FEditNet: Few-shot Editing of Latent Semantics in GAN Spaces

Mengfei Xia, [1] **Yezhi Shu,** [1] **Yuji Wang,** [1] **Yu-Kun Lai,** [2]
Qiang Li, [3] **Pengfei Wan,** [3] **Zhongyuan Wang,** [3] **Yong-Jin Liu** [1*]
[1]Department of Computer Science, BNRist, Tsinghua University
[2]School of Computer Science and Informatics, Cardiff University
[3]Kuaishou Technology
{xmf20, shuyz19, yuji-wan20}@mails.tsinghua.edu.cn, LaiY4@cardiff.ac.uk
{liqiang03, wanpengfei,wangzhongyuan}@kuaishou.com, liuyongjin@tsinghua.edu.cn

Abstract

Generative Adversarial networks (GANs) have demonstrated their powerful capability of synthesizing high-resolution images, and great efforts have been made to interpret the semantics in the latent spaces of GANs. However, existing works still have the following limitations: (1) the majority of works rely on either pretrained attribute predictors or large-scale labeled datasets, which are difficult to collect in most cases, and (2) some other methods are only suitable for restricted cases, such as focusing on interpretation of human facial images using prior facial semantics. In this paper, we propose a GAN-based method called FEditNet, aiming to discover latent semantics using very few labeled data without any pretrained predictors or prior knowledge. Specifically, we reuse the knowledge from the pretrained GANs, and by doing so, avoid overfitting during the few-shot training of FEditNet. Moreover, our layer-wise objectives which take content consistency into account also ensure the disentanglement between attributes. Qualitative and quantitative results demonstrate that our method outperforms the state-of-the-art methods on various datasets including CelebA, FFHQ and LSUN.

Introduction

Based on a min-max game between a generator and a discriminator, Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) learn a nonlinear mapping from a random distribution to the domain of real data, which have exhibited high fidelity in synthesizing high-resolution images (Karras, Laine, and Aila 2019; Karras et al. 2020; Brock, Donahue, and Simonyan 2019).

Despite great success in achieving amazing quality of synthesized images, it remains unclear how GANs construct semantics in the latent space during training. To address this issue, some recent works have made efforts to interpret the meaningful semantic and disentanglement properties in the latent spaces of GANs (Shen et al. 2020; Shen and Zhou 2021; Plumerault, Borgne, and Hudelot 2020). One representative strategy to analyze the semantics is to add a learned direction on the latent code (Radford, Metz, and Chintala 2016). This strategy makes use of geometrical properties of hyperplanes in latent spaces, which opens the door to exploring interpretable latent spaces.

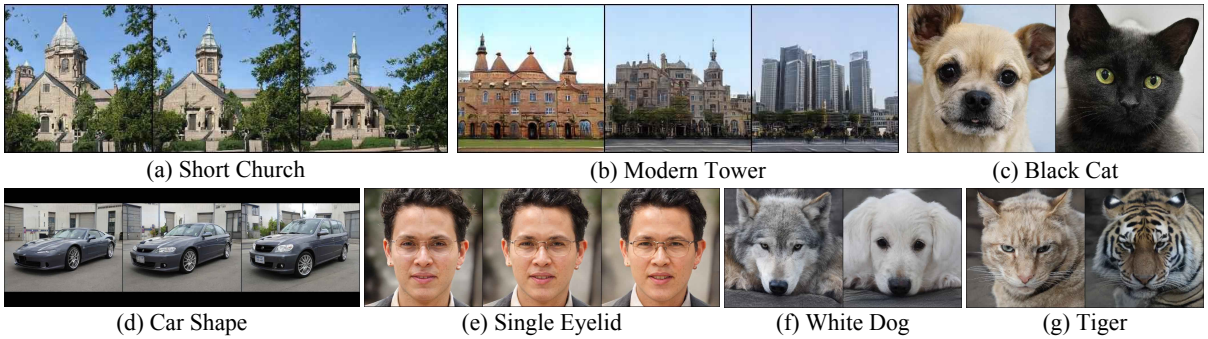
*Corresponding author.

With this strategy, two types of methods have been developed to interpret latent spaces of GANs: unsupervised and supervised. To formulate the latent semantics, a representative unsupervised way is to use principal component analysis (PCA) (Shen and Zhou 2021; Härkönen et al. 2020; Zhu et al. 2022). These methods apply PCA on either the model parameters or feature maps. Then by choosing the eigenvectors associated with the largest few eigenvalues, the achieved directions have the most significant effects on the synthesized images and semantics. However, they cannot precisely edit the user-desired attributes. As a comparison, the directions found by supervised methods usually have better controllability and disentanglement property. The cost paid for this advantage is that most supervised methods use either a pretrained attribute assessor (Goetschalckx et al. 2019; Zhuang, Koyejo, and Schwing 2021) or a large amount of data containing the target attributes to learn the semantics in the latent space (Yang, Shen, and Zhou 2021; Yang et al. 2021b; Shen et al. 2020), which restricts applications of supervised methods, such as the scenarios where labeling a lot of data is infeasible and unlabeled datasets have to be used.

In this paper, we aim to explore interpretable semantics using limited labeled data. To achieve this goal, we propose a GAN-based method called *FEditNet* (**F**ew-shot Attribute **Edit**ing **Net**work), which can perform efficient attribute editing and latent space disentanglement by training with very few data samples. Unlike previous supervised methods which use a pretrained assessor or train attribute assessor from scratch with extra efforts, our method reuse the knowledge from the discriminator (which is pretrained to compete with the given generator) to train the assessor (Yang et al. 2021a). To further ensure the disentanglement from undesired attributes, we introduce a feature contrastive loss as the regularizer to fix the attributes that users do not want to edit. Therefore, our FEditNet is able to achieve high attribute disentanglement and state-of-the-art editing quality with few-shot data, avoiding the vulnerablity and overfitting in few-shot learning. Some results are shown in Fig. 1.

In this paper, we make three contributions:

- We propose a learning method with limited training data, to explore interpretable latent spaces of a pretrained GAN and the editing direction of desired attributes.
- We introduce a feature contrastive loss as a regularizer in training to facilitate disentanglement in latent semantics,



(a) Short Church (b) Modern Tower (c) Black Cat

(d) Car Shape (e) Single Eyelid (f) White Dog (g) Tiger

Figure 1: Interpretable latent semantics of latent spaces in StyleGAN2 (Karras et al. 2020) discovered from very few data samples by our proposed FEditNet. In images (a), (b), (d) and (e), the middle one is the original synthesis while the left and right ones are images from latent codes moving in the forward and backward directions respectively. In each of images (c), (f) and (g), the left one is the original image while the right is the editing result along the direction. The subfigure captions present the intended attributes for manipulation.

which helps to learn the internal structure of latent space.

- We study the layer-wise effects of GANs using the feature contrastive loss, and achieve further performance improvement. We also propose to apply the Bayesian optimization for adaptive loss weights in feature contrastive loss, which can achieve good performance while maintaining time efficiency.

Related Work

Generative Adversarial Networks

GAN (Goodfellow et al. 2014) has played a significant role in promoting the development of image synthesis (Brock, Donahue, and Simonyan 2019; Karras, Laine, and Aila 2019; Arjovsky, Chintala, and Bottou 2017; Karras et al. 2018). A typical GAN consists of two parts: a generator and a discriminator. The generator maps a randomly sampled latent code to a high-fidelity image while the discriminator tries to distinguish the real distribution from the fake/generated data. Conventionally, GANs are based on deep neural networks where the latent code is fed into the convolutional layers after an affine transformation (Arjovsky, Chintala, and Bottou 2017; Radford, Metz, and Chintala 2016). Style-based GANs like StyleGAN (Karras, Laine, and Aila 2019) and StyleGAN2 (Karras et al. 2020) transform the latent codes to layer-wise style codes and feed them to each convolutional layer using Adaptive Instance Normalization (AdaIN) (Huang and Belongie 2017). This operation ensures that each convolutional layer receives sufficient information of the latent code, and thus helps GAN to generate high-quality images.

Semantic Editing on GANs

Early GANs generate images from random latent codes, whose attributes cannot be directly edited. Recently, much effort has been made on learning semantics in the latent space of a fixed GAN model, which do not need to retrain the model for attribute editing. In this direction, unsupervised methods do not need labeled data and can achieve high-quality editing results, but it is difficult to choose the specified semantics (Shen and Zhou 2021; Härkönen et al. 2020; Wu, Lischinski, and Shechtman 2021; Zhu et al. 2022; Voynov and Babenko 2020). As a comparison, supervised methods are more accurate in decoupling and characterizing attributes desired by users; however, they usually need pre-trained attribute assessors or large amount of labeled data, which restricts their applications (Goetschalckx et al. 2019; Shen et al. 2020; Yang et al. 2021b; Zhuang, Koyejo, and Schwing 2021; Jiang et al. 2021).

Contrastive Representation Learning

Contrastive representation learning (CRL) is a state-of-the-art unsupervised leaning technique, which aims to maximize the mutual information and has shown its capability to outperform data-compression methods (Hinton and Salakhutdinov 2006). Representative CRL methods can introduce mutual information of representations between an image and itself (He et al. 2020; Wu et al. 2018; Shrivastava et al. 2011) or between an image and its transformed version (Chen et al. 2020; Misra and van der Maaten 2020; Park et al. 2020). A typical work is CUT (Park et al. 2020), which introduced the InfoNCE (Van den Oord, Li, and Vinyals 2018) to the image translation task and achieved high-quality translation.

Method

We first present our view on the role of large-scale data labeled with target attributes that are required in supervised methods. Based on this view, we present our solution to use limited labeled data in our FEditNet.

Role of Large-Scale Data in Supervised Methods

For those supervised methods relying on training an attribute assessor (Yang et al. 2021b; Yang, Shen, and Zhou 2021), if only a few data are used, the assessor trained from scratch will rapidly memorize the data rather than recognizing the target attribute. For example, if we want to train an assessor which can distinguish a target attribute with *one single sample* of data, then the trained assessor may output a *True* prediction *only if* the test image looks similar to the training

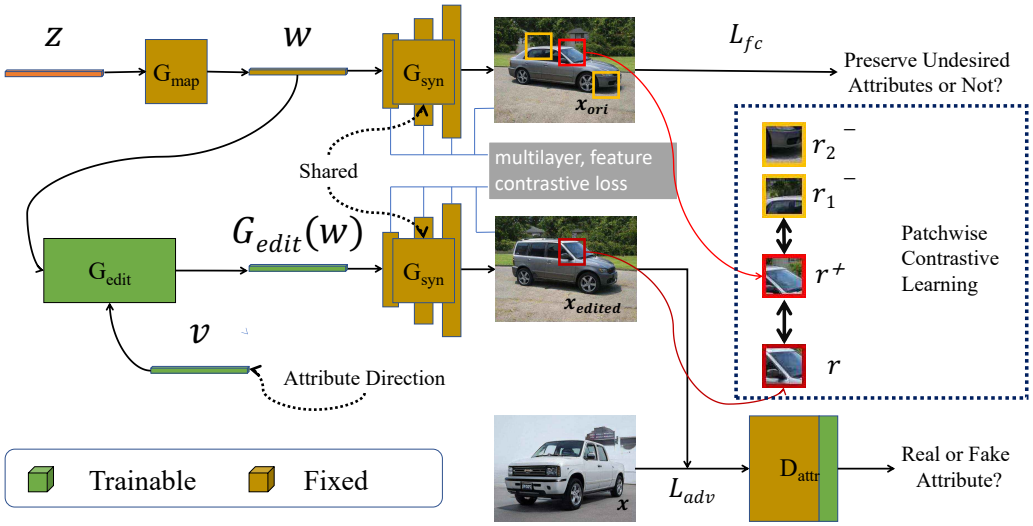


Figure 2: The overview of our FEditNet pipeline. Given a pretrained and fixed StyleGAN model, we aim to discover the latent semantic and the editing direction θ in the latent space, which manipulates the target attribute from x_{ori} to x_{edited} while keeping other attributes fixed. To this end, we fix the backbone of the pretrained discriminator and equip it with a light linear classifier as $D_{attr}(\cdot)$. The novel feature contrastive loss $\mathcal{L}_{fc}$ is also introduced for training.

data sample due to overfitting. As a result, the learning of latent semantics lacks correct supervision. This overfitting problem directly leads to poor diversity of image editing, in which all edited images have high similarity to the few training data.

In addition to the difficulty of learning the target attribute, attribute disentanglement is another challenge in the setting of only a few available training data. We note that the attribute assessor is essentially a classifier, which does not provide supervision on the attributes that need to be fixed during the editing process. In other words, regardless of whether the non-target attributes change or not, the attribute assessor will output a *True* prediction as long as the input image contains the target attribute. Therefore, traditional supervised methods need a large set of diverse data to achieve attribute disentanglement and to ensure the attributes other than the target one are fixed during image editing.

Few-shot Semantic Exploration in GANs

Given a pretrained and fixed generator, one representative way to explore latent semantics is to learn an editing direction v by the GAN framework on a labeled dataset together with an attribute assessor (Yang et al. 2021b; Yang, Shen, and Zhou 2021). As discussed above, training an assessor from scratch is vulnerable when using a few training data. To mitigate model overfitting, some works on GAN transfer have been proposed to reduce trainable parameters (Mo, Cho, and Shin 2020; Robb et al. 2020). We observe that during the training of GANs, the discriminator is trained using thousands of images and thus its backbone has a strong capability of extracting good image features. We follow the idea in (Yang et al. 2021a) to adequately reuse the knowledge of the pretrained discriminator together with the given generator, which can provide sufficient supervision and prevent the model from overfitting in our few-shot settings.

Note that the attribute assessor does not provide supervision for attribute disentanglement. Even with a well-trained attribute assessor, we cannot guarantee a good semantic disentanglement in the latent space. Such entanglement in latent spaces is unacceptable in the application of local image editing. To address this issue, we improve the patch-wise contrastive loss in (Park et al. 2020) and propose our specially designed feature contrastive.

Architecture of FEditNet

Based on the investigation in previous sections, here we propose to explore the latent semantics of pretrained style-based GANs (Karras, Laine, and Aila 2019; Karras et al. 2020) using a few labeled data. The generator of these pretrained GANs is composed of a mapping network $G_{map}(\cdot)$ and a synthesis network $G_{syn}(\cdot)$. $G_{map}(\cdot)$ constructs a function mapping the latent code $z \in \mathcal{Z}$ sampled from Gaussian distribution to the intermediate latent code $w \in \mathcal{W}$, which is also called style code. $G_{syn}(\cdot)$ maps the style code w to the high-resolution synthesized image x, i.e.,

$$w = G_{map}(z), \tag{1}$$
$$x = G_{syn}(w). \tag{2}$$

Direction Generator. Our editing direction generator $G_{edit}(\cdot)$ focuses on learning an attribute-corresponding direction v for all style codes w, and the manipulated image can be represented as:

$$G_{edit}(w) = w + l \cdot v, \tag{3}$$
$$x_{edited} = G_{syn}(G_{edit}(w)), \tag{4}$$

where l is a fixed length for manipulation. Here we use a fixed length during the training to ensure that our generator has a significant and consistent effect on image editing.

Attribute assessor. As aforementioned, training an attribute assessor from scratch using a few data is vulnerable. Given that we fix the pretrained StyleGAN generator and only apply a linear translation to the style space, the generated images before and after editing are in the same domain of the given generator, where the discriminator still works for feature extraction and classification. Therefore, we reuse the pretrained discriminator together with the given generator (Yang et al. 2021a). In detail, we use the backbone $d(\cdot)$ of the discriminator and freeze its parameters as a feature extractor. Then we equip $d(\cdot)$ with a light linear classifier $\phi(\cdot)$ as our attribute assessor $D_{attr}(\cdot)$. The assessor outputs the probability p for the given image x, which measures how likely it contains the target attribute, i.e.,

$$p = D_{attr}(x) = \phi \circ d(x). \tag{5}$$

Objectives for Learning Directions

The loss function $\mathcal{L}(G_{edit}, D_{attr})$ consists of two parts: (1) the adversarial loss $\mathcal{L}_{adv}$ which pushes the direction generator G_{edit} to compete against the attribute assessor $D_{attr}(\cdot)$, and (2) our specially-designed feature contrastive loss $\mathcal{L}_{fc}$ which acts as a regularizer to force the other attributes to be unchanged. We use a global loss weight on $\mathcal{L}_{fc}$ to adjust the strength of the regularizer, i.e.,

$$\mathcal{L}(G_{edit}, D_{attr}) = \mathcal{L}_{adv} + \lambda_{fc}\mathcal{L}_{fc}, \tag{6}$$

where λ_{fc} is the global loss weight. We also introduce a layer-wise loss weight adjustment and Bayesian-optimization-based adaptive loss weights, whose details are presented in the experiment section and

Adversarial Loss. We apply the adversarial loss to both direction generator $G_{edit}(\cdot)$ and the attribute assessor $D_{attr}(\cdot)$, which promotes the min-max competition. This drives $D_{attr}(\cdot)$ to make correct prediction and $G_{edit}(\cdot)$ to learn the latent semantics of the target attribute. The adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{x\in\mathcal{X}}[\log\left(D_{attr}(x)\right)] \\ &- \mathbb{E}_{z\in\mathcal{Z}}[\log\left(1 - D_{attr}\left(x_{edited}\right)\right)], \end{aligned} \tag{7}$$

$$\mathcal{L}_G = -\mathbb{E}_{z\in\mathcal{Z}}[\log\left(D_{attr}\left(x_{edited}\right)\right)], \tag{8}$$

$$\mathcal{L}_{adv} = \mathcal{L}_D + \mathcal{L}_G, \tag{9}$$

where x_{edited} is the image editing result from z by the pretrained generator and $G_{edit}(\cdot)$, i.e., using Eqs. (1) and (3).

Feature Contrastive Loss. Recall that exploring latent semantics only with a few data cannot provide sufficient supervision to train a direction generator that can fix the attributes other than the target attribute. To address this issue, we introduce the feature contrastive loss as a regularizer to help with attribute disentanglement and to freeze the other attributes.

Following the setting of CUT (Park et al. 2020), we maximize the mutual information between images before and after image editing. The idea of contrastive learning is to correlate two signals, i.e., a *query* and its *positive* example, in contrast to other *negative* examples. The query, its positive example and N negative examples are mapped to K dimensional vectors $u, u^+ \in \mathbb{R}^K$, $u^- \in \mathbb{R}^{N\times K}$ respectively. We can mathematically represent the probability of the positive example being selected over the negative examples as:

$$l(u, u^+, u^-) = -\log \frac{\exp\frac{u\cdot u^+}{\tau}}{\exp\frac{u\cdot u^+}{\tau} + \sum_{n=1}^{N}\exp\frac{u\cdot u_n^-}{\tau}} \tag{10}$$

Notice that style-based generators feed the style code to all convolutional layers which are organized in a hierarchical structure (Karras, Laine, and Aila 2019; Karras et al. 2020), we make use of the layer-wise feature stacks to design the contrastive loss as follows. We select L layers and pass the feature maps through a small MLP H_l as used in SimCLR (Chen et al. 2020), producing a stack of features $\{r_l\}_L = \{H_l(f_l(x_{ori}))\}_L$, where f_l represents the feature map of the l-th convolutional layer. We index these layers as $l \in \{1, 2, \cdots, L\}$ and $s \in \{1, 2, \cdots, S_l\}$, where S_l is the number of spatial locations in each layer. We refer to the corresponding feature as $r_l^s \in \mathbb{R}^{C_l}$ and other features as $r_l^{S\setminus s} \in \mathbb{R}^{(S_l-1)\times C_l}$, where C_l is the number of channels of each layer. Similarly, we encode the edited image x_{edited} as $\{\hat{r}_l\}_L = \{H_l(f_l(x_{edited}))\}_L$. Then the feature contrastive loss can be formulated as:

$$\mathcal{L}_{fc} = \frac{1}{L}\sum_{l=1}^{L}\sum_{s=1}^{S_l} l(\hat{r}_l^s, r_l^s, r_l^{S\setminus s}). \tag{11}$$

In other words, we refer to the patches at the same location as the positive one, while negative patches are all patches at different locations.

Given that (1) different layers in style-based generators correspond to different attributes and (2) the patch-wise contrastive loss promotes content consistency of the input representations, the feature contrastive loss is able to help fix the attributes other than the target attribute.

Experiments

We summarize datasets, implementation details, baselines, evaluation metrics, qualitative and quantitative experimental results in this section.

Datasets. We evaluated FEditNet on:

- facial datasets - CelebA dataset (Liu et al. 2015), FFHQ dataset (Karras, Laine, and Aila 2019) and Danbooru2018 dataset (Anonymous, community, and Branwen 2021),
- animal dataset - AFHQ dataset (Choi et al. 2020),
- scene datasets - LSUN dataset (Yu et al. 2015) including church, tower and car.

In each dataset, we manually select 30 synthesized images containing the target attribute as our training dataset.

Implementation details. We trained FEditNet on the platform of PyTorch (Paszke et al. 2019), in a Linux environment with an Nvidia A100 PCIe GPU. The whole 30,000 training steps were completed in 6 hours to obtain editing directions of the best quality. We apply FEditNet to style-based GANs (Karras, Laine, and Aila 2019; Karras et al.

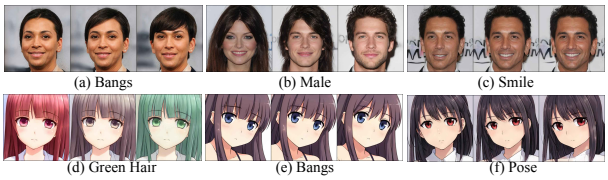


Figure 3: Image editing on binary attributes with StyleGAN2 (Karras et al. 2020). (a)-(c) are trained on CelebA dataset, and (d)-(f) are trained on Danbooru2018 dataset.

Figure 4: Image editing on multiple attributes with StyleGAN2 (Karras et al. 2020) trained on FFHQ dataset.

2020), then use the Adam (Nothaft et al. 2015) optimizer to simutaneously optimize the direction generator $G_{edit}(\cdot)$ and attribute assessor $D_{attr}(\cdot)$. The learning rate of $G_{edit}(\cdot)$ is set to $2 \cdot 10^{-3}$ while that of $D_{attr}(\cdot)$ is set to $2 \cdot 10^{-4}$. We use a fixed length of $l = 5$ in Eq. (3).

Baselines. As we are not aware of few-shot methods for GAN-based attribute editing, we compare FEditNet with three state-of-the-art latent semantic discovery methods: AdvStyle (Yang et al. 2021b), Latent2im (Zhuang, Koyejo, and Schwing 2021) and InterFaceGAN (Shen et al. 2020). Since AdvStyle only provided pretrained editing directions while Latent2im and InterFaceGAN released codes, we compare FEditNet with them in the following way. The editing directions of FEditNet are trained with only 30 samples, while the editing directions of AdvStyle are trained with the full dataset, Latent2im pretrained an attribute classifier on CelebA and InterFaceGAN is trained using 30 pairs of data (including both positive and negative samples).

Evaluation metrics. To quantitatively compare the editing results, we follow (Shen et al. 2020) to apply the re-scoring analysis on 2,000 images containing different facial attributes. We also compare the four methods in a user study, where users were asked to score edited results on three dimensions, i.e., *quality* (the quality of the results), *adequateness* (the significance of the editing on the target attribute) and *consistency* (whether non-target attributes are fixed).

Latent Semantics and Attributes Editing

FEditNet can provide interpretable latent semantics for the target attribute with only a few training data, and our proposed feature contrastive loss provides strong constraints to keep other attributes fixed during image editing. Fig. 3 shows some qualitative results on editing attributes in human face and anime, demonstrating that high-quality facial attribute manipulation can be achieved by FEditNet. Unlike existing methods that train the attribute assessor from scratch, FEditNet introduces the backbone $d(\cdot)$ of pretrained discriminator, which can better extract features from the images and obtain the latent semantics of fine attributes such as eyelids (see the editing results in Fig. 1e for some examples).

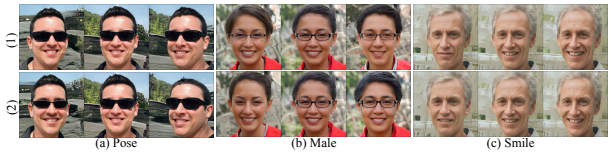


Figure 5: Qualitative comparison of image editing learned by (1) our FEditNet trained with 30 samples, (2) AdvStyle (Yang et al. 2021b) trained with the full dataset. The pretrained model uses StyleGAN (Karras, Laine, and Aila 2019) trained with the FFHQ dataset.

	Bangs	Male	Smile	Pose
Male	-0.11	0.41	0.29	0.01
Smile	-0.02	-0.04	0.23	0.01
Pose	-0.02	0.04	0.00	0.44

(a) FEditNet with 30 data

	Bangs	Male	Smile	Pose
Male	0.01	0.12	-0.07	-0.11
Smile	-0.02	-0.02	0.20	0.01
Pose	-0.02	0.07	-0.11	0.42

(b) AdvStyle with full dataset

Table 1: Re-scoring analysis of (a) our FEditNet, (b) AdvStyle (Yang et al. 2021b) from StyleGAN (Karras, Laine, and Aila 2019) trained on FFHQ dataset. Each row shows how semantic score varies of images before and after editing with a target attribute direction by different predictors.

FEditNet is able to acquire latent semantics from a few training samples, which can be easily obtained by labeling a small amount of data from unlabeled datasets. Some results are shown in Fig. 1. Given only 30 training samples, FEditNet can effectively manipulate images using the editing direction of the target attribute while fixing other attributes.

Comparisons with State of the Arts

In order to keep non-target attributes fixed, previous methods introduce an extra classifier which distinguishes image identities before and after editing. This solution is based on the fact that synthesized images before and after editing share most subjective semantic knowledge (e.g., a human face image is still a human face after manipulation). However, the image editing in these solutions may fail when two cases occur: (1) there is a gap between the domains of synthesized images before and after editing (e.g., editing from cat to dog), in which the identity classifier may no longer work; (2) the domain of the synthesized images has either low diversity (e.g. all tigers have a similar appearance) or few common structural semantics (e.g., editing in the LSUN church dataset). In other words, the training of the identity classifier may fail and make the learning of latent semantics incorrect. As a comparison, FEditNet does not make use of the identity classifier. Instead, our layer-wise feature contrastive loss provides strong supervision hierarchically to keep the attributes other than the target attribute fixed through the content consistency. Therefore, FEditNet can effectively handle the above two cases. We conduct comprehensive experiments to show the proposed FEditNet surpasses the state-of-the-art methods on editing quality, especially the capability of preserving non-target attributes.

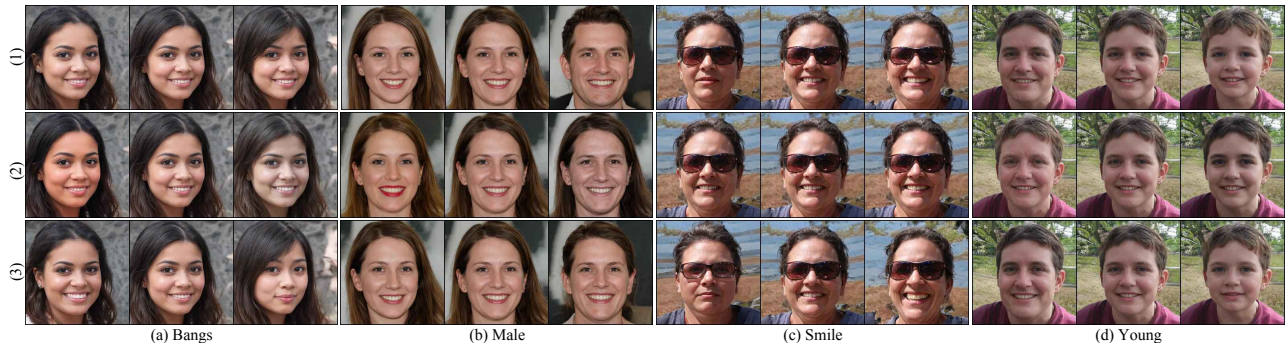


Figure 6: Qualitative comparison of latent semantics learned by (1) our FEditNet trained with 30 data samples, (2) Latent2im (Zhuang, Koyejo, and Schwing 2021) trained with pretrained attribute assessor and (3) InterFaceGAN (Shen et al. 2020) trained with 30 pairs of data samples. All the models use StyleGAN2 (Karras et al. 2020) trained on FFHQ dataset.

	Bangs	Male	Smile	Pose	Young
Bangs	0.38	-0.09	0.01	-0.01	-0.05
Male	-0.03	0.33	-0.05	0.06	-0.13
Smile	-0.04	0.00	0.30	0.02	0.00
Young	0.04	-0.09	-0.08	-0.03	0.11

(a) FEditNet with 30 data

	Bangs	Male	Smile	Pose	Young
Bangs	0.23	-0.10	0.20	0.03	-0.04
Male	-0.04	0.08	0.16	0.01	0.03
Smile	-0.05	-0.04	0.27	0.03	0.08
Young	-0.05	-0.07	0.14	0.01	0.15

(b) Latent2im with pretrained classifier

	Bangs	Male	Smile	Pose	Young
Bangs	0.32	-0.13	-0.29	-0.07	0.14
Male	-0.04	0.17	0.17	0.03	-0.19
Smile	-0.03	0.09	0.28	0.08	0.02
Young	0.00	-0.16	-0.51	-0.07	0.17

(c) InterFaceGAN with 30 pairs of data

Table 2: Re-scoring analysis of (a) our FEditNet, (b) Latent2im (Zhuang, Koyejo, and Schwing 2021) and (c) InterFaceGAN (Shen et al. 2020) from StyleGAN2 (Karras et al. 2020) trained on FFHQ dataset. Each row shows how semantic score of images varies before and after editing with a target attribute direction by different predictors.

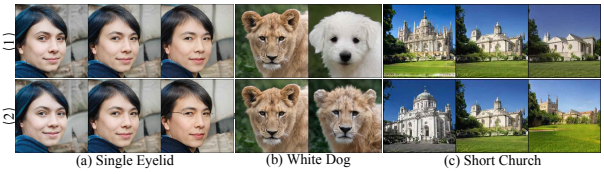


Figure 7: Qualitative comparison of editing latent semantics learned by (1) FEditNet, (2) InterFaceGAN (Shen et al. 2020). The three columns are trained on FFHQ dataset, AFHQ dataset, church from LSUN dataset respectively.

Comparison with AdvStyle Since (1) AdvStyle is based on StyleGAN while Latent2im and InterFaceGAN are based on StyleGAN2, and (2) AdvStyle only provides pre-selected latent semantics, we compare them separately. First we compare FEditNet with AdvStyle using four editing tasks and some qualitative results are shown in Fig. 5. We observe that AdvStyle fails to fix some attributes other than the target attribute and cannot achieve significant editing effects as FEditNet does. The quantitative results on re-scoring analysis summerized in Table 1 and user study in Table S2 (supplementary material) also deomonstrate this observation.

Comparison with InterFaceGAN and Latent2im Second, we compare FEditNet with InterFaceGAN and Latent2im, and qualitative results are illustrated in Fig. 6. We observe that (1) even with little training data, reusing the pretrained discriminator makes FEditNet work well on editing the target attribute, and our feature contrastive loss ensures other attributes to be fixed after editing, (2) although Latent2im makes use of an attribute assessor to explore the latent semantics, it is difficult to edit some subtle attributes such as smile and bangs, and (3) InterFaceGAN suffers severely from the attribute entanglement. We can also conclude from Table 2 and the user study in Table S2 (supplementary material) that the manipulation quality of our FEditNet surpasses the SOTA methods.

Extra Comparison We further compare editing quality of FEditNet and InterFaceGAN on non-labeled or non-facial datasets (note that neither AdvStyle nor Latent2im can edit images by training on these datasets). Some qualitative results are shown in Fig. 7. We observe that (1) FEditNet has a stronger attribute disentanglement capability, i.e., FEditNet successfully edit the target attribute of *single eyelid* and simultaneously keep other attributes fixed, while InterFaceGAN fails to fix non-target attributes, (2) InterFaceGAN is highly likely to fail on the editing tasks between domains with gaps such as converting lions into white dogs in Fig. 7b; as a comparison, FEditNet can easily handle domain gaps and even edit subtle spatial attributes such as shortening the height of a church in Fig. 7c.

Ablation Studies

In our proposed method, we introduce the reuse of the pretrained discriminator and a novel feature contrastive loss. Fig. 8 and Table S3 (supplementary material) show that the

Layer	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
Bangs	1.00	1.07	1.04	1.26	1.18	1.00	0.99	1.05	1.15	1.13	1.12	1.09	1.10	1.05	1.00	1.01
Smile	1.00	1.02	1.04	1.32	1.32	1.14	1.12	1.12	1.17	1.14	1.19	1.20	1.14	1.17	1.12	1.02
Male	1.00	1.12	1.27	1.45	1.49	1.30	1.35	1.38	1.92	1.73	1.97	1.88	2.14	1.59	1.45	1.07

Table 3: For each layer and attribute, we show the ratio r_{fc} between the feature contrastive loss of latent codes near separation boundary and far away from the boundary along the forward direction of the attribute.

Figure 8: Ablation study of the reuse of the pretrained discriminator on the CelebA dataset of (1) the baseline of FEditNet, (2) without the reuse of the pretrained discriminator.

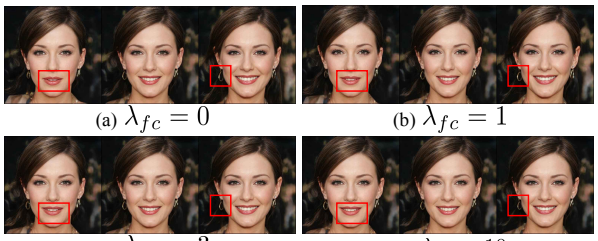


Figure 9: Ablation study of the loss weight λ_{fc} on the CelebA dataset.

reuse of the discriminator helps FEditNet capture the target attribute "Male", so that the editing results are more significant. We also conclude from Fig. 9 and Table S4 (Supplementary material) that (1) larger λ_{fc} better emphasizes the attribute consistency while smaller λ_{fc} promotes more significant editing effects, hence it is expected to set a smaller λ_{fc} for the editing tasks between domains with big gaps, (2) smaller λ_{fc} relaxes the constraints responsible for freezing the non-target attributes, especially the earring part during the manipulation on "Smile" attribute.

Layer-wise Analysis of Representations in GANs

Inspired by the works (Karras, Laine, and Aila 2019; Yang, Shen, and Zhou 2021), we analyze each convolutional layer using the feature contrastive loss. Noting that the feature contrastive loss builds strong constraints on content consistency of layer-wise feature map in the generator, it naturally reflects the response of each convolutional layer to the target attribute during editing.

It is widely recognized that editing is more conspicuous near the separation boundary than far from the boundary along the editing direction (Shen et al. 2020). Hence, the image content is more difficult to preserve near the separation boundary. For the feature map of each convolutional layer, denote by r_{fc} the ratio between the feature contrastive loss near the boundary and that far away from the boundary along the editing direction. The larger the r_{fc} is, the more

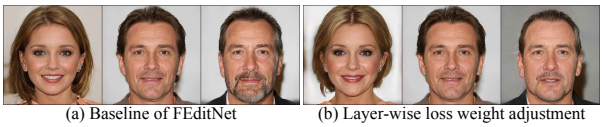


Figure 10: Qualitative comparison of FEditNet baseline and with layer-wise loss weight adjustment from StyleGAN2 (Karras et al. 2020) and CelebA dataset. We better disentangle "Age" and "Smile" attributes from the target one with this adjustment technique.

responsive this layer is to the target attribute and vice versa.

To sample latent codes near or far from the separation boundary, we use the pretrained StyleGAN2 on the CelebA dataset. First, we synthesize $400K$ images randomly. Then we assign scores using pretrained attribute predictors for all images, and choose $10K$ samples with the highest scores as the latent codes far from the separation boundary. Since we cannot locate the exact boundary, for the latent codes near the boundary, we randomly sample $10K$ samples from all images, aiming to reduce the influence of the latent codes far away from the boundary in the backward editing direction. The results are summarized in Table 3. Given that the ratio r_{fc} reflects the response of one layer to the target attribute, one can add additional constraint (or relax constraints) for those layers with small (or large) r_{fc} respectively.

To validate our analysis, we train latent semantics after adjusting the layer-wise loss weights and then compare with the baseline of FEditNet through attribute manipulation and re-scoring analysis. From Fig. 10 and Table S5 (supplementary material), we observe that the adjustment promotes the disentanglement between attributes and helps explore accurate latent semantics. Furthermore, we introduce an adaptive loss weight through Bayesian optimization (Mockus 2012) for fast adjusting loss weights and improving the performance, which are detailed in the supplementary material.

Conclusions

In this paper, we propose a GAN-based approach, called FEditNet, to explore latent semantics for attribute editing. We address the limitations of large-scale labeled dataset or pretrained attribute predictors by only using a few training data to achieve attribute disentanglement. Our method does not need to use prior knowledge of semantics which is commonly used in facial manipulation task; instead, we reuse the knowledge of the given pretrained GANs, which also helps understand the training process of GANs. Our specially-designed feature contrastive loss also provides a novel way to interpret the hierarchical structure in style-based models.

Acknowledgements

This work was partially supported by the Natural Science Foundation of China (61725204), Tsinghua University Initiative Scientific Research Program and Beijing Natural Science Foundation (L222008).

References

Anonymous; community, D.; and Branwen, G. 2021. Danbooru2020: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. https://www.gwern.net/Danbooru2020. Accessed: DATE.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 214–223. PMLR.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8185–8194.

Goetschalckx, L.; Andonian, A.; Oliva, A.; and Isola, P. 2019. GANalyze: Toward Visual Definitions of Cognitive Image Properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5743–5752.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27, 2672–2680. Curran Associates, Inc.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. GANSpace: Discovering Interpretable GAN Controls. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9841–9850. Curran Associates, Inc.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1510–1519.

Jiang, Y.; Huang, Z.; Pan, X.; Loy, C. C.; and Liu, Z. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13799–13808.

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 (CVPR)*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107–8116.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3730–3738.

Misra, I.; and van der Maaten, L. 2020. Self-Supervised Learning of Pretext-Invariant Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6706–6716.

Mo, S.; Cho, M.; and Shin, J. 2020. Freeze the Discriminator: a Simple Baseline for Fine-Tuning GANs. In *CVPR AI for Content Creation Workshop*.

Mockus, J. 2012. *Bayesian approach to global optimization: theory and applications*, volume 37. Springer Science & Business Media.

Nothaft, F. A.; Massie, M.; Danford, T.; Zhang, Z.; Laserson, U.; Yeksigian, C.; Kottalam, J.; Ahuja, A.; Hammerbacher, J.; Linderman, M.; Franklin, M.; Joseph, A. D.; and Patterson, D. A. 2015. Rethinking Data-Intensive Science Using Scalable Analytics Systems. In *Proceedings of the 2015 International Conference on Management of Data (SIGMOD '15)*, 631–646. ACM.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *European Conference on Computer Vision*, 319–345.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32, 8024–8035. Curran Associates, Inc.

Plumerault, A.; Borgne, H. L.; and Hudelot, C. 2020. 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*.

Robb, E.; Chu, W.-S.; Kumar, A.; and Huang, J.-B. 2020. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*.

Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9240–9249.

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 (CVPR)*, 1532–1540.

Shrivastava, A.; Malisiewicz, T.; Gupta, A.; and Efros, A. A. 2011. Data-driven visual similarity for cross-domain image matching. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, 1–10.

Van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, arXiv–1807.

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 (CVPR)*, 12863–12872.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3733–3742.

Yang, C.; Shen, Y.; Zhang, Z.; Xu, Y.; Zhu, J.; Wu, Z.; and Zhou, B. 2021a. One-Shot Generative Domain Adaptation. *arXiv preprint arXiv:2111.09876*.

Yang, C.; Shen, Y.; and Zhou, B. 2021. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, 129(5): 1451–1466.

Yang, H.; Chai, L.; Wen, Q.; Zhao, S.; Sun, Z.; and He, S. 2021b. Discovering Interpretable Latent Space Directions of GANs Beyond Binary Attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12177–12185.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhu, J.; Shen, Y.; Xu, Y.; Zhao, D.; and Chen, Q. 2022. Region-Based Semantic Factorization in GANs. In *International Conference on Machine Learning (ICML)*, 27612–27632.

Zhuang, P.; Koyejo, O. O.; and Schwing, A. 2021. Enjoy Your Editing: Controllable {GAN}s for Image Editing via Latent Space Navigation. In *International Conference on Learning Representations*.