The Visual Computer (2023) 39:6005–6014
https://doi.org/10.1007/s00371-022-02708-7

ORIGINAL ARTICLE

# Disentangled face editing via individual walk in personalized facial semantic field

Chengde Lin[1] · Shengwu Xiong[1,2] · Xiongbo Lu[1]

Accepted: 12 October 2022 / Published online: 3 November 2022


## Abstract

Recent generative adversarial networks (GANs) can synthesize high-fidelity faces and the closely followed works show the existence of facial semantic field in the latent spaces. This motivates several latest works to edit faces via finding semantic directions in the universal facial semantic field of GAN to walk along. However, several challenges still exist during editing: identity loss, attribute entanglement and background variation. In this work, we first propose a personalized facial semantic field (PFSF) for each instead of a universal facial semantic field for all instances. The PFSF is built via portrait-masked retraining of the generator of StyleGAN together with the inversion model, which can preserve identity details for real faces. Furthermore, we propose an individual walk in the learned PFSF to perform disentangled face editing. Finally, the edited portrait is fused back into the original image with the constraint of the portrait mask, which can preserve the background. Extensive experimental results validate that our method performs well in identity preservation, background maintenance and disentangled editing, significantly surpassing related state-of-the-art methods.

**Keywords** Face editing · Personalized facial semantic Field · Identity preservation · Disentangled facial manipulation · Generative adversarial networks · StyleGAN

## 1 Introduction

Face editing task aims to manipulate a facial attribute toward the desired status, such as adding age or adding smiling (see Fig. 1). Real facial semantic editing is needed in extensive applications, but there are still challenges [1]. Extensive works have been proposed: such as earlier algorithms [2,3] based on three-dimensional morphable face models [4] (3DMMs) and methods [5,6] based on improved conditional generative adversarial networks (CGANs).

Recently, generative adversarial networks (GANs) [7–10] have made impressive strides in generating realistic high-resolution face images. Walking in the latent space of a pretrained facial GAN in appropriate directions can result in facial attribute variation. This phenomenon implies that there is abundant facial semantic information determined by the latent space together with GAN models, which is termed as facial semantic field (FSF) in this paper. Thus, several recent studies [1,11–17] propose to edit face based on the priors from pretrained famous GANs. These works show effectiveness in face editing via finding a semantic direction and then moving toward it, but three challenges remain: identity loss, background alteration and entangled manipulation. For example, adding eyeglasses via InterFaceGAN [11] may remove the beard and background by mistake (see the first row of Fig. 2). In addition, adding facial age via InterFaceGAN [11] may also put on eyeglasses (see the second row of Fig. 2). The possible reason is that prior works utilize fixed pretrained GANs to edit all faces, ignoring individuals' differences. For each facial attribute, the universal walking direction is insufficient for disentangled editing for different individuals.

✉ Shengwu Xiong
xiongsw@whut.edu.cn

Chengde Lin
linchengde@whut.edu.cn

Xiongbo Lu
luxiongbo01@gmail.com

[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, 122 Luoshi Road, Wuhan 430070, Hubei, China

[2] Sanya Science and Education Innovation Park of Wuhan University of Technology, Wuhan University of Technology, Yazhouwan, Sanya 572000, Hainan, China

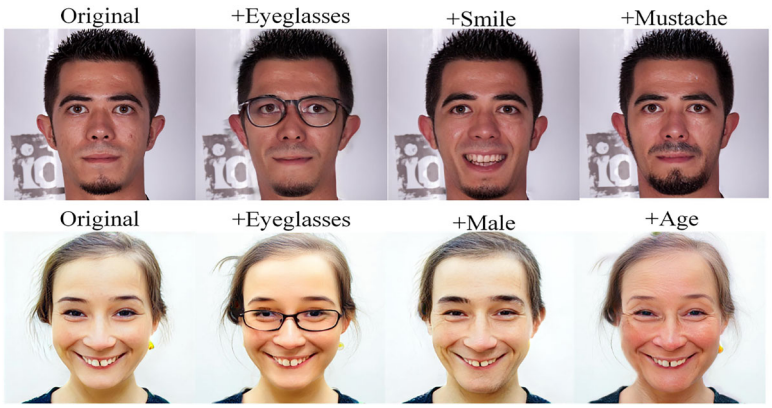


**Fig. 1** Different editing examples (in resolution 1024 × 1024) from our method. The background and identity are kept well after disentangled editing. More edited examples are public on our project page at https://github.com/lcd21/PFSF

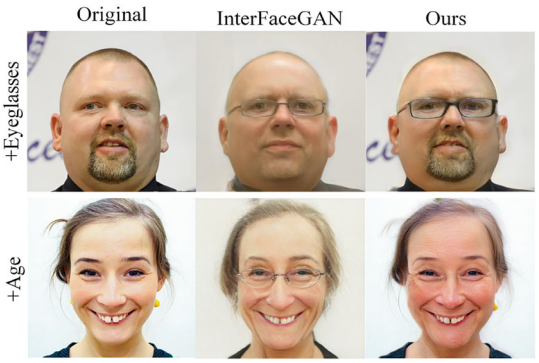


**Fig. 2** Visual comparison on different editing results. Notice the background and other attributes

To address the above limitations, we propose to perform disentangled face editing by learning an individual walk in a personalized facial semantic field instead of a universal facial semantic field for all faces as peers [11,18,19]. The motivation is that there are personalized differences during facial semantic changing. For example, old people may smile in a quite different manner in contrast to children. Thus, personalized walk can leverage more individual characteristics, stimulating more identity preservation and semantic disentanglement. Our personalized walk framework consists of three steps: (1) the generator of StyleGAN and the inversion model are retrained, producing a personalized facial semantic field. The PFSF is built via portrait-masked constraints with more personalized facial details for each instance. Sampling from the personalized facial semantic field, the generator can synthesize similar faces to the original ones compared to the universal facial semantic field from the fixed GANs model used by peers [11,18,19]. In other words, with the help of the learned personalized facial semantic field, more identity information is preserved. (2) For individual semantic manipulation with disentanglement, we learn to walk in the personalized facial semantic field to manipulate the objective attribute but preserve the other attributes. This disentangled semantic walk is supervised via an attributes predictor. (3) The edited portrait is integrated into the background of the given image with the constraints of a portrait mask, which can preserve the background.

Broad experiments are performed on the public datasets FFHQ [9] and CelebaHQ [7] to assess the proposed method. Results demonstrate that our method can preserve the background well and surpass recent state-of-the-art works on identity preservation and disentangled editing.

The major contributions of this work are listed as follows.

1. A personalized facial semantic field is built for individuals retraining the GAN model with the retention of identity and perception as optimizing constraints. This help to preserve more facial identity during inversion.
2. The portrait mask is introduced as a constraint to both the PFSF building and the edited image fusion. This can maintain the background and preserve facial details.
3. An individual walk in the PFSF is proposed to perform disentangled semantic manipulation.
4. A framework consisting of the above components is constructed for real face editing, surpassing existing excellent methods in both quantitative and qualitative evaluation.

## 2 Related work

GAN inversion searches for a latent vector in the latent space of a pretrained GAN, from which a new image similar to the given image can be generated. Current works for GAN inversion are in three types. The first applies optimizing strategy [20–25] to find the right latent code by optimizing an initial latent code directly via irritation. Image2StyleGAN [22] is a typical optimization-based work, optimizing on the latent vector via gradient descent. Optimization-based methods can achieve decent inversions, but they are time-consuming. The second is based on learning strategy [26–30]. They train a network to encode the given image into the latent vector by minimizing the difference between the inverted image and the

original one. pSp [31] trains an network to embed vectors of different styles in W+ space of StyleGAN. Inversion works based on learning are fast in the inference stage, but they lose identity for wild real faces. The last applies hybrid techniques [32–35] to combine the above two strategies. IDInv [32] trains a domain-guided network to encode given images into the latent vectors, which are used as the initialization for the following optimization. These existing inversion works only perform on a fixed universal semantic field, but we consider the personalized differences, and the proposed method performs the inversion and retraining of the GANs simultaneously, producing a personalized semantic field for precise attribute editing.

Recent works [1,11–17,36–38] use the priors in pretrained GANs for face editing. InterfaceGAN [11] searches for a hyperplane to divide a specific attribute from one status to another (e.g., from male to female) and moves to the orientation perpendicular to the hyperplane. SGF [14] and HijackGAN [13]design surrogate networks to learn the semantic gradient in the latent space of pretrained generators, showing disentangled editing to some extent, but suffer the loss of identity and background. GANSpace [12] is a typical unsupervised method, searching for the attribute direction via Principal Component Analysis (PCA). IALS [18] learns attribute-level direction for face editing and proposes a disentanglement transformation to achieve disentanglement in pairs, but entanglements still exist among multiple attributes. Trans4edit [19] trains a transformer to map semantic variations into the latent vector variations. FacialVideoEditing [36] conducts face editing in videos at high resolution via combining GAN inversion and attributes manipulation. Multi-view-face [38] can generate multi-view faces via unpaired images, avoiding large-scale data collection and annotation.

## 3 Proposed method

Our facial editing framework for real faces is composed of building a personalized facial semantic field (PFSF), learning individual walks for editing, and edited face fusion. An overview of our framework is shown in Fig. 3. First, a personalized facial semantic field for each real face is built by optimizing the inversion and retraining the GAN model (e.g., StyleGAN) together, preserving more individual information. Then, an individual walk in the above personalized semantic field is conducted by searching the target semantic direction step by step, guided by the pretrained facial attribute predictor. Finally, the edited face is fused into the original image via a portrait mask.

Current facial editing methods usually fail to real face images not generated by GANs. They are based on fixed inversion and fixed GAN, resulting in a universal facial semantic field. This implies that it is a challenge to embed real face images into a universal semantic field, which motivates us to conduct GAN inversion and GAN retraining simultaneously, building a personalized facial semantic field for the given image. Methods based on the UFSF ignore the personalized difference between individuals and walk in a fixed universal semantic field along a single linear path for all faces. By contrast, our method learns a personalized facial semantic field for each face independently and then walks in it.

Facial attributes $A = \{a_1, a_2, \cdots, a_N\}$ can be quantified as semantic scores $S = \{s_1, s_2, \cdots, s_N\}$, where $s_i \in R$, $N$ denotes the considered attributes number. For each attribute, different scores mean different semantic intensities. Taking attribute aging for an example, the higher attribute score means the older face. The gradient of the semantic scores $\nabla S$ is related to the inversion module $I$ and GAN model $G$, then the personalized facial semantic field can be defined as

$$F(I, G) = \nabla S. \tag{1}$$

For a given face $x$, its personalized facial semantic field $F$ is built by learning personalized inversion module $I$ and GAN model $G$. Then, the latent vector of $x$ in the $F$ is $z = I(x)$. We need to find a semantic direction $dz$ for $z$ to walk from $z_m$ to $z_n$ step by step in the learned personalized facial semantic field $F$, driving the corresponding attribute score altered from the original attribute score $s_m$ to the target one $s_n$ gradually. This semantic walk is expressed as

$$s_n - s_m \sim \sum_{i=m}^{n} (z_i + dz_i)\,. \tag{2}$$

### 3.1 Building personalized facial semantic field

GAN inversion $I$ aims to find a latent vector $z = I(x)$, where the corresponding generator $G$ can reconstruct a new image $\hat{x} = G(I(x))$, whose difference from the original image $x$ is as little as possible. Since the background of the face image needs not to be changed during editing, it is not appropriate to treat the portrait and the background equally. So we propose to focus on the portrait, neglecting the background. We introduce the portrait matting method M [39] to extract the portrait mask $m$ of the original image $x$: $m = M(x)$. Then, the portrait in the original image $x$ can be represented as $x_p = x \cdot m$, and the portrait in the reconstructed image $\hat{x}$ can be expressed as $\hat{x}_p = \hat{x} \cdot m$. The pixel reconstruction loss and perceptual loss are used to minimize the differences between the inverted image $\hat{x}$ and the original image $x$. The objective is expressed as

$$\begin{aligned} L_{\text{inv}} &= L_{\text{pixl}} + \lambda_1 L_{\text{pert}} \\ &= L_2(x, \hat{x}) + \lambda_1 L_{pert}(x, \hat{x}) \end{aligned}$$

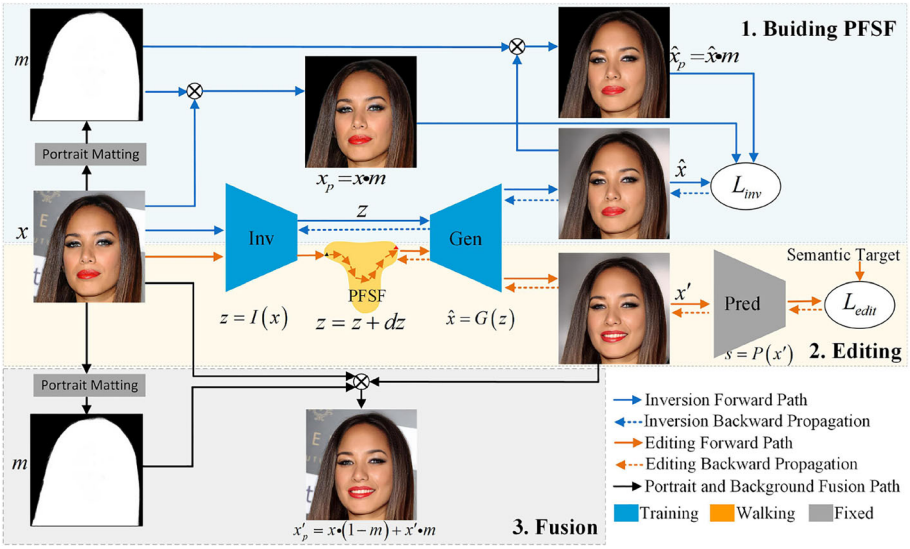


**Fig. 3** Framework of our method. For a given face image, the personalized facial semantic field (PFSF) is built by inverting it into the latent space of GAN and retraining GAN together, which is colored blue. After the PFSF is built, the generator of GAN is fixed, and a pretrained facial attribute classifier is used as supervision for searching disentangled semantic direction in the PFSF. This individual walk aims to walk in the right semantic direction, making the given face image edited according to the semantic target (e.g., smiling). The editing process is colored in orange. Finally, the edited portrait is fused with the background of the original image

$$= L_2(M(x)\cdot x, M(x)\cdot G(I(x))) + \lambda_1 L_{pert}(M(x)\cdot x, M(x)\cdot G(I(x))) \quad (3)$$

where the perceptual loss $L_{pert}$ is from work [40], and the reconstruction $L2$ denotes the mean square error (MSE). Building PFSF is marked as light blue at the top of Fig. 3.

The inversion module $I$ and generator $G$ are parameterized by $\theta_I$ and $\theta_G$, respectively, then the personalized facial semantic field can be learned by the following objective function.

$$\theta_I^*, \theta_G^* = \arg\min_{\theta_I,\theta_G} (L_{\text{inv}}) \quad (4)$$

### 3.2 Individual semantic walk for facial editing

After the personalized facial semantic field is built for the given image $x$, the personalized generator $G$ and corresponding latent vector $z = I(x)$ are learned. The generated image $G(z)$ shows little difference to the original face $x$. Then, the synthetic image $G(z)$ can be manipulated into the semantic target (e.g., smiling or aging) via $z$'s walk in the personalized facial semantic field in the proper semantic direction. For disentangled and precise facial editing, we propose to walk individually, instead of the universal and linear manner used by current methods.

Our strategy for disentangled semantic direction consists of two steps: first, preliminary semantic directions for extensive instances sampled from the personalized facial semantic field are obtained by the gradient of attributes classifiers. Then, the average of the above preliminary semantic directions is seen as the final direction. The disentangled semantic direction $dz$ in the personalized facial semantic field should guarantee that while a latent vector $z$ walks along it, the generated image $G(z + dz)$ should change only one attribute as desired and preserve the rest. N binary facial attributes ($a \sim \{0, 1\}$) are listed as $A = \{a_1, a_2, \cdots, a_N\}$. For example, $a_3 9 = 1$ means young and $a_3 9 = 0$ means old. The attributes $A$ of a face image $x$ is recognized via the facial analysis model $P$ [41]. This step is expressed as $A = P(x)$.

The cross-entropy loss $\mathcal{L}_{\text{tar}}$ is used to drive the selected attribute from the original status to the objective status.

$$\mathcal{L}_{\text{tar}} = -y_t \log(a_t) - (1 - y_t)\log(1 - a_t), \quad (5)$$

where $a_t = P(G(z + dz))[t]$ is the attribute probability predicted by $P$, and $y_t$ means the objective label of the target attribute.

The MSE loss is adopted to guarantee other attributes unchanged, which is critical for disentangled editing. Let $A_{other} = \{a_1, a_2, \cdots, a_i\}_{i \neq t}$ be other attributes except target attribute $a_t$ and $Y_{other} = \{y_1, y_2, \cdots, y_i\}_{i \neq t}$ be the corresponding label values. Then, the loss to preserve other attributes can be represented as

$$L_{\text{other}} = \frac{1}{N-1} \| A_{\text{other}} - Y_{\text{other}} \|_2 . \quad (6)$$

The individual semantic direction can be searched via the following objective.

$$dz_x = \arg\min_{dz} (\lambda_{\text{tar}} L_{\text{tar}} + \lambda_{\text{other}} L_{\text{other}}), \quad (7)$$

where $\lambda_{\text{tar}}$ and $\lambda_{\text{other}}$ denote the weights to control the two losses' contributions.

Then, the instance semantic direction $dz$ can be searched by minimizing the loss $L_{\text{nav}}$. And the final semantic direction can be obtained via averaging those instance semantic

directions from multiple samples $z_s$ as

$$dz = \frac{1}{S}\sum_{s=1}^{S} dz_s, \tag{8}$$

where S means the number of images sampled from the SFPF.

### 3.3 Portrait fusion

After the individual semantic walk, the given face has been edited, but its background has also been altered. We extract the background of the original image $x$ and the portrait of the edited image $x'$ as $x \cdot (1-m)$ and $x' \cdot (m)$, respectively. Then, the finally edited image with the background maintained can be illustrated as

$$x'_p = x \cdot (1-m) + x' \cdot m, \tag{9}$$

where $m$ is the portrait mask of the original image. The portrait fusion is marked as light gray at the bottom of Fig. 3.

The overall pipeline of our method is summarized in Algorithm 1.

**Algorithm 1** Disentangled Face Editing via Walk in Personalized Facial Semantic Field

**Require:**
A real face image $x$, a GAN generator $G(\cdot)$ parameterized by$\theta_G$, a pretrained inversion model $I$ parameterized by$\theta_I$, a facial attributes classifier $P$, portrait matting model $M$.
The portrait mask $m = M(x)$
**while** *not converged* **do**
  $L_{\text{inv}} = L_{L2}(m \cdot x, m \cdot G_{\theta_G}(I_{\theta_I}(x))) + \lambda_1 L_{pert}(m \cdot x, m \cdot G_{\theta_G}(I_{\theta_I}(x)))$
  $\theta_I^*, \theta_G^* = \arg\min_{\theta_I, \theta_G}(L_{\text{inv}})$
**end while**0
**while** $s < S$ **do**
  $\hat{x} = G_{\theta_G^*}(z_s)$
  $a_t = P(G(z_s))[t]$
  $\mathcal{L}_{\text{tar}} = -y_t \log(a_t) - (1-y_t)\log(1-a_t)$
  $A_{other} = \{a_1, a_2, \cdots, a_i\}_{i \neq t}$ be other attributes except target attribute
  $a_t$ and $Y_{other} = \{y_1, y_2, \cdots, y_i\}_{i \neq t}$ be the corresponding label values
  $L_{\text{other}} = \frac{1}{N-1}\|A_{\text{other}} - Y_{\text{other}}\|_2$
  $dz_s += \arg\min_{dz}(\lambda_{\text{tar}} L_{\text{tar}} + \lambda_{\text{other}} L_{\text{other}})$
  $s = s + 1$
**end while**
The target semantic direction $dz = \frac{1}{S}\sum_{s=1}^{S} dz_s$
The inverted vector $z = I_{\theta_I^*}(x)$
The edited image $x' = \theta_G^*(z + k * dz)$
**Ensure:** The fused image after edited $x'_p = x \cdot (1-m) + x' \cdot m$

## 4 Experiments

### 4.1 Experimental settings

Our experiments are performed on the subsets from the face datasets FFHQ [9] and dataset CelebaHQ [7]. The dataset FFHQ consists of 70,000 real faces in 1024 × 1024 resolution. The dataset CelebaHQ is made up of 30,000 real faces in 1024 × 1024 resolution. The facial attribute classifier is obtained from the pretrained ResNet50 [42] for facial attributes recognition by the work [41]. While building the PFSF, the inversion is conducted via optimization as Image2Style [20]. The learning rate and iteration times for fine-tuning the generator are valued as 0.1 and 100, respectively. The total number of the attributes is 40 as the dataset CelebA [43]. The hyperparameters $\lambda_{tar}$, and $\lambda_{pre}$ in Eq. (7) are assigned 1 and 2. Our experiment is on a single NVIDIA 1080Ti GPU.

**Benchmarks** The proposed method is compared to the recent famous works, including InterfaceGAN [11], GANSpace [12], IALS [18] and Trans4edit [19]. The pretrained generator of StyleGAN [9] is adopted as the backbone.

**Metrics** Our method is evaluated in terms of identity preservation and disentanglement. Identity preservation is assessed in the same way as InterfaceGAN [11], adopting cosine similarity between the identity features of the original and the edited image. The identity features are obtained from the face recognition framework [44]. A higher identity preservation score means that more identity information is preserved. The disentanglement is evaluated by the cross-entropy of the edited attribute scores and the target attribute scores. A lower disentanglement score means that attributes are less entangled.

### 4.2 Qualitative evaluation

The visual comparisons among edited images from the proposed approach and the decent benchmarks are demonstrated in Fig. 4. We can see our method achieves much better performance in terms of identity preservation. After editing, the personalized facial details are preserved well by our method but are easy to be lost by compared works (e.g., see the mouth, makeup, headdress and painting from the first row to the fourth row, respectively). This benefits from editing on the personalized facial semantic field built for each face instead of editing on a UFSF for all faces as peers [11,18,19]. Fine-tuning the generator can enhance its ability to reconstruct personalized features that even do not appear in the original training dataset (see Fig. 5). Furthermore, our method shows better disentanglement. For instance, aged results in the fifth row and the sixth row show that GANSpace [12] and InterfaceGAN [11] both add age accompanied with

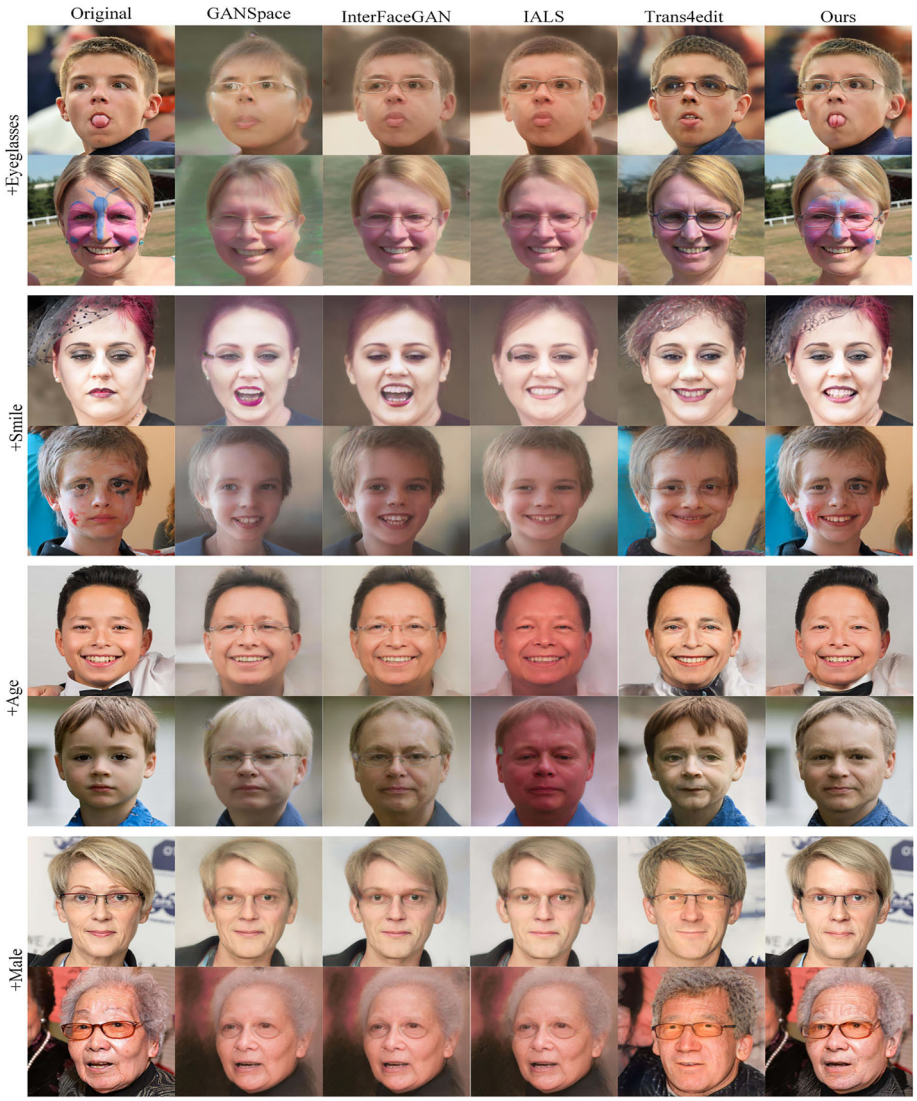


**Fig. 4** Visual comparison. Each column means edited image from the following method, respectively: (2) GANSpace [12], (3) InterfaceGAN [11], (4) IALS [18], (5) Trans4edit [19], (6) Our method

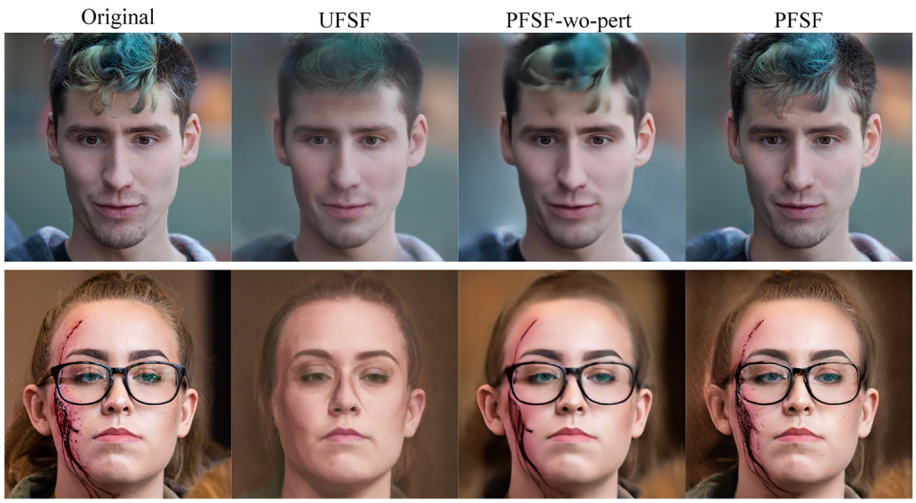


**Fig. 5** Reconstructed examples of inversions from the universal facial semantic field (FSF) and personalized facial semantic fields (PFSF) before the background fusion. It is obvious that reconstructions from PFSF (Ours) preserve more detailed characteristics and show more similarity to the original examples

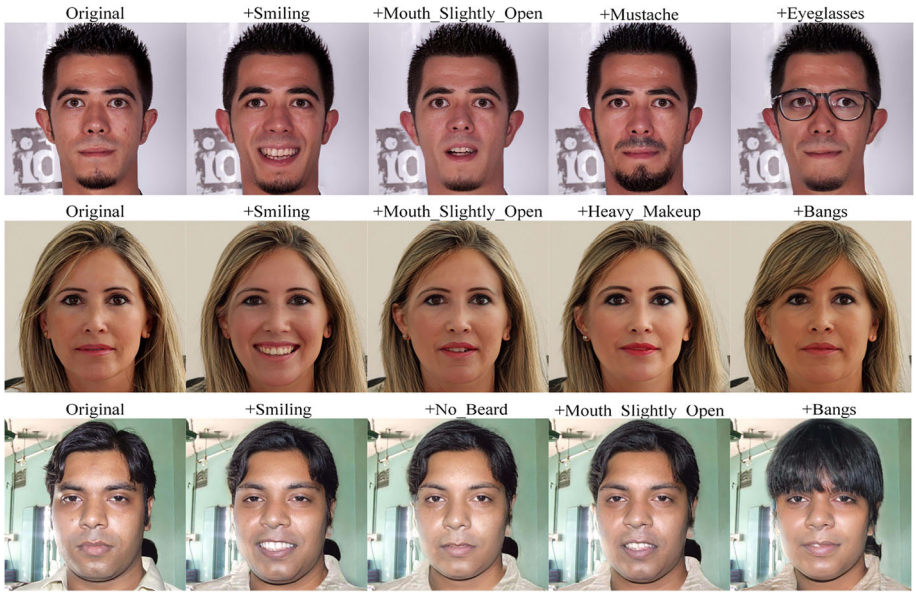


**Fig. 6** Different edited examples from our method. Notice that the other attributes and the background are preserved well

adding eyeglasses; by contrast, our method can make faces aging without eyeglasses added. When manipulating gender, our method maintains eyeglasses and smiling (the last two rows) unchanged, however, the compared works fail to keep them. Results from GANSpace [12] seem not as competitive as other works, because GANSpace searches for semantic direction in an unsupervised manner while the others in a supervised manner. In addition, from the first two rows and the last two rows, we can see that the background of edited faces from our method is the same as the original images; however, edited results from compared works fail to reconstruct the corresponding background well. More edited examples are listed in Fig. 6. All these qualitative findings are consistent to the quantitative evaluation in the next section.

### 4.3 Quantitative evaluation

To perform a fair quantitative comparison on the subset of datasets FFHQ and CelebaHQ, the same four attributes as the benchmarks are listed. The symbol – denotes that the method is often unable to edit the objective attribute as desired. Tables 1 and 2 demonstrate that identity preservation scores of edited images from our method are much higher than those compared methods. This denotes that more identity features are kept by the proposed method. Among different attributes editing results from each method, it is obvious that adding smiling obtains the best identity preservation score, but manipulating gender gets the lowest one. This is because changing the gender needs to change the whole face, however, adding smiling just needs small variations on the mouth and eyes.

From Tables 3 and 4, our method demonstrates higher disentanglement scores than benchmarks. This denotes that our method can preserve the other attributes well, while the target attribute is edited as desired. Except for our approach, Trans4edit [19] also obtains competitive performance in disentangled editing, which benefits from the disentanglement loss designed to train the latent transformer. Among different edited attributes, disentanglement scores are the worst while adding age. Since editing gender needs massive changes of facial appearances, inevitably altering some other attributes. On the other hand, adding eyeglasses only needs to change

**Table 1** Quantitative evaluation of identity preservation on data from FFHQ [9]. A–D in the first row mean adding eyeglasses, adding smiling, aging and adding male, respectively

| Method | A | B | C | D |
|---|---|---|---|---|
| GANSpace [12] | – | 0.432 | 0.335 | 0.358 |
| InterfaceGAN [11] | 0.411 | 0.493 | 0.452 | 0.367 |
| IALS [18] | 0.405 | 0.522 | 0.478 | 0.348 |
| Trans4edit [19] | 0.498 | 0.612 | 0.514 | 0.386 |
| Ours | **0.551** | **0.707** | **0.643** | **0.494** |

Bold values indicate the best results

**Table 2** Quantitative evaluation of identity preservation on data from CelebaHQ [7]. A–D in the first row mean adding eyeglasses, adding smiling, aging and adding male, respectively

| Method | A | B | C | D |
|---|---|---|---|---|
| GANSpace [12] | – | 0.465 | 0.301 | 0.298 |
| InterfaceGAN [11] | 0.402 | 0.525 | 0.309 | 0.318 |
| IALS [18] | 0.397 | 0.529 | 0.361 | 0.291 |
| Trans4edit [19] | 0.432 | 0.570 | 0.394 | 0.315 |
| Ours | **0.497** | **0.654** | **0.459** | **0.371** |

Bold values indicate the best results

**Table 3** Quantitative comparison on semantic disentanglement on dataset FFHQ [9]. A–D in the first row mean adding eyeglasses, adding smiling, aging and adding male, respectively

| Method | A | B | C | D |
|---|---|---|---|---|
| GANSpace [12] | – | 0.442 | 0.410 | 0.474 |
| InterfaceGAN [11] | 0.266 | 0.412 | 0.270 | 0.459 |
| IALS [18] | 0.269 | 0.416 | 0.254 | 0.447 |
| Trans4edit [19] | 0.251 | 0.382 | 0.249 | 0.428 |
| Ours | **0.213** | **0.357** | **0.227** | **0.391** |

Bold values indicate the best results

**Table 4** Quantitative comparison on semantic disentanglement on dataset CelebaHQ [7]. A–D in the first row mean adding eyeglasses, adding smiling, aging and adding male, respectively

| Method | A | B | C | D |
|---|---|---|---|---|
| GANSpace [12] | – | 0.418 | 0. 459 | 0.641 |
| InterfaceGAN [11] | 0.276 | 0.383 | 0.438 | 0.597 |
| IALS [18] | 0.283 | 0.396 | 0.439 | 0.632 |
| Trans4edit [19] | 0.268 | 0.369 | 0.420 | 0.577 |
| Ours | **0.235** | **0.339** | **0.390** | **0.539** |

Bold values indicate the best results

the local region around the eyes and does not affect other regions, so editing this attribute wins the best disentanglement.

### 4.4 Ablation analysis

We conduct the ablation experiment by performing three versions of our method to test the effectiveness of different components. To test the portrait mask's contribution to identity preservation, we set the portrait mask as a unit matrix to remove its effect. This version of our method is marked as PFSF-wo-mask. In addition, we set the $\lambda_1$ as 0 to remove the loss $L_{pert}$ in Eq. (3) to test the perceptual loss's contribution to identity preservation. This version is labeled as PFSF-wo-pert. The corresponding experimental results in terms of identity preservation are listed in Table 5. It is obvious from the first two rows of Table 5. that our method with PFSF can improve the identity preservation score by a wide margin in contrast to the version with UFSF. The last three rows in Table 5. demonstrate that the introduced portrait mask, and the perceptual loss can improve identity preservation, respectively. It is accordant with the visual comparison in Fig. 5. The third and the fourth column in Fig. 5. show that our method with the perceptual loss $L_{pert}$ preserves more personalized features (e.g., the colored hair and the scar).

Furthermore, we set $\lambda_{\text{other}}$ in Eq. (7) as 0 to test the contribution of the loss $L_{other}$ to disentanglement. Our method without the loss $L_{other}$ is marked as PFSF-wo-Lother, and the experimental results are listed in Table 6. Table 6. shows

**Table 5** Identity preservation comparison on different versions of our method. A–D in the first row mean adding eyeglasses, adding smiling, aging and adding male, respectively

| Method | A | B | C | D |
|---|---|---|---|---|
| UFSF | 0.405 | 0.522 | 0.478 | 0.348 |
| Ours (PFSF-wo-pert) | 0.519 | 0.667 | 0.617 | 0.453 |
| Ours (PFSF-wo-mask) | 0.535 | 0.688 | 0.630 | 0.477 |
| Ours | **0.551** | **0.707** | **0.643** | **0.494** |

Bold values indicate the best results

**Table 6** Facial attribute disentanglement comparison on different versions of our method. A–D in the first row mean adding eyeglasses, adding smiling, aging and adding male, respectively

| Method | A | B | C | D |
|---|---|---|---|---|
| UFSF | 0.269 | 0.416 | 0.254 | 0.447 |
| Ours (PFSF-wo-Lother) | 0.255 | 0.407 | 0.249 | 0.424 |
| Ours | **0.213** | **0.357** | **0.227** | **0.391** |

Bold values indicate the best results

that disentanglement scores are improved significantly with constrain $L_{other}$ added. This means the constraint $L_{other}$ contributes to keeping other attributes during editing the target attributes, and it is valid for disentangled editing.

### 4.5 Limitations

Several limitations exist in this work. First, limited by the 40 binary annotations of facial attributes in the face dataset CelebA [43], our method cannot edit other semantic attributes (e.g., pose or illumination) beyond the above annotations. In addition, our method needs to retrain the generator for each face to build the personalized facial semantic fields. It takes 60 s to build a personalized facial semantic field and 30 min to search the semantic direction (when the sampling number $S = 5,000$ in Eq. 8). This is time-consuming, and it is not appropriate for real-time application. Since the experiment is conducted on a single NVIDIA GTX 1080Ti GPU, this time-consuming limitation would be alleviated on GPUs with more memory or with higher performance.

## 5 Conclusion

This paper presented a disentangled face editing method via walking in the personalized facial semantic field. We build personalized facial semantic fields for individuals via retraining the GAN model with the retention of identity and perception as optimizing constraints. Then, individual walk in the personalized facial semantic field is conducted to perform disentangled semantic manipulation, with the objective attribute manipulated but the others preserved. Experiments

validate that the proposed method can surpass existing works in terms of identity and background preservation and disentangled editing. One future work is to edit other attributes beyond the annotations in dataset CelebA, such as editing facial poses via extending our idea to 3D engineering fields inspired by the related works [45–47]. Another potential future work is extending this method to the task of multimodal-driven face editing (such as voice-driven face editing [48,49])

**Acknowledgements** This work was in part supported by NSFC (Grant No. 62176194, Grant No. 62101393), the Major project of IoV (Grant No. 2020AAA001), Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2021KF0031), CSTC(Grant No. cstc2021jcyj-msxmX1148) and the Open Project of Wuhan University of Technology Chongqing Research Institute (ZL2021-6).

## References

1. Zhuang, P., Koyejo, O., Schwing, A.G.: Enjoy your editing: controllable GANs for image editing via latent space navigation. In: international conference on learning representations (2021)
2. Kemelmacher-Shlizerman, I., Suwajanakorn, S., Seitz, S.M.: Illumination-aware age progression. In: conference on computer vision and pattern recognition. p. 3334–3341 (2014)
3. Egger, B., Smith, W.A.P., Tewari, A., Wuhrer, S., Zollhöfer, M., Beeler, T., et al.: 3D Morphable Face Models - Past, Present, and Future. ACM Trans. Graph. **39**(5), 1–38 (2020)
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Waggenspack, W.N., (ed.), proceedings of annual conference on computer graphics and interactive techniques p. 187–194 (1999)
5. Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE conference on computer vision and pattern recognition p. 8789–8797 (2018)
6. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: AttGAN: Facial Attribute Editing by Only Changing What You Want. IEEE Transct. Img. Process. **28**(11), 5464–5478 (2019)
7. Karras, T., Aila, T., Laine, S., Lehtinen, J.: progressive growing of GANs for improved quality, stability, and variation. In: international conference on learning representations (2018)
8. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN training for high fidelity natural image synthesis. In: international conference on learning representations (2018)
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE conference on computer vision and pattern recognition. p. 4401–4410 (2019)
10. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: analyzing and improving the image quality of StyleGAN. In: IEEE conference on computer vision and pattern recognition. p. 8107–8116 (2020)
11. Shen, Y., Yang, C., Tang, X., Zhou, B.: InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. IEEE Trans. Pattern. Anal. Mach. Intell. **44**(4), 2004–2018 (2022)
12. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: GANSpace: Discovering interpretable GAN controls. In: annual conference on neural information processing systems (2020)
13. Wang, H., Yu, N., Fritz, M.: Hijack-GAN: unintended-use of pretrained, black-box GANs. In: IEEE conference on computer vision and pattern recognition, p. 7872–7881 (2021)
14. Li, M., Jin, Y., Zhu, H.: Surrogate gradient field for latent space manipulation. In: IEEE conference on computer vision and pattern recognition. p. 6529–6538 (2021)
15. Viazovetskyi, Y., Ivashkin, V., Kashin, E.: StyleGAN2 Distillation for feed-forward image manipulation. In: computer vision in european conference. vol. 12367, p. 170–186 (2020)
16. Yang, G., Fei, N., Ding, M., Liu, G., Lu, Z., Xiang, T.: L2M-GAN: Learning to manipulate latent space semantics for facial attribute editing. In: IEEE conference on computer vision and pattern recognition. p. 2951–2960 (2021)
17. Ju, Y., Zhang, J., Mao, X., Xu, J.: Adaptive semantic attribute decoupling for precise face image editing. Vis Comput. **37**(9–11), 2907–2918 (2021)
18. Han, Y., Yang, J., Fu, Y.: Disentangled face attribute editing via instance-aware latent space search. In: Proceedings of the thirtieth international joint Conference on artificial intelligence. p. 715–721 (2021)
19. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A latent transformer for disentangled face editing in images and videos. In: IEEE international conference on computer vision. p. 13789–13798 (2021)
20. Abdal, R., Qin, Y., Wonka, P.: Image2Style: How to embed images into the StyleGAN latent space? In: IEEE international conference on computer vision. p. 4431–4440 (2019)
21. Creswell, A., Bharath, A.A.: Inverting the generator of a generative adversarial network. IEEE Trans. Neural Netw. Learn Syst. **30**(7), 1967–1974 (2019)
22. Abdal, R., Qin, Y., Wonka, P.: Image2StyleGAN++: How to edit the embedded images? In: IEEE conference on computer vision and pattern recognition. p. 8293–8302 (2020)
23. Ma, F., Ayaz, U., Karaman, S.: Invertibility of convolutional generative networks from partial measurements. In: annual conference on neural information processing systems. p. 9651–9660 (2018)
24. Lipton, Z.C., Tripathi, S.: Precise recovery of latent vectors from generative adversarial networks. In: international conference on learning representations (2017)
25. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-Code GAN prior. In: IEEE conference on computer vision and pattern recognition. p. 3009–3018 (2020)
26. Zhu, J., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: European conference on computer vision. vol. 9909, 597–613 (2016)
27. Bau, D., Zhu, J.Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., et al.: Inverting layers of a large generator. In: ICLR workshop. vol. 2, p. 4 (2019)
28. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible Conditional GANs for image editing (2016). arXiv preprint arXiv:1611.06355
29. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H., Pérez, P., et al.: StyleRig: Rigging StyleGAN for 3D control over portrait images. In: IEEE conference on computer vision and pattern recognition. p. 6141–6150 (2020)
30. Xu, Y., Shen, Y., Zhu, J., Yang, C., Zhou, B.: Generative hierarchical features from synthesizing Images. In: IEEE conference on computer vision and pattern recognition. p. 4432–4442 (2021)
31. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., et al.: Encoding in style: a StyleGAN encoder for image-to-image translation. In: IEEE conference on computer vision and pattern recognition. p. 2287–2296 (2021)
32. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain GAN inversion for real image editing. In: European conference on computer vision. vol. 12362. p. 592–608 (2020)
33. Bau, D., Zhu, J., Wulff, J., Peebles, W.S., Zhou, B., Strobelt, H., et al.: seeing What a GAN cannot generate. In: IEEE international conference on computer vision. p. 4501–4510 (2019)

34. Guan, S., Tai, Y., Ni, B., Zhu, F., Huang, F., Yang, X.: Collaborative learning for Faster StyleGAN embedding. (2020). arXiv preprint arXiv:2007.01758
35. Yang, N., Zhou, M., Xia, B., Guo, X., Qi, L.: Inversion based on a detached dual-channel domain method for styleGAN2 embedding. IEEE Signal Process Lett. **28**, 553–557 (2021)
36. Lin, C., Xiong, S.: Controllable face editing for video reconstruction in human digital twins. Img. Vision Comput. **125**, 104517 (2022)
37. Lin, C., Xiong, S., Chen, Y.: Mutual information maximizing GAN inversion for real face with identity preservation. J. Visual Communicat. Image Represent. **87**, 103566 (2022)
38. Wang, S., Zou, Y., Min, W., Wu, J., Xiong, X.: Multi-view face generation via unpaired images. Vis Comput. **38**(7), 2539–2554 (2022)
39. Li, J., Ma, S., Zhang, J., Tao, D.: Privacy-preserving portrait matting. In: ACM multimedia conference, Virtual Event. p. 3501–3509 (2021)
40. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-Resolution. In: European conference on computer vision. vol. 9906; p. 694–711 (2016)
41. Wang, R., Chen, J., Yu, G., Sun, L., Yu, C., Gao, C., et al.: Attribute-specific Control Units in StyleGAN for Fine-grained image manipulation. In: ACM multimedia conference. p. 926–934 (2021)
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition. p. 770–778 (2016)
43. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning Face attributes in the Wild. In: IEEE international conference on computer vision. p. 3730–3738 (2015)
44. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: IEEE conference on computer vision and pattern recognition. p. 4690–4699 (2019)
45. Song, Y., He, F., Duan, Y., Liang, Y., Yan, X.: A kernel correlation-based approach to adaptively acquire local features for learning 3D point clouds. Comput. Aided Des. **146**, 103196 (2022)
46. Xu, H., He, F., Fan, L., Bai, J.: D3AdvM: a direct 3D adversarial sample attack inside mesh data. Comput. Aid. Geometric Design. **97**, 102122 (2022)
47. Liang, Y., He, F., Zeng, X., Luo, J.: An improved loop subdivision to coordinate the smoothness and the number of faces via multi-objective optimization. Integr. Comput. Aided Eng. **29**(1), 23–41 (2022)
48. Fang, Z., Liu, Z., Liu, T., Hung, C., Xiao, J., Feng, G.: Facial expression GAN for voice-driven face generation. Vis Comput. **38**(3), 1151–1164 (2022)
49. Huang, X., Wang, M., Gong, M.: Fine-grained talking face generation with video reinterpretation. Vis Comput. **37**(1), 95–105 (2021)




**Chengde Lin** received his MS degree in mechanical and electronic engineering from Guilin University of Electronic Technology, Guilin, in 2012. He is currently pursuing the PhD degree at the School of Computer Science and Technology, Wuhan University of Technology, China. His main research interests include computer vision and machine learning.

**Shengwu Xiong** received the BSc degree in computational mathematics and the MSc and PhD degrees in computer software and theory from Wuhan University, China, in 1987, 1997, and 2003, respectively. He is currently a professor with the School of Computer Science and Technology, Wuhan University of Technology, China. His research interests include intelligent computing, machine learning and pattern recognition.

**Xiongbo Lu** received the BS degree from the Hebei University of Technology, China, in 2014, and the MS degree from the Wuhan University of Technology, China, in 2019, where he is currently pursuing the PhD degree with the School of Computer Science and Technology. His main research interests include image generation and style transfer.