLU layer after an intermediate layer of the generator, where the size of feature maps $f_G(z)$ is 32×32. Then, we added our proposed loss $Loss(W, \mathbf{Q}, G)$ to the output of the added ReLU layer. The generator and the discriminator of the StyleGAN were jointly fine-tuned, because they were progressively trained in (Karras, Laine, and Aila 2019).

Qualitative Evaluation

Visualization of feature maps. Based on the method in (Zhang, Wu, and Zhu 2018), we visualized the receptive fields (RFs) corresponding to a filter's feature maps, which were scaled up to the image resolution. Fig. 3 shows the RFs of filters in our interpretable GANs. In our interpretable GANs, each filter consistently generated image regions corresponding to the same visual concept when generating different images. Different filters generated image regions corresponding to different visual concepts. We also compared RFs between the group center and filters in this group, as shown in Fig. 4 (a). Moreover, we explored the number of visual concepts represented by filters in our interpretable GAN. Fig. 4 (b) illustrates the proportions of filters representing different visual concepts when setting $C = 24$. Results show that 512 filters totally represented 11 visual concepts. Besides, as shown in Fig. 4 (c), when setting different values of C, GANs with a larger value of C learned more detailed concepts.

Modifying visual concepts on images. Our interpretable GAN enabled us to modify specific visual concepts on generated images. For example, we exchanged a specific visual concept between pairs of images by exchanging the corresponding feature maps in the target layer (*i.e.* the convolutional layer that was modified to an interpretable layer). Fig. 5 shows the results of exchanging the mouth, hair and nose between pairs of images. Note that our method only changed the shape of a specific visual concept. For StyleGAN, the color of a specific visual concept was mainly controlled by styles in higher-resolution layers, as discussed in (Karras, Laine, and Aila 2019). Fig. 5 also shows the difference between the modified images and the original images, where at every pixel location we calculated the squared distance in RGB space. Our method only modified a localized visual concept without changing other unrelated regions. Besides, we also exchanged whole faces between pairs of images, as shown in Fig. 6.

Improving the realism of images. To improve the realism of some bad generated images, we used Langevin dynamics to sample revised latent vectors. As shown in Fig. 7 (a), revised latent vectors sampled from Langevin dynamics generated more realistic images than the original latent vectors.

Besides, we also performed Langevin dynamics to improve the realism of modified images. Specifically, given two latent vectors z_a and z_b, we exchanged a certain group of feature maps between $f_G(z_a)$ and $f_G(z_b)$. Let $f(z_a)'$ and $f(z_b)'$ denote the exchanged feature maps of z_a and z_b. Then, we performed Langevin dynamics to sample revised latent vectors. Specifically, z_a was updated as follows: $z_a^{\tau+1} = z_a^{\tau} + \frac{\delta^2}{2}\frac{\partial}{\partial z_a}(P_W(f(z_a^{\tau})'|\mathbf{Q}) + P_W(f(z_b^{\tau})'|\mathbf{Q})) + \delta U^{\tau}$. z_b was updated in the same way. In this way, the exchanged feature maps $f(z_a)'$ and $f(z_b)'$ had higher probabilities and could generate more realistic images. Fig. 7 (b) shows the results of the modified images after performing Langevin dynamics.

Quantitative Analysis

Human perception evaluation. We conduct a user study to evaluate the results of modifying a specific visual concept on generated images. Specifically, we exchanged the mouth, chin and eyes between pairs of images as three tasks.

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

Table 2: Human evaluation scores.

Model	Face verfication accuracy(%)
SimSwap (2020)	87.40
FaceShifter [1] (2020)	85.45
FSGAN (2019)	89.20
Interpretable StyleGAN	90.25

Table 3: Face verfication accuracy. All methods was tested by images generated by our Interpretable StyleGAN.

We randomly chose 200 pairs of test images for each task respectively. For each task, given an original image and a modified image, 10 volunteers were asked to choose which image contained the exchanged visual concept on the modified image among four choices. Table 2 shows the results of human evaluation scores. Each score represents the average percentage of the correctly-answered questions among all volunteers. We used the methods proposed in (Collins et al. 2020) and (Suzuki et al. 2018) as baselines. Our method outperformed the above methods in the user study.

Identity preserving evaluation. We performed a face verification experiment to evaluate the results of face swapping. For one pair of images, we replaced the face of the original image with the face of the source image to generate the modified image. Then we tested whether the face of the modified image and the face of the source image were of the same identity. Specifically, we selected 2K pairs of faces and used ArcFace (Deng et al. 2019) (99.52% on LFW (Huang et al. 2008)) to test the results. Table 3 shows the accuracy of the face verification. Our method was superior to other state-of-the-art face swapping methods for identity preserving.

Locality evaluation. 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 in RGB space. Specifically, we manually annotated segmentation masks for specific visual concepts on 100 generated images respectively. Then, we measure the ratio of the Out-MSE and In-MSE for each pair of images, *i.e.* the MSE outside the region of a specific visual concept and MSE inside the region of a specific visual concept. Let $x \in R^D$ and $x' \in R^D$ denote the original image and the modified image. $G^c(x) \in \{0,1\}^D$ denote the hand-annotated segmentation mask of the c-th visual concept on image x ($c = 1, \cdots, C$). $\hat{G}^c(x) \in \{0,1\}^D$ denotes the reverse mask, *i.e.* $\hat{G}^c_u(x) = \mathcal{I}(G^c_u(x) = 0)$, where $\mathcal{I}(\cdot)$ is the indicator function ($u = 1, \cdots, D$). The In-MSE and Out-MSE for the c-th visual concept is calculated as follows: $In - MSE_c = \frac{\sum_{u=1}^{D} G^c_u(x)(x_u - x'_u)^2}{\sum_{u=1}^{D} G^c_u(x)}$, $Out - MSE_c = \frac{\sum_{u=1}^{D} \hat{G}^c_u(x)(x_u - x'_u)^2}{\sum_{u=1}^{D} \hat{G}^c_u(x)}$. The locality metric of the modification for the c-th visual concept is calculated as follows: $Locality_c = \frac{Out - MSE_c}{In - MSE_c}$. A small number of this metric indicates that our modification mainly changes the regions related to a specific visual concept. Table 4 shows the results of our locality metric for each visual concept. Our method had better localization, *i.e.* less change outside the region of a specific visual concept.

[1] Using code in https://github.com/denis19973/faceshifter_tornado, because the original paper has not released the code yet.

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

Table 4: Locality 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

Table 5: Fréchet Inception Distance (FID) between ground truth images and generated images of GANs. † represents performing Langevin dynamics on generated images.

Realism evaluation. To measure the realism of generated images, we used the Fréchet Inception Distance (FID) (Heusel et al. 2017), which compares the distribution of two sets of images in the feature space of a deep CNN layer. The smaller FID is, the more realistic generated images usually are. Table 5 shows the results of FID between the ground truth images and 50K generated images of GANs. This table indicates that forcing filters to encode disentangled visual concepts decreased the realism of generated images a bit. Surprisingly, performing Langevin dynamics achieved worse results, although Fig. 7 shows qualitatively that the realism of generated images was improved through Langevin dynamics. This reemphasizes that correctly and automatically measuring the realism of generated images is still difficult.

Conclusion

In this paper, we have proposed a generic method to modify a traditional GAN into an interpretable GAN, which forces each filter in an intermediate layer of the generator to represent a meaningful visual concept. Specifically, we design a loss to push each filter in the intermediate layer to consistently generate image regions corresponding to the same visual concept when generating different images, and different filters to generate image regions corresponding to different visual concepts. Experiments have demonstrated that our method enables people to modify a specific visual concept on generated images, such as changing the appearance of this visual concept.

Acknowledgments

This work is partially supported by the National Nature Science Foundation of China (No. 61906120, U19B2043), Shanghai Natural Science Fundation (21JC1403800,21ZR1434600), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Key Research Project of Zhejiang Lab (No. 2021PC0AC02). Besides Dr. Chao Li, Dr. Quanshi Zhang is also a corresponding author. He is with the John Hopcroft Center and the MoE Key Lab of Artificial Intelligence, AI Institute, at the Shanghai Jiao Tong University, China.

References

Abdal, R.; Zhu, P.; Mitra, N. J.; and Wonka, P. 2021. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3): 1–21.

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Zhou, B.; Tenenbaum, J. B.; Freeman, W. T.; and Torralba, A. 2018. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. In *International Conference on Learning Representations*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *MM '20: The 28th ACM International Conference on Multimedia*, 2003–2011. ACM.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2180–2188.

Cherepkov, A.; Voynov, A.; and Babenko, A. 2021. Navigating the gan parameter space for semantic image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3671–3680.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Collins, E.; Bala, R.; Price, B.; and Susstrunk, S. 2020. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5771–5780.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Gao, G.; Huang, H.; Fu, C.; Li, Z.; and He, R. 2021a. Information Bottleneck Disentanglement for Identity Swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3404–3413.

Gao, R.; Lu, Y.; Zhou, J.; Zhu, S.-C.; and Wu, Y. N. 2018. Learning generative convnets via multi-grid modeling and sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9155–9164.

Gao, Y.; Wei, F.; Bao, J.; Gu, S.; Chen, D.; Wen, F.; and Lian, Z. 2021b. High-Fidelity and Arbitrary Face Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16115–16124.

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.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved Training of Wasserstein GANs. In *NIPS*.

Härkönen, E.; Hertzman, A.; Lehtinen, J.; and Paris, S. 2020. GANSpace: Discovering Interpretable GAN controls. In *IEEE Conference on Neural Information Processing Systems;*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, 172–189.

Jahanian, A.; Chai, L.; and Isola, P. 2019. On the" steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Lample, G.; Zeghidour, N.; Usunier, N.; Bordes, A.; Denoyer, L.; and Ranzato, M. 2017. Fader networks: Generating image variations by sliding attribute values. In *Advances in Neural Information Processing Systems*, 5963–5972.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5549–5558.

Li, L.; Bao, J.; Yang, H.; Chen, D.; and Wen, F. 2020. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5074–5083.

Li, Z.; Jiang, R.; and Aarabi, P. 2021. Continuous Face Aging via Self-estimated Residual Age Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15008–15017.

Ma, L.; Sun, Q.; Georgoulis, S.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 99–108.

Maas, A. L.; Hannun, A. Y.; Ng, A. Y.; et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, 3. Citeseer.

Natsume, R.; Yatagawa, T.; and Morishima, S. 2018. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*.

Nijkamp, E.; Hill, M.; Zhu, S.-C.; and Wu, Y. N. 2019. Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model. *NeurIPS 2019*.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7184–7193.

Plumerault, A.; Le Borgne, H.; and Hudelot, C. 2019. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations*.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2287–2296.

Shen, W.; and Liu, R. 2017. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4030–4038.

Shen, W.; Wei, Z.; Huang, S.; Zhang, B.; Fan, J.; Zhao, P.; and Zhang, Q. 2021. Interpretable Compositional Convolutional Neural Networks. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2971–2978. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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.

Shen, Y.; and Zhou, B. 2021. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1532–1540.

Singh, K. K.; Ojha, U.; and Lee, Y. J. 2019. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6490–6499.

Suzuki, R.; Koyama, M.; Miyato, T.; Yonetsuji, T.; and Zhu, H. 2018. Spatially controllable image synthesis with internal representation collaging. *arXiv preprint arXiv:1811.10153*.

Voynov, A.; and Babenko, A. 2020. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, 9786–9796. PMLR.

Wu, Z.; Lischinski, D.; and Shechtman, E. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12863–12872.

Xiao, T.; Hong, J.; and Ma, J. 2018. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, 168–184.

Zhang, G.; Kan, M.; Shan, S.; and Chen, X. 2018. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)*, 417–432.

Zhang, Q.; Wu, Y. N.; and Zhu, S.-C. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8827–8836.

Zheng, Y.; Huang, Y.-K.; Tao, R.; Shen, Z.; and Savvides, M. 2021. Unsupervised Disentanglement of Linear-Encoded Facial Semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3917–3926.

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.

Zhu, S. C.; and Mumford, D. 1998. Grade: Gibbs reaction and diffusion equations. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, 847–854. IEEE.

Zhu, Y.; Li, Q.; Wang, J.; Xu, C.-Z.; and Sun, Z. 2021. One Shot Face Swapping on Megapixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4834–4844.