FEditNet++: Few-Shot Editing of Latent Semantics in GAN Spaces With Correlated Attribute Disentanglement

Ran Yi, Teng Hu, Mengfei Xia, Yizhe Tang, and Yong-Jin Liu, *Senior Member, IEEE*

***Abstract*—Generative Adversarial Networks have achieved significant advancements in generating and editing high-resolution images. However, most methods suffer from either requiring extensive labeled datasets or strong prior knowledge. It is also challenging for them to disentangle correlated attributes with few-shot data. In this paper, we propose FEditNet++, a GAN-based approach to explore latent semantics. It aims to enable attribute editing with limited labeled data and disentangle the correlated attributes. We propose a layer-wise feature contrastive objective, which takes into consideration content consistency and facilitates the invariance of the unrelated attributes before and after editing. Furthermore, we harness the knowledge from the pretrained discriminative model to prevent overfitting. In particular, to solve the entanglement problem between the correlated attributes from data and semantic latent correlation, we extend our model to jointly optimize multiple attributes and propose a novel decoupling loss and cross-assessment loss to disentangle them from both latent and image space. We further propose a novel-attribute disentanglement strategy to enable editing of novel attributes with unknown entanglements. Finally, we extend our model to accurately edit the fine-grained attributes. Qualitative and quantitative assessments demonstrate that our method outperforms state-of-the-art approaches across various datasets, including CelebA-HQ, RaFD, Danbooru2018 and LSUN Church.**

***Index Terms*—Attribute disentanglement, StyleGAN latent space, few-shot attribute editing.**

Manuscript received 4 October 2023; revised 30 June 2024; accepted 18 July 2024. Date of publication 23 July 2024; date of current version 5 November 2024. This work was supported in part by the NationalNatural Science Foundation of China under Grant 62302297, Grant U2336214, Grant 62332019, Grant 72192821, and Grant 62272447, in part by Shanghai Sailing Program under Grant 22YF1420300, in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001, in part by Beijing Natural Science Foundation under Grant L222008 and Grant L222117, in part by the the Fundamental Research Funds for the Central Universities under Grant YG2023QNB17 and Grant YG2024QNA44, in part by ShanghaiMunicipal Science and Technology Major Project under Grant 2021SHZDZX0102, and in part by Shanghai Science and Technology Commission under Grant 21511101200. Recommended for acceptance by T. M. Hospedales. *(Corresponding author: Yong-Jin Liu.)*

Ran Yi, Teng Hu, and Yizhe Tang are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ranyi@sjtu.edu.cn; hu-teng@sjtu.edu.cn; tangyizhe@sjtu.edu.cn).

Mengfei Xia is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: xmf20@mails.tsinghua.edu.cn).

Yong-Jin Liu is with the MOE-Key Laboratory of Pervasive Computing, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: liuyongjin@tsinghua.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3432529, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3432529

I. Introduction

BASED on a min-max game between a generator and a discriminator, Generative Adversarial Networks (GANs) [1] learn a nonlinear mapping from random input distributions to real data domains, demonstrating high fidelity in generating high-resolution images [2], [3], [4].

Despite great success in achieving amazing quality of synthesized images, the mechanisms of how GANs construct semantics within their latent spaces during training remain unclear. To address this problem, some recent research endeavors have aimed to interpret semantics and disentangle properties within GANs' latent spaces [5], [6], [7]. One representative strategy to analyze the semantics is to add a learned direction on the latent code [8]. This strategy leverages geometrical properties of hyperplanes in latent spaces, enabling the possibility to explore interpretable latent spaces.

This strategy has led to the development of two types of methods for interpreting latent spaces of GANs: unsupervised and supervised. To formulate the latent semantics, a representative unsupervised way is to use principal component analysis (PCA) [6], [9], [10]. These methods apply PCA to either the model parameters or feature maps, selecting eigenvectors associated with the largest eigenvalues to identify directions with the most significant impact on synthesized images and semantics. However, they may lack precision in editing user-desired attributes. In comparison, supervised methods typically provide a better control and disentanglement but often require either a pretrained attribute classifier [11], [12] or large amounts of labeled data containing the target attributes [5], [13], [14], which restricts applications of supervised methods, such as in scenarios where labeling large amounts of data is infeasible and unlabeled datasets have to be used.

In order to address the challenge of accurately manipulating semantics with limited labeled data, in our preliminary conference work [15], we proposed Few-Shot Attribute Editing Network (FEditNet), which can efficiently edit target attributes using a small amount of labeled data, while also effectively decoupling multiple correlated attributes to prevent mutual interference among them. FEditNet employs a pre-trained StyleGAN trained on a large amount of unlabeled data as the generative network, and utilizes very few labeled data to explore the relationship between the StyleGAN latent space and attribute editing. By reusing the knowledge from a pre-trained

discriminator, FEditNet can better extract the feature of the editing targets within a limited dataset, thereby enhancing the editing of the target attributes more effectively. With the proposed feature contrastive loss, FEditNet adaptively learns to preserve features uncorrelated to the editing attribute, ensuring image content consistency before and after editing.

It is worth noting that in our research, we observed the entanglement phenomena during few-shot attribute editing, and this problem becomes more pronounced as the editing intensity increases. We summarize these entanglement challenges into two primary factors: 1) Data correlation, where certain attributes in the data tend to co-occur or disappear simultaneously, causing the model to treat the co-occurred attributes as a single attribute; and 2) Semantic latent correlation, where the similar semantics between attributes make the model difficult to edit one attribute without concurrently affecting the related attributes. These entanglement problems are prevalent in attribute editing, and although some existing methods [5], [16] aim to disentangle the coupled attributes, they still face challenges when dealing with very limited data, where attribute entanglement is much more severe.

Therefore, building upon FEditNet [15], we propose FEditNet++, an improved attribute editing model capable of effectively disentangling correlated attributes based on a very small amount of data. Different from FEditNet, which optimizes the attribute direction for individual attributes separately, FEditNet++ additionally incorporates relevant information of coupled attributes and jointly optimizes the editing directions for them at the same time. By proposing a novel decoupling loss, FEditNet++ suppresses the correlation among the editing directions of correlated attributes, thereby enabling the model to decouple the entanglement of attribute editing directions in the latent space. Furthermore, FEditNet++ also uses a new cross-assessment loss, which keeps the invariance of the correlated attributes when editing the target attribute, facilitating the further disentanglement of coupled attributes in the image space. By implementing dual decoupling in both latent space and image space, FEditNet++ can effectively achieve accurate editing of individual attributes while preserving the integrity of coupled attributes, enhancing the efficacy of few-shot attribute editing. Furthermore, to deal with the editing of novel attributes with unknown entanglements, we propose a novel-attribute disentanglement strategy that disentangles the novel attribute with one randomly sampled existing attribute in each iteration.

Finally, we extend our FEditNet++ to deal with fine-grained attribute editing tasks, which requires the control of attributes with subtle changes and the accurate editing of attributes with specified intensities. We experiment on one representative fine-grained editing task, i.e., editing of facial action units (AUs), which decomposes human expressions into multiple facial action units. We regard each AU as a target attribute and extend FEditNet++ for accurate and disentangled editing of AU. To edit the AUs accurately, we propose a coefficient regression network to predict the coefficients for each attribute direction and employ an AU sampling method to further disentangle the data correlation.

To validate the effectiveness of our FEditNet++, we evaluate our model on two face datasets (CelebA-HQ [17] and RaFD [18] datasets), an anime dateset (Danbooru2018 [19]), and a scene dataset (LSUN Church dataset [20]). We compare the few-shot semantic editing with state-of-the-art methods InterFaceGAN [5] and Latent2im [12]. In addition, we also evaluate the decoupling capabilities of FEditNet++ with respect to multiple attributes and demonstrate that, in comparison to FEditNet, FEditNet++ exhibits significantly enhanced attribute decoupling and editing capabilities. We further conduct experiments on fine-grained expression editing, demonstrating the effectiveness of FEditNet++ in disentangling over ten attributes and accurately editing fine-grained attributes.

This paper builds upon our previous conference paper FEditNet [15], which is published in *AAAI Conference on Artificial Intelligence, 2023*. In this journal paper, we propose FEditNet++, an attribute-disentanglement-based few-shot semantic editing model, and have made significant improvements and extensions. Firstly, we extend FEditNet to disentangle correlated attributes by jointly training editing directions of multiple correlated attributes. Secondly, we propose a decoupling loss and cross-assessment loss that disentangle the correlated attributes in both latent and image space. Thirdly, we propose a novel-attribute disentanglement strategy that disentangles the novel attribute with one randomly sampled existing attribute each time. At last, we extend our FEditNet++ to edit fine-grained attributes with high editing accuracy, which further validates the significance of our FEditNet++ in attribute editing and disentanglement.

To sum up, the main contributions of our few-shot attribute editing model are as follows:

- We propose FEditNet++, a disentanglement-based few-shot attribute editing model, which jointly trains and disentangles the correlated attributes and achieves a good few-shot editing performance.
- To disentangle the edited attributes from each other, we propose a novel decoupling loss and a novel cross-assessment loss, which reduce the correlation between coupled attributes in both latent and image space, enhancing the editing accuracy and avoiding the influence of the correlated attributes.
- We further propose a novel-attribute disentanglement strategy that disentangles the novel attribute with one randomly sampled existing attribute in each iteration to deal with the editing of novel attributes with unknown entanglements.
- We further extend FEditNet++ to deal with fine-grained-attribute editing tasks. In particular, we apply FEditNet++ on fine-grained expression editing with action units, and propose coefficient regression network and AU sampler for more accurate and disentangled editing, achieving a high fine-grained editing accuracy.

II. Related Work

A. Generative Adversarial Networks

GANs [1] have played a significant role in promoting the development of image synthesis [2], [4], [17], [21]. 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 [8], [21]. Style-based GANs like StyleGAN [2] and StyleGAN2 [3] transform the latent codes to layer-wise style codes and feed them to each convolutional layer using Adaptive Instance Normalization (AdaIN) [22]. This operation ensures that each convolutional layer receives sufficient information about the latent code and thus helps GAN generate high-quality images. In this paper, we focus on exploring the latent semantics of the layer-wise style codes.

B. Semantic Editing on GANs

The GAN-based semantic editing methods can be classified into three groups [23]: **Image domain translation-based methods**, **Semantic segmentation-based methods**, and **Latent space navigation-based methods**. For a detailed survey, the reader is referred to [23]. **1) Image domain translation-based methods** [24], [25], [26] edit images by translating the image from one domain to another domain, but they suffer from changing the identity during domain translation; **2) Semantic segmentation-based methods** [27], [28], [29], [30] improve the controllability of different semantics by decomposing different semantics into different embeddings; and **3) Latent space navigation-based methods** [16], [31], [32], [33], [34], [35], [36] learn semantics in the latent space of a fixed GAN model, and do not need to retrain the model for attribute editing. In the area of latent space navigation, unsupervised methods can achieve high-quality editing results without labeled data, but it is difficult to specify a target semantic [6], [9], [10], [34], [35], [37], [38], [39], [40]. As a comparison, supervised methods are more accurate in decoupling and characterizing attributes desired by users; however, they usually need pretrained attribute assessors or a large amount of labeled data, which restricts their applications [5], [11], [12], [14], [16], [31], [32], [33], [34], [41]. To decouple the correlated attributes, most of the supervised methods [11], [12], [16] rely on the disentangling ability of the pretrained attribute assessors. In contrast, InterFaceGAN [5] proposes a projection-based method to solve the entanglement problem in latent space, but may influence the editing of the target attribute and cause the entanglement with other attributes. Meanwhile, EditGAN [35] enables learning semantics from paired image and semantic mask data in DatasetGAN [42], but requires manually labeled fine-grained semantic masks as inputs and is incapable of editing global attributes like "Age" and "Gender".

Few-shot Domain Adaptation (Global Editing): Some methods [43], [44], [45], [46] aim to use few-shot training data to adapt generative models from the source domain to the target domain with the desired attributes, which can be considered as a type of global editing. Researchers have discovered that altering only a portion of the network weights [47], [48] and employing various regularization techniques [49], [50] in conjunction with batch statistics [51] can mitigate the overfitting problem under few-shot data. Recently, CDC [44], RSSA [45], and PCF [52] have introduced novel loss functions to maintain the structure of the generated distribution. Furthermore, DCL [46] proposes Dual Contrastive Learning, which can help avoid overfitting. Additionally, StyleGAN-NADA [43] harnesses the robust prior of the CLIP model [53] and introduces a new directional CLIP loss to adapt the model to the target domain. Although these few-shot model adaptation approaches can achieve global attribute editing with limited data, they have limited ability in local editing and all suffer from altering unrelated attributes, which are more suitable for scenarios where identity consistency is not a significant concern, such as style transfer.

TABLE I
COMPARISON WITH THE EXISTING ATTRIBUTES EDITING METHODS

Method	Attribute Disentanglement	Few-shot Editing	Continuous Editing	Global Editing	Local Editing	NO Additional Condition
EditGAN	✓	✓	✗	✗	✓	Semantic Mask
Prompt-to-Prompt	✗	✓	✗	✓	✗	Paired Prompt
StyleGAN-NADA	✗	✓	✗	✓	✓	✓
StyleCLIP	✗	✓	✓	✓	✓	✓
InterFaceGAN	○	✗	✓	✓	✓	✓
Latent2im	✗	✗	✓	✓	✓	✓
FEditNet	✗	✓	✓	✓	✓	✓
FEditNet++	✓	✓	✓	✓	✓	✓

TABLE II
THE DATA CORRELATION IN CELEBA-HQ DATASET

$Attr1$	$Attr2$	$\frac{Num(Attr1\&2)}{Num(Attr1)}$	$\frac{Num(Attr1\&2)}{Num(Attr2)}$	$Average$
Heavy_Makeup	Wearing_Lipstick	98.4 %	80.0%	**89.2%**
No_Beard	Wearing_Lipstick	69.3 %	**99.9%**	84.6%
High_Cheekbones	Smiling	85.3 %	83.8%	84.5%
No_Beard	Young	81.4 %	84.8%	83.1%
Attractive	Young	93.3 %	68.7%	81.0%
Heavy_Makeup	No_Beard	**100.0%**	56.3%	78.1%
Mouth_Slightly_Open	Smiling	76.5 %	76.8%	76.6%
Wearing_Lipstick	Young	89.0 %	64.2%	76.6%
Attractive	Wearing_Lipstick	75.0 %	76.6%	75.8%
Attractive	No_Beard	88.8 %	62.9%	75.8%

Some attributes usually appear at the same time, causing the trained attribute directions to be entangled with each other. The bold values show the maximum frequency of co-occurrence for attribute pairs.

Comparison with the Existing Editing Methods: A good attribute editing method should be equipped with the abilities for attribute disentanglement, few-shot editing, continuous editing, global and local editing, and do not require additional conditions. We summarize the representative attribute editing methods and compare them with our FEditNet++ in Table I on these editing abilities. Among all the comparison methods, only EditGAN [35] can accomplish attribute disentanglement due to its semantic-mask-based editing framework, but it requires additional semantic masks to edit target attributes. InterFaceGAN [5] has proposed a coarse projection-based disentanglement method, but the disentanglement ability is limited, and still suffers from changing unrelated attributes. Moreover, InterFaceGAN and Latent2im [12] rely on large-scale data for attribute editing, limiting their performance in few-shot attribute editing. Meanwhile, EditGAN, Prompt-to-prompt [54], and StyleGAN-NADA [43] do not support continuous editing, which cannot accurately control the intensity of the target attributes. In contrast, our FEditNet++ is the only method that supports attribute disentanglement, few-shot editing, continuous editing, global and local editing, and does not require additional conditions.

C. Contrastive Representation Learning

Contrastive representation learning (CRL) is a state-of-the-art unsupervised learning technique, which aims to maximize the mutual information and has shown its capability to outperform data-compression methods [55]. Representative CRL methods can introduce mutual information of representations between an image and itself [56], [57], [58] or between an image and its transformed version [59], [60], [61]. A typical work is CUT [61], which introduced InfoNCE [62] to the image translation task and achieved high-quality translation.

III. Few-Shot Challenge

We first analyze the challenges in few-shot attribute editing based on latent semantics of GANs. We summarize the challenges under a few-shot setting in three aspects: model overfitting, uncorrelated attribute preservation, and correlated attribute disentanglement.

A. Model Overfitting Under Few-Shot Setting

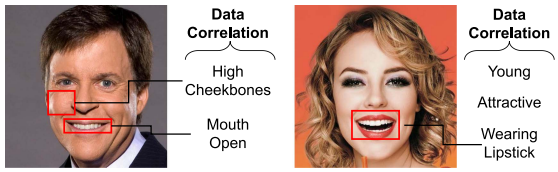
As illustrated in Section II-B, supervised methods are more accurate in attribute editing, but they heavily rely on a large-scale labeled dataset. For those supervised methods relying on training an attribute assessor [13], [14], if only a few data are used, the assessor trained from scratch will rapidly memorize the data rather than recognize 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 data sample due to overfitting. As a result, the learning of latent semantics lacks correct supervision. This overfitting problem directly leads to poor image editing diversity, in which all edited images have high similarity to the few training data.

B. Challenges in Uncorrelated Attribute Preservation Under Few-Shot Setting

In addition to the difficulty of learning the target attribute, uncorrelated attribute preservation is another challenge in settings with limited training data. We note that the attribute assessor is essentially a classifier that does not provide supervision on the contents that need to be fixed during the editing process. In other words, regardless of whether the uncorrelated 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 ensure the uncorrelated attributes are kept unchanged during image editing.

C. Challenges in Correlated Attribute Disentanglement Under Few-Shot Setting

In attribute editing, apart from the necessity to preserve uncorrelated attributes, there exist some attributes that are closely correlated and hard to disentangle. There are two sources for the attribute entanglement: 1) Data correlation: some attributes like "wearing_lipstick" and "young" usually appear at the same time in the dataset, which is caused by data correlation. 2) Semantic latent correlation: some attributes like "bangs" and "hairs", both pertain to facial hair, which may lead to entanglement in latent space.



(a) Attribute: Smile (b) Attribute: No Beard

Fig. 1. Data correlations manifest in various ways. For instance, certain attributes tend to co-occur, such as "Smile" often being accompanied by "High Cheekbones" and "Mouth Open". Similarly, the "No Beard" is frequently associated with attributes like "Young", "Attractive", and "Wearing Lipstick". Such correlations can lead to attribute entanglement, constraining the efficacy of attribute editing methods.

1) Data Correlation: Firstly, data correlation stems from the co-occurrence of coupled attributes in the dataset, leading to the discriminator mistakenly treating two attributes as the same. In the real world, the data correlation problem consists of *correlation in real distribution* and *correlation from data bias*. 1) For the correlation in real distribution, a representative example is the pair of "Smile" and "High_Cheekbones" (shown in Fig. 1(a)), where they are the natural attributes in real human faces and should not be disentangled. 2) As for the data correlation from data bias, e.g., "No Beard" and "Young" (shown in Fig. 1(b)), which may not necessarily appear at the same time, it is our goal to disentangle these attributes from each other.

We conduct a statistical analysis of co-occurring attributes within the CelebA-HQ dataset [17], and report the most serious co-occurred attribute pairs in Table II. For each pair of attributes $Attr_1$ and $Attr_2$, we compute the frequency of their co-occurrence ($Freq_1$ & $Freq_2$) by dividing the number of images with both attributes appearing together by the number of images with one of the attributes:

$$Freq_1 = \frac{Num(Attr_1 \,\&\, Attr_2)}{Num(Attr_1)},$$
$$Freq_2 = \frac{Num(Attr_1 \,\&\, Attr_2)}{Num(Attr_2)}, \tag{1}$$

where $Num(Attr_1 \,\&\, Attr_2)$ represents the number of simultaneous occurrences of both attributes in the CelebA-HQ dataset, and $Num(Attr_1)$ and $Num(Attr_2)$ represent the number of occurrences of the two attributes, respectively.

We exhibit the attribute-occurrence problem in (1), where the third and fourth columns show the co-occurrence frequency $Freq_1$ and $Freq_2$, and the fifth column is the average frequency of the third and fourth columns. In (1), we show ten pairs of attributes of the highest averaged co-occurrence frequency, representing the most serious data correlation problem in the dataset. Consequently, when training these attributes, the discriminator would mistakenly treat the co-occurred attributes as part of the editing attribute, resulting in changes in co-occurred attribute when editing the target attribute.

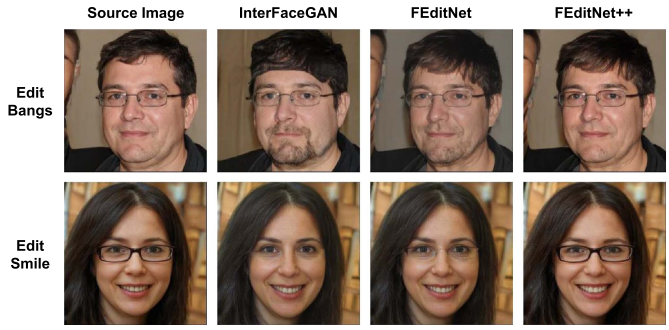


Fig. 2. The attribute entanglement problem encountered in InterFaceGAN [5]: 1) When editing "Bangs", the edited results coincide with the appearance of goatee; 2) When Editing "Smile", the eyeglasses disappear. With feature contrastive loss, FEditNet better preserves content but still suffers from entanglement when the editing intensity is large. With the novel decouple and cross-assessment losses, our FEditNet++ can disentangle the correlated attributes well when editing "Bangs" and "Smile".

2) Semantic Latent Correlation: The attribute entanglement phenomenon may also stem from the semantic latent correlation within the StyleGAN latent space. For instance, when editing the bangs attribute in human faces, facial hair significantly increases, often resulting in the simultaneous appearance of a beard. This attribute entanglement problem also leads to imprecise attribute editing. We show more examples of this problem in the editing results from InterFaceGAN in Fig. 2 (second column). When editing the "Bangs" attribute, the edited results usually coincide with the appearance of "Goatee". Moreover, when editing the "Smile" attribute, the eyeglasses in the edited results usually disappear.

IV. Method

In this paper, we propose a GAN-based method called FEditNet++, aiming to discover latent semantics and enable attribute editing using very few labeled data, without any pretrained predictors. In our previous conference paper FEditNet [15] (Fig. 3), we train a latent direction for each attribute under very few data, along which the corresponding attribute strengthens in the edited image. To solve the overfitting problem with limited data, we introduce an attribute assessor composed of a lightweight MLP, leveraging feature extracted from a pretrained StyleGAN discriminator. Furthermore, we introduce a feature contrastive loss, which effectively preserves the uncorrelated attributes during attribute editing.

In this journal paper, we propose FEditNet++ (Fig. 4), which further addresses the challenge of attribute entanglement problem arising from data correlation and semantic latent correlation (Section III-C). We extend our exploration to incorporate a correlated attribute disentanglement approach, which simultaneously optimizes latent directions and attribute assessors for multiple correlated attributes while keeping them independent of each other by a novel decoupling loss and cross-assessment constraint. With the proposed disentanglement approaches, FEditNet++ effectively enhances FEditNet's ability to disentangle correlated attributes, resulting in notable improvements in attribute editing performance.

A. Architecture for Few-Shot Attribute Editing

1) Attribute Direction: Based on the investigation in previous sections, here we propose to explore the latent semantics of pretrained style-based GANs [2], [3] using a few labeled data. The generator of these pretrained GANs is composed of a mapping network $G_{map}(\cdot)$ and a synthesis network $G_{syn}(\cdot)$. The mapping network $G_{map}(\cdot)$ 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 the style code. $G_{syn}(\cdot)$ maps the style code w to the high-resolution synthesized image x, i.e.,

$$\begin{aligned} w &= G_{map}(z), \\ x &= G_{syn}(w). \end{aligned} \tag{2}$$

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 [13], [14]. We propose an editing direction generator $G_{edit}(\cdot)$ that focuses on learning an attribute-corresponding editing direction v for all style codes w, and the manipulated image can be represented as:

$$\begin{aligned} G_{edit}(w) &= w + l \cdot v, \\ \hat{x} &= G_{syn}(G_{edit}(w)), \end{aligned} \tag{3}$$

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.

2) Attribute Assessor: In GAN-based attribute editing, the role of the attribute assessor is to distinguish images with the target attribute from the images without the attribute, which is trained adversarially with the attribute editing direction. As discussed in Section III-A, training an attribute assessor from scratch is vulnerable when using limited training data. To mitigate model overfitting, some works on GAN transfer have been proposed to reduce trainable parameters [47], [63]. 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 features. We follow the idea in [64] 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 setting.

When 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 [64]. In detail, we use the backbone $E(\cdot)$ of the discriminator and freeze its parameters as a feature extractor. Then, we equip $E(\cdot)$ with a lightweight two-layer 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:

$$p = D_{attr}(x) = \phi \circ E(x). \tag{4}$$

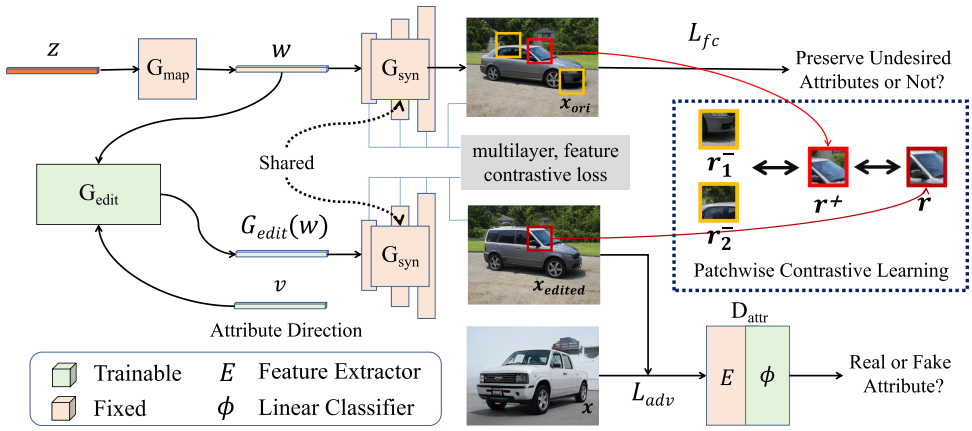


Fig. 3. The overview of FEditNet proposed in our preliminary conference paper [15]. Given a pretrained and fixed StyleGAN model, we discover the latent semantic and the editing direction θ in the latent space, which manipulates the target attribute from x_{ori} to x_{edited} while keeping other attributes unchanged. We fix the backbone of the pretrained discriminator and equip it with a light linear classifier $D_{attr}(\cdot)$. The novel feature contrastive loss $\mathcal{L}_{fc}$ is also introduced for training.

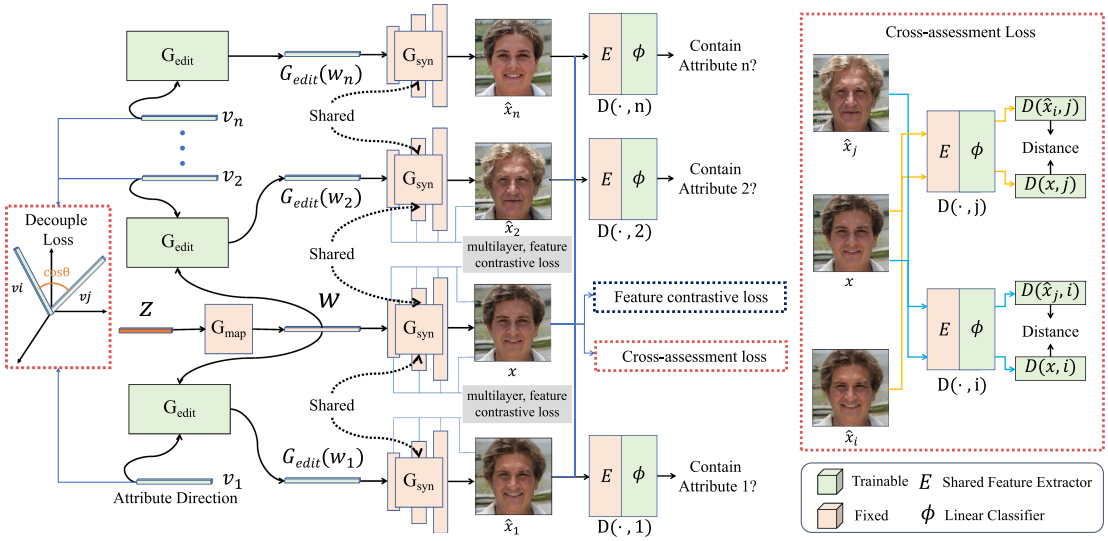


Fig. 4. The overview of our correlated-attribute disentanglement-based FEditNet++ pipeline and the disentanglement-based loss terms, where $\hat{x}_i$ and $\hat{x_j}$ are the edited results for attribute i and j. Our model achieves more precise attribute editing effects by jointly optimizing multiple attributes and decoupling them from each other.

B. 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.,

$$\mathcal{L}(G_{edit}, D_{attr}) = \mathcal{L}_{adv} + \lambda_{fc}\mathcal{L}_{fc}, \tag{5}$$

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 supplementary material.

1) 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 predictions and $G_{edit}(\cdot)$ to learn the latent semantics of the target attribute. The adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{x\in\mathcal{X}}[\log\left(D_{attr}(x)\right)] \\ &-\mathbb{E}_{z\in\mathcal{Z}}[\log\left(1 - D_{attr}\left(\hat{x}\right)\right)],\end{aligned} \tag{6}$$

$$\mathcal{L}_G = -\mathbb{E}_{z\in\mathcal{Z}}[\log\left(D_{attr}\left(\hat{x}\right)\right)], \tag{7}$$

$$\mathcal{L}_{adv} = \mathcal{L}_D + \mathcal{L}_G, \tag{8}$$

where $\hat{x}$ is the image editing result from z by the pretrained generator and $G_{edit}(\cdot)$, i.e., using (2) and (3).

2) Feature Contrastive Loss: Recall that exploring latent semantics only with a few data points cannot provide sufficient supervision to train a direction generator that can keep the unrelated attributes unchanged while editing the target semantics. 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 [61], 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:

$$l(u,u^+,u^-) = -\log \frac{\exp \frac{u\cdot u^+}{\tau}}{\exp \frac{u\cdot u^+}{\tau} + \sum_{n=1}^{N} \exp \frac{u\cdot u_n^-}{\tau}}. \quad (9)$$

Notice that style-based generators feed the style code to all convolutional layers, which are organized in a hierarchical structure [2], [3], 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 [59], 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 $l \in \{1,2,\ldots,L\}$ and $s \in \{1,2,\ldots,S_l\}$, 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 $\hat{x}$ as $\{\hat{r}_l\}_L = \{H_l(f_l(\hat{x}))\}_L$. Then, the feature contrastive loss can be formulated as:

$$\mathcal{L}_{fc} = \frac{1}{L}\sum_{l=1}^{L}\sum_{s=1}^{S_l} l(\hat{r}_l^s, r_l^s, r_l^{S\setminus s}). \quad (10)$$

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 keep the unrelated semantics unchanged during editing the target attribute

3) Identity Loss: The proposed feature contrastive loss can help keep the image patches that are unrelated to the target attributes unchanged when editing the target attribute. For facial attribute editing, some minor modifications to the facial part may change the identity significantly. Therefore, we further incorporate an identity preservation loss $\mathcal{L}_{id}$ by leveraging a pretrained facial feature extraction network on facial datasets [65]. This involves extracting identity features from both the source images x and edited images $\hat{x}$, followed by computing their cosine similarity. Maximizing this cosine similarity serves as an identity preservation loss, thereby promoting consistency in facial identity before and after editing:

$$\mathcal{L}_{id} = \langle R(x), R(\hat{x})\rangle, \quad (11)$$

where $R(\cdot)$ is the identity feature extractor.

The aforementioned few-shot learning of attribute direction and assessor with adversarial and feature contrastive loss constitutes our FEditNet, which achieves single attribute editing by exploring the latent semantics of pretrained StyleGAN using few-shot data. In the next section, we introduce our FEditNet++, which further addresses the multi-attribute entanglement challenge based on FEditNet.

C. Disentanglement Among Correlated Attributes

1) Attribute Entanglement Problem: In attribute editing, there often arises an attribute entanglement problem between some pairs of correlated attributes as illustrated in Section III-C. Although our FEditNet has better content preservation than InterFaceGAN with the help of our feature contrastive loss, it still faces the attribute entanglement problem to some extent due to the limited dataset, especially when the editing intensity grows, as shown in Fig. 2 (third column). This attribute entanglement problem poses a challenge to accurate editing and it is crucial to disentangle the editing directions of correlated attributes from each other.

2) Limitations of Existing Attribute Decoupling Approach: InterFaceGAN [5] has introduced a decoupling method via subspace projection. Given two attribute editing directions v_1 and v_2 with unit length, InterFaceGAN projects v_1 onto v_2 and then subtracts the projection from v_1 to get the disentangled direction $v_1' = v_1 - (v_1^T v_2)v_2$, which is perpendicular and hence independent to v_2. To edit two attributes v_1 and v_2 separately, there are two ways: *1)* calculate the new attribute direction (v_1', v_2') for one attribute (w.l.o.g, we choose v_1 here) and remain the other direction unchanged: $v_1' = v_1 - (v_1^T v_2)v_2$, $v_2' = v_2$; *2)* calculate the new attribute directions (v_1', v_2') for both the two attribute directions: $v_1' = v_1 - (v_1^T v_2)v_2$, $v_2' = v_2 - (v_2^T v_1)v_1$. However, both ways cannot fully address the attribute entanglement problem:

For the first decoupling method, the projection method can avoid the influence of direction v_2 when editing v_1' in the latent space. However, since v_2 is unchanged, there still remains entanglement in the semantic space when editing v_2. For the second decoupling method, both v_1' and v_2' are decoupled with the original v_1 and v_2 separately, i.e., $v'^T_1 v_2 = 0$ and $v'^T_2 v_1 = 0$. However, the new directions v_1' and v_2' are not independent with each other:

$$\begin{aligned} v'^T_1 v_2' &= (v_1 - (v_1^T v_2)v_2)^T (v_2 - (v_2^T v_1)v_1) \\ &= (v_1^T v_2)(-\|v_2\|_2 - \|v_1\|_2) + v_1^T v_2 + (v_1^T v_2)^3 \\ &= (v_1^T v_2)^3 - v_1^T v_2 \\ &\neq 0, \quad \text{if } v_1^T v_2 \neq 0, \pm 1. \end{aligned} \quad (12)$$

Therefore, the existing projection-based decoupling methods can not fully solve the attribute entanglement problem.

3) Disentanglement by Multi-Attribute Joint Training: Multi-attribute Joint Training Framework: Concerning the attribute entanglement problem arising from data correlation or semantic latent correlation, it is difficult to disentangle the target attribute from the correlated attributes without introducing additional information about the correlated attributes. To solve the attribute entanglement problem, we extend our original FEditNet by introducing the correlated attributes into our model, jointly training multiple attribute directions $\{v_1, v_2 \cdots v_K\}$ for these correlated attributes, and designing decoupling loss and cross-assessment to disentangle the correlated attributes from each other during training. The training framework is shown in Fig. 4.

Specifically, considering K correlated attributes $\{Attr_1, Attr_2, \cdots Attr_K\}$. For each attribute $Attr_i$, we assign it with an attribute editing direction v_i, along which the corresponding attribute strengthens. To disentangle each attribute from other correlated attributes, we jointly train K attribute directions and introduce decoupling loss (detailed below). Moreover, we train a shared attribute assessor $D(\cdot, i)$ to assess each target attribute i, to guide the training of the corresponding attribute directions and introduce cross-assessment to further disentangle.

4) Disentanglement Based on Decoupling Loss and Cross-Assessment: Decoupling Loss: To disentangle the correlated attributes, we further propose a decoupling loss. We disentangle two or more attributes from each other by concurrently editing them, and ensuring that the edited results for any attribute do not exhibit entangling problems with the correlated attributes. The proposed decoupling loss allows us to achieve pairwise decoupling between multiple attributes within the latent space, further enhancing the correlated attribute disentanglement capability of our FEditNet++.

Specifically, our decoupling loss constrains the editing directions corresponding to various attributes to be as orthogonal as possible, which is calculated by:

$$\mathcal{L}_{decouple} = \sum_{i,j} \langle v_i, v_j \rangle, \tag{13}$$

where v_i and v_j are the attribute-corresponding directions, and the cosine distance between each pair of attribute-corresponding directions is minimized to be close to 0.

This decoupling loss reduces the correlation between the editing directions of different attributes, and minimizes the extent to which editing one attribute affects the changes in other attributes. It maximally untangles the coupling between attributes within the latent space.

Cross-assessment Loss: The proposed decoupling loss can increase the independence between the target attribute direction and correlated attribute direction in the latent space, thus preventing the mutual influence of these attribute directions. To further improve the disentanglement between the correlated attributes, we propose a cross-assessment loss, which keeps the invariance of correlated attributes when editing the target attribute. Specifically, with the source image x and edited image $\hat{x}_i$ of attribute i, we compute the assessment score by the assessor $D(\cdot, j)$ of the correlated attribute j before and after editing. To avoid the attribute entanglement in image space, we constrain the invariance of the assessment scores by:

$$\mathcal{L}_{ca} = \sum_{i \neq j} \|D(x, j) - D(\hat{x}_i, j)\|. \tag{14}$$

Total training loss for disentanglement: To disentangle K correlated attributes, we jointly train K attribute directions at the same time with decoupling loss and cross-assessment. To discriminate the target attributes, we employ a shared attribute assessor $D(\cdot, i)$ to assess each target attribute i by modifying the output channel of the last linear layer to the number of attributes followed by a softmax operation, which can avoid increasing the parameter numbers too much. For each iteration, we randomly sample a batch of images from different few-shot attribute datasets $\mathcal{X}_i$ to train the corresponding attribute assessors. Then, we randomly sample a batch of latent codes to train the attribute directions. For each latent code w, we randomly sample an attribute direction v_i (for attribute i) and manipulate the image by:

$$\begin{aligned} G_{edit}(w, i) &= w + l \cdot v_i, \\ \hat{x}_i &= G_{syn}(G_{edit}(w, i)), \end{aligned} \tag{15}$$

where l is a fixed length for manipulation. With the edited image $\hat{x}$, we input it into the shared attribute assessor $D(\cdot, i)$ to maximize the assessing score of the i-th attribute. The adversarial loss is formulated as:

$$\begin{aligned} \mathcal{L}_D = &-\mathbb{E}_{i \in U(1,K), x \in \mathcal{X}_i}[\log(D(x, i))] \\ &- \mathbb{E}_{z \in \mathcal{Z}}[\log(1 - D(\hat{x}_i, i)))], \end{aligned} \tag{16}$$

$$\mathcal{L}_G = -\mathbb{E}_{i \in U(1,K), z \in \mathcal{Z}}[\log(D(\hat{x}_i, i)))], \tag{17}$$

$$\mathcal{L}_{adv} = \mathcal{L}_D + \mathcal{L}_G. \tag{18}$$

We jointly train K attribute directions and the shared attribute assessor by the new adversarial loss in (18), the feature contrastive loss in (10), the identity loss (11), the proposed decoupling loss in (13) and cross-assessment loss in (14). The total training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{fc}\mathcal{L}_{fc} + \lambda_{id}\mathcal{L}_{id} + \lambda_{decouple}\mathcal{L}_{decouple} + \lambda_{ca}\mathcal{L}_{ca}. \tag{19}$$

Based on the new training framework and objectives, our model achieves a good performance in both attribute editing and disentanglement.

D. Disentanglement for Novel Attributes

When there is a novel attribute that we are not aware of which attributes it is entangled with, we propose a new strategy to minimize its entanglement problem. By disentangling the novel attributes $Attr_k$ with all the existing attributes $\{Attr_1, \cdots Attr_n\}$, we can get an attribute direction v_k for the novel attribute, which is disentangled from each known attribute. However, directly following the multi-attribute joint training framework may cost a long training time and unnecessary computations, since all the existing attribute editing directions are already disentangled and contain the desired semantics. Therefore, for novel-attribute disentanglement, in each iteration, we randomly sample one existing attribute $Attr_i$ to disentangle with the novel attribute

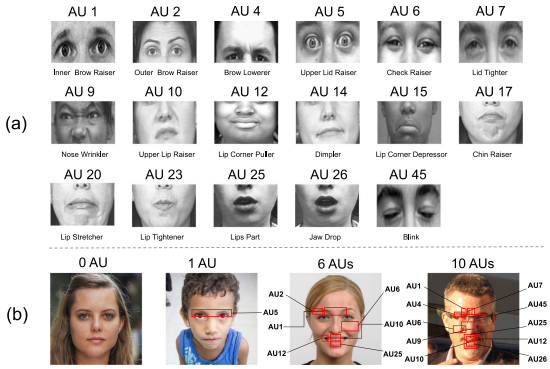


Fig. 5. (a) AU information: 17 action units and their corresponding facial actions. (b) The activated AUs of some example expressions. Most of the expressions are composed of multiple activated AUs.

TABLE III
THE MOST SERIOUS CO-OCCURRENCE RELATIONSHIP OF AU PAIRS IN THE FFHQ AND CELEBA-HQ DATASETS

$AU\ i$	$AU\ j$	$\frac{Num(AU\ i\ \&\ j)}{Num(AU\ i)}$	$\frac{Num(AU\ i\ \&\ j)}{Num(AU\ j)}$	$Average$
AU 12	AU 14	85.03%	85.06%	**85.04%**
AU 6	AU 12	**93.53%**	74.48%	84.00%
AU 12	AU 25	75.05%	**86.08%**	80.57%
AU 6	AU 10	78.45%	80.06%	79.26%
AU 10	AU 12	88.45%	69.01%	78.73%
AU 6	AU 14	87.20%	69.46%	78.33%
AU 6	AU 7	77.81%	77.85%	77.83%
AU 6	AU 25	80.82%	73.81%	77.31%
AU 10	AU 14	85.70%	66.89%	76.30%
AU 7	AU 12	84.28%	67.08%	75.68%

AU 12 and 14 co-occur in 85.04% cases, and when AU 6 is activated, AU 12 is activated simultaneously in 93.53% cases.
The bold values show the maximum frequency of co-occurrence for AU pairs.

$Attr_k$ by the loss in (19). After training, we have a novel attribute editing direction v_k that is disentangled from all the known attributes, which can effectively decrease the coupling problem. The experiments in Section VI-C validate the effectiveness of this strategy, which can minimize the coupling problem to the greatest extent.

V. FINE-GRAINED-ATTRIBUTE EDITING

Based on the correlated-attribute disentanglement-based editing method, we further extend our FEditNet++ for fine-grained attribute editing. In this context, fine-grained refers to two aspects: 1) the changes in edited attributes at the image level are subtle; 2) it has higher requirement for the precision in attribute editing. Fine-grained editing model needs to accurately adjust the corresponding attributes to a specified intensity, in contrast to previous models that can only edit the attributes with relative intensities.

For fine-grained attribute editing, a well-known challenge is to edit facial action units (AUs) in human faces. In the Facial action coding system (FACS) [66], facial muscle movements are decomposed into multiple action units, each with an intensity determined by the activation level of the corresponding muscle. Different combinations of intensities for different action units constitute different facial expressions. Different from facial expression editing based on basic expression types, action units-based editing allows for editing of more subtle expressions and more accurate editing. In our experiments, we primarily focus on fine-grained editing of action units to explore the capability of FEditNet++ for editing fine-grained attributes.

A. Preliminaries - Action Units and AU Co-Occurrence

FACS and AU: Facial action coding system (FACS) [66] was proposed by Friesen and Ekman to encode facial muscle movements by their appearance in the face. FACS decomposes facial expressions into action units (AUs). Each AU describes a contraction or relaxation of one or more muscles. In this paper, we use 17 widely used AUs to represent facial movements and their corresponding facial actions are shown in Fig. 5(a). A facial expression involves one or more muscle movements, and can be represented by one or more activated AUs (Fig. 5(b)).

AU Co-occurrence: In real-face photos, some AUs are strongly correlated. For example, AU 6 (cheek raiser) and AU 12 (lip corner raiser) usually appear together. AU 12 (lip corner raiser) and AU 14 (dimpler) usually exist in the same expression. Table III shows the most serious co-occurrence relationship of some AU pairs, where $Num(AU\ i\ \&\ j)$ represents the number of simultaneous occurrences of AU i and AU j in FFHQ and CelebA-HQ datasets, and $Num(AU\ i)$ represents the number of occurrences of AU i. The last column is the average number of the third and fourth columns. It can be seen that AU 12 and AU 14 co-occur in 85.04% cases, and when AU 6 is activated, AU 12 is activated simultaneously in 93.53% cases. The co-occurrence relationship of these AUs in real face photos makes them difficult to disentangle, which means when editing one target AU, the intensities of its co-occurred AUs will also change, contrary to the purpose of fine-grained expression editing (e.g., single-editing of AU 6).

In this section, we extend our FEditNet++ in fine-grained attribute editing. Our model takes a latent code w with its AU intensities y_s and the target AU intensities y_t as inputs, and edits w with the learned AU directions $\{v_{AUi}\}$ to get the edited latent code w_y. Then, we input w_y into the pretrained StyleGAN generator to synthesize the output image $\hat{x} = G(w)$, which shares the same identity with the source image $x = G(w)$ and has the target AU intensity y_t.

B. Fine-Grained Direction Generator

StyleGAN [3] utilizes $\mathcal{W}$ space to generate images. It has been demonstrated that the mapping from $\mathcal{W}$ space to image features (e.g., hair, age, illumination) can be more linear than the original $\mathcal{Z}$ space [3]. But the mapping from $\mathcal{W}$ space to image features still can not be treated absolutely linear. *E.g.*, if moving the latent code v along v_{AUi} with distance α increases the intensity of AU i with Δy_i, moving w with 2α may not increase the intensity of AU i with $2\Delta y_i$.

Therefore, instead of directly learning an AU editing direction v_{AUi} for each AU i and using the AU intensity as the manipulation length as in (15), we further predict a coefficient for more accurate control. We design a coefficient regression network

(CRN) to regress a coefficient $C_i \in \mathbb{R}$ for each AU direction v_{AUi} ($i = 1, 2, \ldots, 17$). CRN takes the extracted features from an PSP [67] encoder and the relative AU intensity $y_r = y_t - y_0$ as inputs (y_t is the target AU intensity and y_0 is the source AU intensity), and outputs the coefficient for each AU direction. In this way, we can control the intensity of each AU by weighted addition of the AU directions, and obtain AU-conditioned latent direction $v_{dir} \in \mathbb{R}^{18\times512}$ as follows:

$$v_{dir} = \sum_{i=1}^{17} (C_i\, y_r^{(i)}) v_{AUi}, \tag{20}$$

where $y_r^{(i)} = y_t^{(i)} - y_0^{(i)}$ represents the i-th AU's relative AU intensity (target minus source).

With the AU-conditioned latent direction v_{dir}, we can get the edited latent code $w_y = w + v_{dir}$. Then, we feed w_y to the generator to obtain the synthesized image $\hat{x} = G(w_y)$, which has the same identity as the source image $x = G(w)$, and the target AU intensity y_t.

C. AU Assessor with Accurate Intensity Prediction

Different from the previous attribute assessor (Section IV-A2), which only judges whether the target attribute exists in the image, in this section, we introduce AU assessor, which 1) predicts the accurate intensities for AUs, rather than only judging the existence of attributes; 2) uses one network to predict the intensities of all 17 AUs, instead of training an assessor for each AU. Our AU assessor D_{AU} takes an image as input and predicts its corresponding AU intensities for the 17 target AUs. Since this is more challenging than only judging the existence of one attribute, our AU assessor does not employ the pretrained discriminator as the backbone. Instead, it trains from scratch with several convolution layers to extract the image features and a fully-connected layer to map the extracted features into 17 AU intensities. To train the AU assessor, we employ OpenFace [68] to label the AU intensities of the dataset as the ground truth and train the AU assessor to minimize the MSE distance between the labeled and predicted AU intensities.

D. AU Sampler

In previous attribute editing tasks, we can randomly pick up one attribute to train the editing direction and assessor for this attribute. However, in the AU editing task, there exists some contradiction between some pairs of AUs. For example, when AU 12 (lip corner puller) is activated, AU 15 (lip corner depressor) cannot be activated at the same time. Therefore, we cannot take a random AU intensity as the target AU y_t to train our model. To ensure the validity of the target AU intensity, we leverage the estimated AU intensities in the real facial dataset and pick up one as the target AU intensity during training.

However, as shown in Table III, some AUs have a severe co-occurrence problem. For example, AU 6 (cheek raiser) and AU 12 (lip corner raiser) are correlated and co-occur in 85.04% cases. To enable accurate editing of each AU, the model needs to disentangle the correlation, otherwise it is difficult for the model to distinguish between these attributes. To help our model disentangle these AUs, we introduce a random modification of the input target AU y_t. We first randomly pick one AU i. Then, its intensity is set to 0 with 50% probability (to simulate the inactivated case), and is sampled from a $(0, 5)$ uniform distribution with 50% probability. This helps our model to break the co-occurrence relationship and disentangle these AUs more easily.

E. Fine-Grained AU Disentanglement Loss

To disentangle co-occurred AUs, we follow the training framework and objectives in Section IV-C. We employ the feature contrastive loss to keep the face identity unchanged after editing. In addition, we also make use of the adversarial loss to keep the authenticity of the edited image. In addition, we employ the decoupling loss (Section IV-C4) to force the 17 AU directions to be orthogonal to each other:

$$\mathcal{L}_{decouple} = \sum_{i,j} \langle v_{AUi}, v_{AUj} \rangle, \tag{21}$$

where the cosine distance between each pair of AU directions is minimized to be close to 0.

Moreover, to ensure AU editing accuracy, we employ the AU assessor D_y to predict the AU intensities from the edited image, and propose an AU loss to measure the L_2 distance between the predicted AU intensities $D_y(\hat{x})$ and the target AU intensities in y_t:

$$\mathcal{L}_{au}(y_t, \hat{x}) = \|y_t - D_y(\hat{x})\|_2, \tag{22}$$

where $\hat{x}$ is the edited image, and $D_y(\hat{x}) \in \mathbb{R}^{17}$ are the predicted AU intensities. Since the AU assessor predicts the intensities for 17 AUs, which can be regarded as 17 attribute assessors (one for each AU), this AU loss can work as the cross-assessment loss between 17 AUs to disentangle them.

The total training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{fc}\mathcal{L}_{fc} + \lambda_{id}\mathcal{L}_{id} + \lambda_{decouple}\mathcal{L}_{decouple} + \lambda_{au}\mathcal{L}_{au}. \tag{23}$$

VI. Experiments

We conduct experiments on multiple datasets, editing tasks and compare with state-of-the-art methods. We summarize datasets, baselines, evaluation metrics, qualitative and quantitative experiments as follows. In addition, we present implementation details, experimental results on fine-grained attribute editing, more ablation studies, and more comparisons in the Supplementary Material.

Datasets: We evaluated FEditNet++ on: 1) Facial datasets - CelebA-HQ dataset [17], and RaFD dataset [18], 2) Anime dataset - Danbooru2018 dataset [19], and 3) Scene dataset - LSUN Church dataset [20]. For the non-fine-grained editing task, we manually select 30 images from each dataset containing the target attribute as our training dataset. For the fine-grained editing task, we mainly conduct experiments on RaFD dataset, with each AU trained with about 85 images on average.

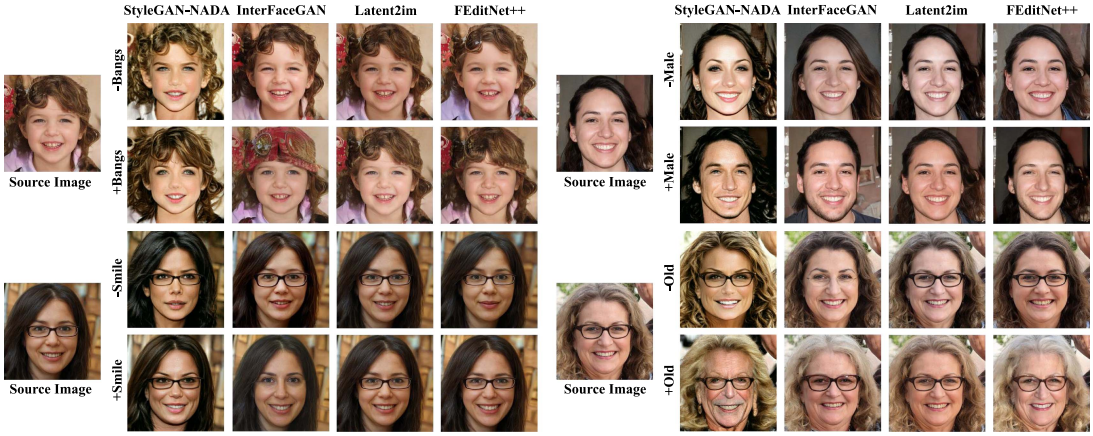


Fig. 6. Qualitative comparison of latent semantics on StyleGAN2 [3] learned by (1) our FEditNet++ trained with 30 data samples, (2) Latent2im [12] trained with pretrained attribute assessor, (3) InterFaceGAN [5] and (4) StyleGAN-NADA (img-guide) trained with 30 pairs of data samples.

TABLE IV
RE-SCORING ANALYSIS OF (A) LATENT2IM [12], (B) INTERFACEGAN [5] FROM A PRETRAINED STYLEGAN2 [3], (C) STYLEGAN-NADA GUIDED BY IMAGES, AND (D) OUR FEDITNET++

	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

(a) Latent2im

	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

(b) InterFaceGAN

	Bangs	Male	Smile	Pose	Young
Bangs	0.27	-0.25	-0.03	0.02	0.46
Male	-0.03	0.41	-0.27	-0.03	0.24
Smile	-0.03	-0.31	0.41	0.06	-0.26
Young	0.01	-0.24	-0.07	-0.05	0.44

(c) StyleGAN-NADA

	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

(d) FEditNet++

Each row shows how the semantic score of images varies before and after editing with a target attribute direction by different predictors.

A. *Few-Shot Attribute Editing*

Comparison Baselines: We compare FEditNet++ with state-of-the-art attribute editing methods: Latent2im [12], InterFaceGAN [5], EditGAN [35], StyleCLIP [36], StyleGAN-NADA [43], AdvStyle [14] and a diffusion-based model combined with Prompt-to-Prompt [54] and Null-text Inversion [69]. The editing directions of FEditNet are trained with only 30 samples, Latent2im pretrained an attribute classifier on CelebA-HQ and InterFaceGAN is trained using 30 pairs of data (including positive and negative samples). StyleGAN-NADA (image-guide) encodes 30 images with target attributes into CLIP space to train an editing model. The details on other comparison methods are provided in Supplementary material.

Evaluation Metrics: To quantitatively compare the editing results, (1) we follow [5] to apply the re-scoring analysis on 2,000 images containing different facial attributes. (2) 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).

We first compare FEditNet with **InterFaceGAN** [5], **Latent2im** [12], and **StyleGAN-NADA (image-guide)** [43], 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 uncorrelated 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"; (3) InterFaceGAN suffers from attribute entanglement severely; and (4) StyleGAN-NADA (image-guide) edits the image in a global manner, changing the identity largely after editing the target attributes. More comparisons with StyleCLIP, StyleGAN-NADA (text-guide), Prompt-to-Prompt, EditGAN, and AdvStyle are in Supplementary material.

We also conduct a quantitative comparison in Table IV and an user study in supplementary material to show that the manipulation quality of our FEditNet surpasses the state-of-the-art methods. We train 5 ResNets as the predictors for attributes Bangs, Male, Smile, Pose and Young, where each predicts the probability of the existence of an attribute and is trained using all images with that attribute in CelebA-HQ dataset. Then, we employ the pretrained ResNets to measure the probability of existence of each attribute, and compare the scores before and after editing. We expect the score of the edited attribute i increases after editing, while the scores of other attributes j, $(j \neq i)$ are unchanged. In Table IV, each row corresponds to the editing of an attribute, and each column reports the difference in scores before and after editing. Our model achieves high score changes of the edited attribute, and low score changes of other attributes, indicating a good editing performance.

We exhibit more attribute-editing experiments on LSUN Church [20] and Danbooru2018 [19] datasets in Fig. 7. We edit the "Short", "Simple", "Tall" and "Complex" attributes for LSUN Church, and edit "Green_Hair", "Mouth_Open", "Green_Eyes" and "No_Bangs" attributes on Danbooru2018. The results demonstrate that our model can edit different attributes in a wide range of datasets with only few-shot training data.

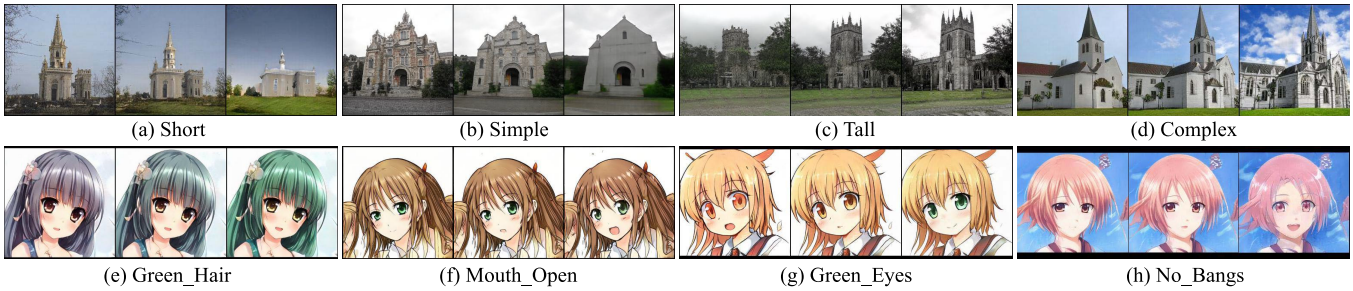


Fig. 7. More attribute editing results on LSUN Church [20] and Danbooru2018 [19] datasets. Our model can achieve a good attribute editing performance on a wide range of datasets.

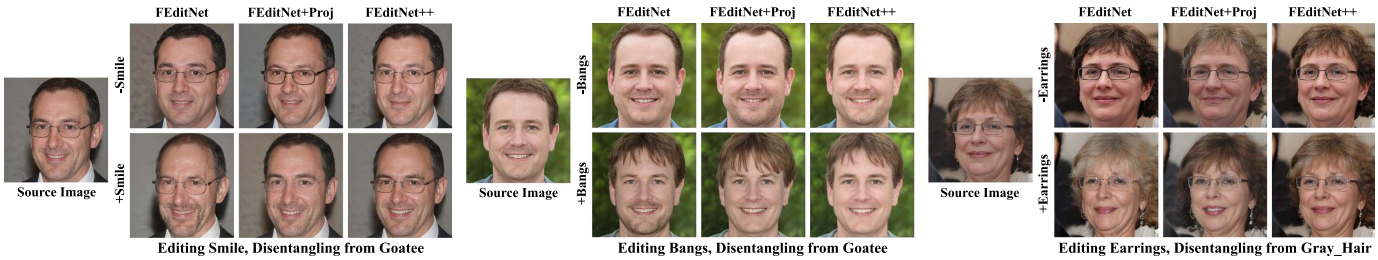


Fig. 8. Comparison on two-attribute disentanglement among (1) FEditNet, (2) FEditNet + Projection, and (3) FEditNet++. FEditNet++ can disentangle the target attributes from the correlated attribute well. The middle one is the original image while the left and right ones are images from latent codes moving in the backward and forward directions, respectively.

B. Correlated-Attribute Disentanglement

As shown in Section VI-A, FEditNet achieved the best performance in few-shot single-attribute editing over the state-of-the-art methods. However, as the editing intensity grows, some correlated attribute entanglement problems may appear. Our FEditNet++ aims to solve the attribute entanglement problem, especially when the editing intensities grow. In this section, we compare our FEditNet++ with FEditNet in editing the entangled attributes. Moreover, as InterFaceGAN has proposed a projection-based decoupling method, we also employ it in FEditNet and compare FEditNet+Projection with our FEditNet++.

1) Disentanglement Between Two Correlated Attributes: To show the effectiveness of FEditNet++ in correlated-attribute disentanglement, we choose the attribute pairs with the most serious entanglement problem, which are (Smile, Goatee), (Bangs, Goatee) and (Wearing_Earrings, Gray_Hair). It is worth noting that some attributes, such as smile and bangs, do not have a severe entanglement problem when the editing intensity is not large (Section VI-A). But as the editing intensity grows, the entanglement problem appears. As shown in Fig. 8, when editing either Smile or Bangs, FEditNet brings some goatee in the human face, and when editing Wearing_Earings, FEditNet turns the hair gray. Therefore, when the editing intensity becomes large, the few-shot editing methods suffer from attribute entanglement. To further validate FEditNet++'s disentangling capability, we employ the projection-based decoupling method in InterFaceGAN in FEditNet++ and compare it with our FEditNet++. As shown in Fig. 8, although FEditNet + projection can disentangle the target attribute from the entangled one, the model suffers from changing other attributes. For example, when editing Smile, FEditNet + projection removes the eyeglasses, and when editing Bangs and Wearing_Earrings, it changes the identity. In comparison, our FEditNet++ disentangles the correlated attributes successfully and maintains the uncorrelated attributes well.

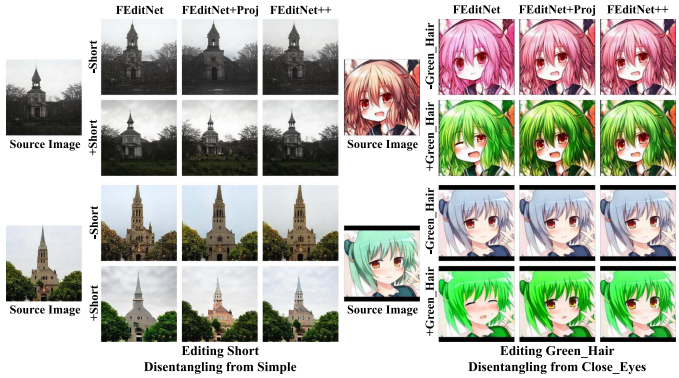


Fig. 9. More disentanglement experiments on the correlated attributes on LSUN Church and Danbooru2018 datasets, with entangled pairs ("Short", "Simple") and ("Green_Hair", "Mouth_Open"). As the editing intensity grows, FEditNet encounters attribute entanglement problem: when editing "Short" on LSUN Church, FEditNet makes the edited church "Simple"; and when editing "Green_Hair" on Danbooru2018, the results are entangled with "Close_Eyes". For FEditNet + Projection, it tends to be entangled with "Complex" in LSUN Church and open the mouths when editing "Green_Hair". In comparison, our FEditNet++ disentangles the correlated attributes well.

We also validate the effectiveness of FEditNet++ in disentangling correlated attributes on LSUN Church and Danbooru2018 datasets in Fig. 9. As the editing intensity grows,

TABLE V
RE-SCORING ANALYSIS OF (A) FEDITNET, (B) FEDITNET + PROJECTION AND (C) OUR FEDITNET++ FOR TWO-CORRELATED-ATTRIBUTE DISENTANGLEMENT

	Smile	Goatee
Smile	0.1677	0.0812
Goatee	-0.3076	0.2847

	Bangs	Goatee
Bangs	0.2889	0.0739
Goatee	-0.0466	0.2847

	Earring	Gray_Hair
Earring	0.1193	0.0860
Gray_Hair	-0.1526	0.3203

(a) FEditNet with 30 data

	Smile	Goatee
Smile	0.1705	0.0397
Goatee	-0.3076	0.2841

	Bangs	Goatee
Bangs	0.2850	-0.0255
Goatee	-0.1060	0.2841

	Earring	Gray_Hair
Earring	0.1339	-0.1219
Gray_Hair	-0.2291	0.3219

(b) FEditNet + projection with 30 data

	Smile	Goatee
Smile	0.2495	0.0025
Goatee	0.0932	0.3370

	Bangs	Goatee
Bangs	0.2901	0.0653
Goatee	-0.0051	0.3440

	Earring	Gray_Hair
Earring	0.1386	0.0009
Gray_Hair	0.0387	0.3611

(c) FEditNet++ with 30 data

Experiments are conducted on three coupled pairs: (Smile, Goatee), (Bangs, Goatee), and (Earring, Gray_Hair). Each row shows how the semantic score of images varies before and after editing with a target attribute direction by different predictors.

the edited results from the original FEditNet tend to be entangled with the correlated attributes. It makes the church "Simple" while only editing the "Short" attribute, and makes the Danbooru2018 "Close_Eyes" while only editing "Green_Hair". With the projection-based decoupling method, FEditNet + Projection tends to be entangled with "Complex" in LSUN Church and open the mouths when editing "Green_Hair". In comparison, our FEditNet++ can disentangle the correlated attributes well and ensure a good editing performance.

We further conduct the quantitative comparison in Table V. We measure the correlated-attribute disentanglement ability of FEditNet, FEditnet + projection-based disentanglement and FEditNet++. In Table V, each row corresponds to the editing of an attribute. A positive score in a column means a higher intensity of the corresponding attribute. From Table V, FEditNet++ outperforms the FEditNet and FEditNet+projection in both target-attribute editing and correlated-attributes disentanglement.

C. Disentanglement for Novel Attributes

In this section, we conduct experiments to demonstrate the effectiveness of the proposed disentanglement strategy for novel attributes in Section IV-D. We conduct an experiment on the four basic attributes "Bangs", "Male", "Smile" and "Old", where we take "Smile" as the novel attribute that we are not aware of its entangling relationship, and the other three attributes as the known attributes. We train the "Smile" attribute in two ways: 1) train by the novel attribute disentanglement strategy (Ours), and 2) train the novel attribute along without disentanglement. The comparison results are shown in Fig. 10. It can be seen that the results of novel attribute disentanglement strategy have a much better editing effect of "Smile", and keep the unrelated attributes ("Male", "Old", "Bangs") unchanged, which means they have been disentangled from the correlated attributes well. Moreover, we also conduct a quantitative experiment in Table VI, which also demonstrates the effectiveness of the novel-attribute disentanglement strategy.

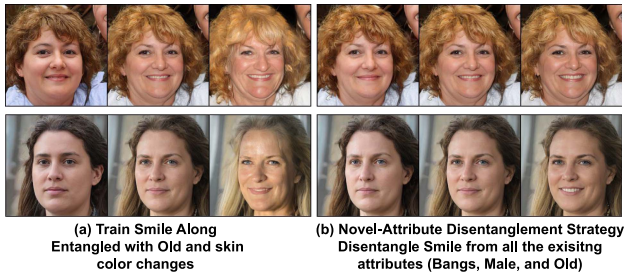


Fig. 10. Novel attribute disentanglement: when disentangling the novel attribute "Smile" with all the existing attributes "Male", "Old", and "Bangs" using our novel-attribute disentanglement strategy, the results are much better than training "Smile" along, which can minimize the entanglement problem.

TABLE VI
DISENTANGLEMENT FOR NOVEL ATTRIBUTES. WE DISENTANGLE THE NOVEL ATTRIBUTE "SMILE" WITH THE THREE EXISTING ATTRIBUTES "MALE", "OLD" AND "BANGS"

Method	Smile (target)	Male	Old	Bangs
Novel-Attribute Disentanglement (Ours)	0.1744	**-0.0539**	**-0.0493**	**-0.0132**
FEditNet	0.1798	-0.0968	0.1066	0.0777

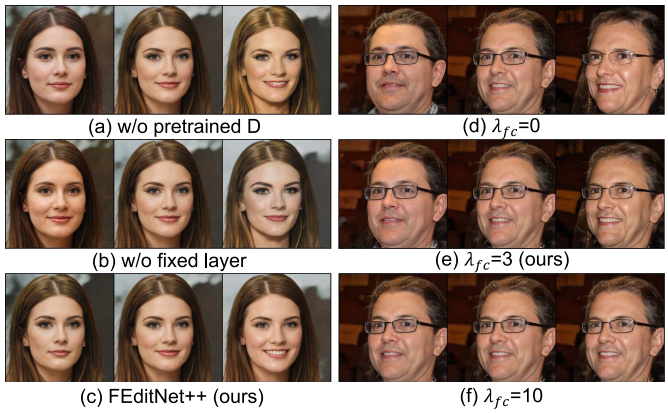


Fig. 11. Ablation study of the pretrained discriminator (a-c) and the weight of the feature contrastive loss λ_{fc} (d-f) on editing "Smile".

D. Ablation Study

1) Ablation Study on Key Components in FEditNet: In this section, we evaluate the effectiveness of each main component in the FEditNet in our preliminary conference work [15], which are: the reuse of the pretrained discriminator and a novel feature contrastive loss.

Fig. 11 compares FEditNet++ with 1) the ablated model without pretrained discriminator D, and 2) the ablated model, which trains the whole attribute assessor without fixing the lower layers. Fig. 11(a-c) shows that the reuse of the discriminator helps FEditNet++ capture the target attribute "Smile", and the freezing of the lower layers can prevent the model from learning too many uncorrelated attributes, so that the editing results are more significant. We also conclude from Fig. 11(d)-(f) that (1) larger λ_{fc} better emphasizes the attribute consistency while

TABLE VII
QUANTITATIVE ABLATION STUDY ON THE PRETRAINED DISCRIMINATOR AND THE WEIGHT OF THE FEATURE CONTRASTIVE LOSS ON EDITING "SMILE"

Editing Smile	Smiling	Bangs	Male	Young	Pose
w/o pretrain	0.1321	-0.0764	-0.1407	-0.1818	0.04
w/o fixed layer	-0.0209	-0.0062	-0.0524	0.1194	0.08
$\lambda_{fc} = 0$	0.3011	-0.126	0.3764	0.3168	0.09
$\lambda_{fc} = 3$ (Ours)	0.3026	-0.0401	0.0044	0.0041	0.02
$\lambda_{fc} = 10$	0.1058	0.0317	-0.0059	0.0038	-0.02

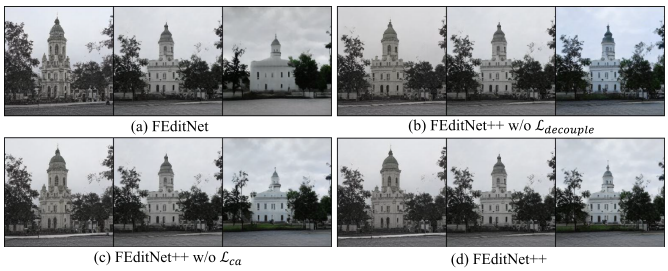


Fig. 12. Ablation study of the decoupling and cross-assessment losses on editing "Short" and disentangling from "Simple".

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 the "Smile" attribute.

Moreover, we also conduct a quantitative comparison on these ablated models in Table VII. It indicates that the model without the pretrained discriminator or without the feature contrastive loss (i.e., $\lambda_{fc} = 0$) cannot keep the correlated attributes unchanged during editing the "Smile" attribute. In addition, the model, without fixing the lower layer of the discriminator, cannot even edit the target attribute. Meanwhile, when the weight of the feature contrastive loss is too high (i.e., $\lambda_{fc} = 10$), the model puts too much attention on keeping the source image feature, thereby cannot edit the target attribute well. In contrast, our model (i.e., $\lambda_{fc} = 3$) edits the "Smile" attribute accurately while keeping the unrelated attributes unchanged.

2) Ablation Study on Difference Between FEditNet and FEditNet++: The major differences between FEditNet++ and FEditNet are joint training of multiple correlated attribute directions with (1) the decoupling, and (2) cross-assessment losses. We perform an ablation study on their effects.

We evaluate the effectiveness of our decoupling loss and cross-assessment loss in attribute disentanglement. We show the editing results of the models without the decoupling loss, cross-assessment loss or both of them in Fig. 12 and Table VIII. (a) Without both of the decoupling and cross-assessment losses, i.e., FEditNet, when editing "Short" on the church the result is entangled with "Simple" severely. (b) For the model without decoupling loss $\mathcal{L}_{decouple}$, it heavily relies on the cross-assessment loss in image space, causing decreased editing effects to keep the "Simple" attribute unchanged. (c) For the model without the cross-assessment loss, it can better maintain the complexity of the original church compared to (a), but still cannot disentangle "Short" from "Simple" thoroughly. (d) In comparison, the model with both losses (Ours) can edit the target attributes well while disentangling the correlated attributes successfully.

TABLE VIII
QUANTITATIVE ABLATION STUDY OF THE DECOUPLING AND CROSS-ASSESSMENT LOSSES ON EDITING "SHORT" AND DISENTANGLING FROM "SIMPLE" ON LSUN CHURCH

Method	Editing Short	
	Δ Short	Δ Simple
w/o $\mathcal{L}_{decouple}$&$\mathcal{L}_{ca}$	0.1225	0.0702
w/o $\mathcal{L}_{decouple}$	0.1224	0.0834
w/o $\mathcal{L}_{ca}$	0.1230	0.0665
FEditNet++	**0.1234**	**0.0575**

Each row shows how the semantic score of images varies before and after editing with a target attribute direction by different predictors.
The bold values show the maximum changes of the semantic score for the target attribute (Short), and the minimum changes of that for the attribute to be disentangled (Simple).

E. Discussions

While our model achieves a good attribute disentanglement performance, our work still has some limitations. For a novel attribute that we are not aware of which attributes it is entangled with, our novel-attribute disentanglement strategy requires enumerating all the existing attributes. However, in real scenarios, we may face challenges to exhaust all potential attributes. In such cases, we may fail to disentangle the novel attributes effectively. This open-set problem is relatively challenging. In future work, we will focus on addressing this type of problem.

VII. CONCLUSION

In this paper, we propose FEditNet++, a GAN-based approach to explore disentanglement-based latent semantics for few-shot 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. We inherit the prior knowledge from the pretrained discriminator network to capture the target attributes and prevent the model from overfitting. Moreover, we propose a novel layer-wise feature contrastive loss to maintain the invariance of the uncorrelated attributes before and after editing.

In particular, FEditNet++ improves upon FEditNet in the following three aspects: (1) To solve the entanglement problem between the correlated attributes caused by the data correlation and latent semantic correlation, we propose to jointly train and disentangle the correlated attributes with new disentanglement losses. (2) We propose a novel decoupling loss and cross-assessment loss to minimize the correlation between correlated attributes in both latent and image space, enhancing the editing accuracy and avoiding the influence of the correlated attributes. We further propose a novel-attribute disentanglement strategy to deal with the editing of novel attributes with unknown entanglements. (3) We extend our FEditNet++ into the fine-grained attribute editing task, enabling our model to edit the subtle attributes in a more accurate way. Extensive quantitative and qualitative experiments have demonstrated the superior capability and great potential of our FEditNet++ in the attribute editing accuracy and correlated attribute disentanglement.

REFERENCES

[1] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of styleGAN," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8107–8116.

[4] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–11.

[5] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9240–9249.

[6] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1532–1540.

[7] A. Plumerault, H. L. Borgne, and C. Hudelot, "Controlling generative models with continuous factors of variations," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–10.

[9] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANspace: Discovering interpretable GAN controls," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9841–9850.

[10] J. Zhu, Y. Shen, Y. Xu, D. Zhao, and Q. Chen, "Region-based semantic factorization in GANs," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 27612–27632.

[11] L. Goetschalckx, A. Andonian, A. Oliva, and P. Isola, "GANalyze: Toward visual definitions of cognitive image properties," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5743–5752.

[12] P. Zhuang, O. O. Koyejo, and A. Schwing, "Enjoy your editing: Controllable GANs for image editing via latent space navigation," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–11.

[13] C. Yang, Y. Shen, and B. Zhou, "Semantic hierarchy emerges in deep generative representations for scene synthesis," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1451–1466, 2021.

[14] H. Yang, L. Chai, Q. Wen, S. Zhao, Z. Sun, and S. He, "Discovering interpretable latent space directions of GANs beyond binary attributes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12177–12185.

[15] M. Xia et al., "Feditnet: Few-shot editing of latent semantics in GAN spaces," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 2919–2927.

[16] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–21, 2021.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.

[18] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cogn. Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[19] D. Anonymous community and G. Branwen, "Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset," 2021, Accessed: Jan. 12, 2021. [Online]. Available: https://www.gwern.net/Danbooru2020

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

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[22] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1510–1519.

[23] Y. Liu, Q. Li, Q. Deng, Z. Sun, and M.-H. Yang, "GAN-based facial attribute manipulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14590–14610, Dec. 2023.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2017, pp. 1125–1134.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[26] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4030–4038.

[27] T. Park et al., "Swapping autoencoder for deep image manipulation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7198–7211.

[28] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.

[29] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN V2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8185–8194.

[30] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, "Gan-control: Explicitly controllable GANs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14083–14093.

[31] J. Zhu et al., "Low-rank subspaces in GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 16648–16658.

[32] N. Spingarn, R. Banner, and T. Michaeli, "GAN "steerability" without optimization," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–11.

[33] Z. He, M. Kan, and S. Shan, "EigenGAN: Layer-wise eigen-learning for GANs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14408–14417.

[34] Y. Wei et al., "Orthogonal jacobian regularization for unsupervised disentanglement in image generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6721–6730.

[35] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler, "EditGAN: High-precision semantic image editing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 16331–16345.

[36] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of styleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2085–2094.

[37] Z. Wu, D. Lischinski, and E. Shechtman, "Stylespace analysis: Disentangled controls for stylegan image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12863–12872.

[38] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the GAN latent space," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9786–9796.

[39] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your GAN: Interactive point-based manipulation on the generative image manifold," in *Proc. ACM SIGGRAPH Conf.*, 2023, pp. 1–11.

[40] J. Zhu, C. Yang, Y. Shen, Z. Shi, D. Zhao, and Q. Chen, "Linkgan: Linking GAN latents to pixels for controllable image synthesis," 2023, *arXiv:2301.04604*.

[41] Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu, "Talk-to-edit: Fine-grained facial editing via dialog," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13799–13808.

[42] Y. Zhang et al., "DatasetGAN: Efficient labeled data factory with minimal human effort," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10145–10155.

[43] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "StyleGAN-nada: Clip-guided domain adaptation of image generators," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–13, 2022.

[44] U. Ojha et al., "Few-shot image generation via cross-domain correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10743–10752.

[45] J. Xiao, L. Li, C. Wang, Z.-J. Zha, and Q. Huang, "Few shot generative model adaption via relaxed spatial structural alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11204–11213.

[46] Y. Zhao, H. Ding, H. Huang, and N.-M. Cheung, "A closer look at few-shot image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9140–9150.

[47] S. Mo, M. Cho, and J. Shin, "Freeze the discriminator: A simple baseline for fine-tuning GANs," in *Proc. CVPR AI Content Creation Workshop*, 2020, pp. 1–6.

[48] Y. Wang, C. Wu, L. Herranz, J. Van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring GANs: Generating images from limited data," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 218–234.

[49] Y. Li, R. Zhang, J. Lu, and E. Shechtman, "Few-shot image generation with elastic weight consolidation," 2020, *arXiv: 2012.02780*.

[50] H. Zhang, Z. Zhang, A. Odena, and H. Lee, "Consistency regularization for generative adversarial networks," 2019, *arXiv: 1910.12027*.

[51] A. Noguchi and T. Harada, "Image generation from small datasets via batch statistics adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2750–2758.
[52] T. Hu et al., "Phasic content fusing diffusion model with directional distribution consistency for few-shot model adaption," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2406–2415.
[53] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[54] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022, *arXiv:2208.01626*.
[55] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[56] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[57] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.
[58] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," in *Proc. SIGGRAPH Asia Conf.*, 2011, pp. 1–10.
[59] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[60] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6706–6716.
[61] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 319–345.
[62] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv: 1807.03748*.
[63] E. Robb, W.-S. Chu, A. Kumar, and J.-B. Huang, "Few-shot adaptation of generative adversarial networks," 2020, *arXiv: 2010.11943*.
[64] C. Yang et al., "One-shot generative domain adaptation," 2021, *arXiv:2111.09876*.
[65] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.
[66] E. Friesen and P. Ekman, "Facial action coding system: A technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, 1978, Art. no. 5.
[67] E. Richardson et al., "Encoding in style: A stylegan encoder for image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2287–2296.
[68] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.
[69] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6038–6047.

Ran Yi received the BEng and PhD degrees from Tsinghua University, China, in 2016 and 2021. She is an assistant professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. Her research interests include computer vision, computer graphics, and computational geometry.

Teng Hu received the BEng degree from the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, in 2022. He is currently working toward the PhD degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research interests include computer vision and computer graphics.

Mengfei Xia received the BSc degree from the Department of Mathematical Science, Tsinghua University, China, in 2020. He is currently working toward the PhD degree with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include deep learning, image processing, and computer vision.

Yizhe Tang received the BEng degree from the School of Computer Science and Engineering, Central South University, China, in 2023. He is currently working toward the master's degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research interests include computer vision and computer graphics.

Yong-Jin Liu (Senior Member, IEEE) received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. He is a tenured full professor with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include computer vision, computer graphics, computer-aided design, intelligent information processing of media, and pattern analysis. For more information, visit http://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.htm.