Information Sciences 660 (2024) 120130

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

ICGNet: An intensity-controllable generation network based on covering learning for face attribute synthesis

Xin Ning [a,c], Feng He [b,d], Xiaoli Dong [a,*], Weijun Li [a,c], Fayadh Alenezi [e], Prayag Tiwari [f,**]

[a] AnnLab, Institute of Semiconductors, Chinese Academy of Sciences, Beijing, 100083, China
[b] University of Science and Technology of China, Hefei, 230026, China
[c] Center of Materials Science and Optoelectronics Engineering School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100049, China
[d] Department of computer science, Yangtze University, Jingzhou, 434023, China
[e] Department of Electrical Engineering, Faculty of Engineering, Jouf University, Sakakah, 72388, Saudi Arabia
[f] School of Information Technology, Halmstad University, Sweden

ARTICLE INFO

Keywords:
Face attribute synthesis
Controllable intensity
Covering learning
Generative adversarial network
Image processing

ABSTRACT

Face-attribute synthesis is a typical application of neural network technology. However, most current methods suffer from the problem of uncontrollable attribute intensity. In this study, we proposed a novel intensity-controllable generation network (ICGNet) based on covering learning for face attribute synthesis. Specifically, it includes an encoder module based on the principle of homology continuity between homologous samples to map different facial images onto the face feature space, which constructs sufficient and effective representation vectors by extracting the input information from different condition spaces. It then models the relationships between attribute instances and representational vectors in space to ensure accurate synthesis of the target attribute and complete preservation of the irrelevant region. Finally, the progressive changes in the facial attributes by applying different intensity constraints to the representation vectors. ICGNet achieves intensity-controllable face editing compared to other methods by extracting sufficient and effective representation features, exploring and transferring attribute relationships, and maintaining identity information. The source code is available at https://github.com/kllaodong/-ICGNet.

- We designed a new encoder module to map face images of different condition spaces into face feature space to obtain sufficient and effective face feature representation.
- Based on feature extraction, we proposed a novel Intensity-Controllable Generation Network (ICGNet), which can realize face attribute synthesis with continuous intensity control while maintaining identity and semantic information.
- The quantitative and qualitative results showed that the performance of ICGNet is superior to current advanced models.

* Corresponding author at: AnnLab, Institute of Semiconductors, Chinese Academy of Sciences, Beijing, 100083, China.
** Corresponding author.
E-mail addresses: dongxiaoli@semi.ac.cn (X. Dong), prayag.tiwari@ieee.org (P. Tiwari).

https://doi.org/10.1016/j.ins.2024.120130
Received 20 March 2023; Received in revised form 21 November 2023; Accepted 10 January 2024

Available online 15 January 2024

1. Introduction

Face attribute synthesis is an important task in computer vision and has many applications. It aims to synthesize a facial image with target attributes and identity information using the available information. With the development of deep learning, especially after the Generative Adversarial Network (GAN) [1], face attribute synthesis has made significant progress in audio–visual entertainment, medical cosmetology [2] and virtual social interaction [3] and has attracted considerable attention from researchers. Most extant face attribute synthesis tasks are performed using GAN, which can be subdivided into two types: those based on style transfer and feature optimization according to the different implementation methods. The former is inadequate for mining attribute information and typically generates low-quality results, whereas the latter is complex and non-robust. These two methods are generally problematic because the attribute intensity cannot be continuously controlled. A process of attribute synthesis based on style transfer begins with constructing two attribute independent data distributions, and the model utilizes spatial pooling or down-sampling to learn. Meanwhile, domain information is manually or automatically added to the model to guide the accurate synthesis of target attributes [4]. However, down-sampling leads to ambiguity and quality degradation of generated results. Although adding a jump connection [5], selective transmission units [6] and local attention mechanisms [7] are advantageous, the image quality of the generated results still unsatisfactory.

Methods based on feature optimization decode an input image into a predefined region in a decoupled manner. Subsequently, directional synthesis of attributes can be realized by applying a linear operation to the latent code [8]. Because the final synthesis results depend on the performance of many aspects such as latent code alignment and attribute representation, it is difficult to achieve accurate attribute control. Instance optimization [8] and progressive synthesis can improve the quality of output results and realize continuous operation on attributes;however, they are very time consuming and often changes identity information.

In facial editing tasks, attribute intensity commonly describes the degree of modification applied to a specific attribute in the facial images and reflects attributes' expressiveness, prominence, and visibility. However, most current facial editing methods fail to achieve stable and controllable attribute editing, preventing the precise and fine-grained modifications of faces. The following challenges exist: (1) Attribute Confusion: There can be confusion or mutual influence among the attributes. For example, modifying the age attribute of a face may inadvertently affect other attributes such as gender or expression, resulting in inaccurate or unexpected editing results. (2) Uncontrollable Attribute Intensity: Some methods only offer discrete control over attribute intensity. This limited control leads to non-smooth variations in the edited results, failing to meet users' demands for precise control over attribute changes. (3) Attribute Distortion: Certain methods introduce image distortions or deformations during editing, producing images that appear unnatural or lack realism. However, these challenges hinder the realization of personalized attributes customization, artistic creation, and design. Therefore, achieving precise editing and personalization of facial attributes holds significant importance. Aiming at the above problem, taking face attributes as the research object, this paper attempts to use the homology continuity principle between homologous things to realize the continuous change of face attribute intensity.

Let's assume there is a set (S) representing things in nature. If two samples (a) and (b) belong to the set (S) and they are homologous but not identical, then there exists a gradual relationship (R) between (a) and (b), where there is a series of intermediate things (c_1, c_2, …) that also belong to set (S), satisfying the following conditions:
There is a relationship (R) between (a) and (c_1), (c_1) and (c_2), (c_2) and (c_3), and so on, up to (c_{n-1}) and (b). The gradual change from (a) to (b) is continuous, without any abrupt transitions or jumps. All intermediate things (c_1, c_2, …) share similarities with both (a) and (b) in terms of their attributes, and belong to the same class as them.
Symbolically:
Let (a), (b), (c_1), (c_2), … be elements of set (S), and let (R) denote the gradual relationship. If (a) and (b) are homologous but not identical, then there exist intermediate things (c_1, c_2, …) satisfying the following conditions:
1. $(R(a,c_1)),(R(c_1,c_2)),\ldots,(R(c_{n-1},b))$, where ($n$) is the number of intermediate things.
2. The change from (a) to (b) is continuous, without abrupt transitions or jumps in attributes.
3. All intermediate things $(c_1,c_2,\ldots)$ belong to the same class as (a) and (b) in terms of their attributes.

Currently, it has achieved good results when applied in facial expression transformation [9], face aging [10] and other applications, and achieved good results. The identity and background information must be consistent when performing face attribute synthesis. The same face with different attributes belongs to other states of the same object. Therefore, it conforms to the general principle of homologous continuity. We designed a new encoder based on this feature and proposed a novel face attribute synthesis method called intensity-controllable generator network (ICGNet), which obtains sufficient and effective face feature representation by mapping face images from different condition spaces into the face feature space. This ensures a connected path between any two face attributes of the same person based on the homology continuity principle and continuous change of attribute intensity. In addition, we analyzed the gradient relationships between the different attribute states and quantified them using a nearly linear function. The decoupling of attribute synthesis and intensity control is realized by transmitting the relationship back to the image space. Finally, we extended the process to progressive generation to make the generation details complete and realistic.

Our work contributes in three main aspects:
1) We designed a new encoder module that can effectively map facial images from different conditions to a facial feature space, allowing for a sufficient and meaningful representation of facial features. By applying the homology continuity principle, we demonstrated that features exhibit continuity in the face feature space, enabling continuous intensity control of individual attributes through vector direction adjustment.
2) Building upon our feature extraction approach, we introduced ICGNet, a novel, straightforward, and robust method for generating facial attributes. It enables the synthesis of facial attributes with continuous intensity control while preserving identity and semantic

information.
3) We conducted extensive experiments to evaluate and assess ICGNet comprehensively and compared its performance with multiple existing methods. Our evaluation included quantitative and qualitative analyses, demonstrating that ICGNet outperforms current advanced models regarding output quality and overall performance.

The rest of this article is organized as follows: Section 2 reviews related work, Section 3 introduces the structure and implementation details of ICGNet, Section 4 provides the experimental configuration of ICGNet and analyzes the results, and Section 5 summarizes the study and provides prospects for future work.

2. Related work

2.1. Face attribute synthesis methods

In recent years, the most popular method for face attribute synthesis is GAN [11], which made face attribute synthesis technology begin to be widely applied in many areas. Generally, GAN is equipped with a generator and discriminator. The generator mainly uses an encoder-decoder architecture. The encoder extracts the input image features and inputs them into the decoder to reconstruct the target image. According to the implementation process, face attribute synthesis can be mainly divided into two approaches, based on style transfer and feature optimization.

Style transfer. Most implementations based on style transfer need to construct two mutually independent distribution domains and independently mine the differences between the two distributions through the network to perform attribute transformation. Choi et al. [12] proposed StarGAN, which added attribute condition information based on CycleGAN [13] to realize the transformation of images with different attributes. In AttGAN, He et al. [5] modified the target attribute accurately by classifying the generated image attribute. To realize controllable face restoration, an SN-PatchGAN discriminator, a UNet-like generator, and gated convolution layer were used in SC-FEGAN [14] to realize accurate interactive face image restoration. To further improve the accuracy of attribute editing, STGAN [6] selected and modified the encoder features with self-adaption to decouple the generator by using a selective transmission unit. The above methods can only synthesize images with discrete attributes and cannot simulate continuous changes of attributes in actual scenes.

Feature optimization. Methods based on feature optimization reconstruct the target image by applying attribute information to the feature using a decoder. Karras et al. [15] proposed a novel styles-based generator architecture, StyleGAN, and achieved the best image quality. Collins et al. [16] proved that the W space used by StyleGAN is decoupled. To reduce optimization time, Zhu et al. [17] used a reverse generator. Bau et al. [18] learned the mapping relationship between image space and latent space by training another encoder. However, the image quality reconstructed by these methods largely depends on the pretrained generator, which can easily cause the loss of input information, leading to the failure of subsequent attribute synthesis tasks.

Methods based on style transfer or feature optimization can realize the synthesis of face target attributes. However, there exist some problems that cannot be ignored. The former can only synthesize results at low resolution. The latter realizes the synthesis of high-definition faces through progressive synthesis, but it easily results in insufficient attribute operation of attribute constraints depending on the pretrained generator. Besides, the mentioned methods cannot well realize the continuous intensity control of face attributes.

On this basis, we proposed ICGNet, which designs a new encoder module to extract sufficient and effective face features to ensure that there is connected path between any two face attributes of the same person based on homology continuity principle. Then, we can manipulate the features of different facial attributes to realize the gradually changing from one attribute to another. Finally, through the implementation of the decoder corresponding to the encoder, the continuous change of face attribute intensity is realized based on the gradient feature, and the face images with different intensity attributes are synthesized. In addition, the image quality is further improved by progressive synthesis operation. It should be noted that the proposed ICGNet is oriented to image generation applications, focusing on the continuous generation of different attributes of the face.

2.2. Face attribute datasets

Facial attributes cover many aspects, such as age, expression, skin color, and hair. We take the most common and attractive age and expression attributes in face synthesis as the research object and perform a detailed and comprehensive investigation on the relevant datasets.

Face expression datasets. A facial expression community contains a variety of expression data. For example, JAFFE [19] dataset was collected from 10 Japanese women in a standardized experimental environment, and facial expression images were then obtained through post processing. It contains 213 face images, and each identity has 7 different expressions. KDEF&AKDEF [20] database was initially developed for psychological and medical research, mainly for experiments of perception, attention, emotion, and memory. It contains 4,900 facial expression images of 35 women and 35 men. The GENKI [21] data, collected by the University of California's Machine Concept Laboratory, can be subdivided into three parts. These images have a wide range of backgrounds, such as different light conditions, geographical locations, personal identity, and ethnicity. RAFD [22] is a high-quality facial database collected by the Nijmegen Institute for Behavioral Sciences at Radboud University. It contains 8,040 images of 20 Caucasian adult males, 19 Caucasian adult females, 4 Caucasian boys, 6 Caucasian girls, and 18 Moroccan male adults. There are eight expression features:

Table 1
Facial expression datasets.

Dataset	Subjects	Total Images	Highlights
JAFFE [19]	10	213	Seven expressions: sad, happy, angry, disgust, surprise, fear, and neutral.
KDEF & AKDEF [20]	70	4,900	Seven expressions, each with five angles.
GENKI [21]	—	15,159	Two types of smiles and no smile.
RaFD [22]	67	8,040	Eight expressions and three gaze directions.
CK [25,26]	137	* video sequence	Contains labels for expressions and basic action units
Fer2013 [27]	—	26,190	Six expressions.
RAF [28]	—	29,672	Five precise key points of the face, age, and gender.
EmotionNet [29]	—	950,000	Contains basic and compound expressions, and labels for expression units.
AffectNet [30]	—	420,299	Eight facial expressions along with the intensity of valence and arousal.
DISFA [31]	—	130,788	Consisting of 27 videos of 4844 frames each, with 130,788 images in total.
FERV39k [32]	—	38,935	Trusted manual labeling process.

Table 2
Age datasets.

Dataset	Subjects	Total Images	Highlights
FGNet [33]	82	1,002	The age range is 0 to 69.
CACD2000 [34]	2,000	163,446	The range is 16 to 62.
Adience [23]	2,284	26,580	Age label: (0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, 60+).
IMDB-WIKI [35]	—	523,051	The image comes from the Internet.
MORPH [24]	13,000+	55,000	The age range is 16 to 77.
FairFace [36]	—	108,501	7 race groups.

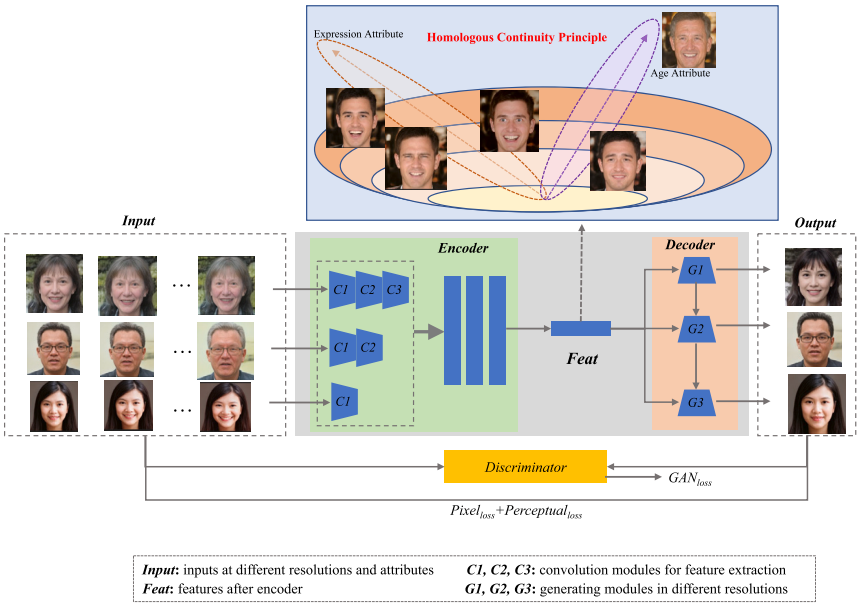

Fig. 1. Schematic of ICGNet network architecture. We used images with different attributes and resolutions as input and three progressive convolution modules for feature extraction and fusion to combine different levels of features to get Feat. Feat can be edited in feature space to realize image generation with varying intensity attributes. The decoder uses the progressive generation process to generate results at each resolution and calculate the corresponding loss through the discriminator and attribute classifier. Finally, the reverse boot generator restores the image at each level and calculates the difference between the two.

anger, disgust, fear, happiness, sadness, surprise, contempt and neutral, and each expression contains three different head postures. The details of datasets are presented in Table 1.

Face age dataset. The Adience [23] dataset was collected by uniform side photography equipment, including 26,580 pieces of nonrepeated data of 2,284 people. Unlike previous datasets, this dataset was marked by age group rather than specific age. IMDB-WIKI [1] is a large human face database with data from various sources. It contains 523,051 pieces of human face data. The age is calculated according to the date of photo taking and the date of birth of the person for labeling. It is the largest age-related dataset at present and of great research value. Morph [24] dataset was published in 2017, which labeled a total of 55,000 pieces of images from 13,000 people using continuous labeling. Detailed information datasets can be found in Table 2.

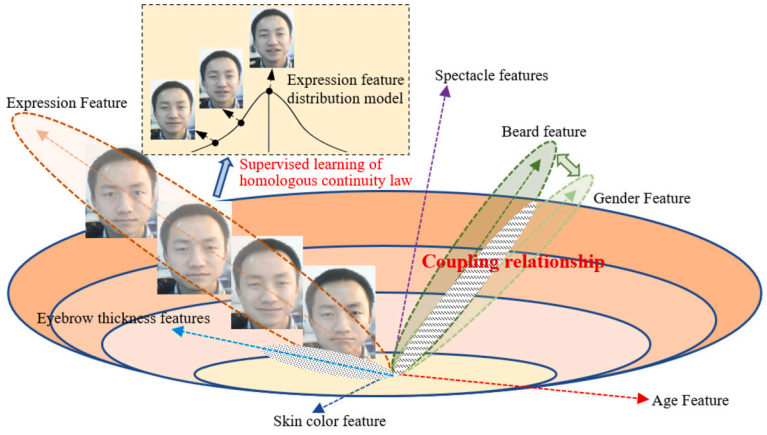

Fig. 2. Example figure of facial attribute manifold distribution. The encoder receives facial attribute images as input and transforms them into feature representations through a series of neural network layers. In this process, the encoder utilizes the principle of homology continuity to map facial images with similar attributes to adjacent regions in the feature space, capturing the continuity between attributes. In the feature space, the encoder decouples different attribute features. Firstly, it extracts expression features that reflect the facial expression states such as smile, anger, or surprise. By independently representing expression features, the encoder can precisely control and edit the facial expression attributes. It also extracts skin color features that reflect the tone and brightness of the skin in facial images. By decoupling the skin color features, it can independently adjust and modify the skin color attribute of the facial image to adapt to different needs and scenarios. Additionally, the encoder extracts age features that reflect the age attribute of the face. By decoupling the age features, the encoder can accurately control and edit the age of the face, enabling personalized handling of facial attributes. It also extracts gender features that represent the gender attribute in facial images. By decoupling the gender features, the encoder can independently recognize and edit the gender attribute of the face, resulting in higher accuracy and controllability of the gender attribute.

3. ICGNET method

3.1. Overview

The overall structure of the neural network ICGNet we proposed is shown in Fig. 1. Based on the traditional GAN, we used a combination of encoders and decoders for the generator module, as Fig. 1 shows. The encoder-decoder architecture is conducive to decoupling different attributes, and a single attribute can be synthesized through an orthogonal operation. Improving these crucial points is essential for enhancing the attribute composition and quality. In addition, continuously changing attributes can be established by controlling the strength of attribute synthesis. To obtain a complete representation of the input information (Feat in Fig. 1), we performed a feature extraction operation that gradually decouples the attribute features of the input images with different attributes for the same person inside the encoder. Based on the homologous continuity of facial attributes, face attribute features in the feature space can realize different strength controls of attributes, which is also the basis for realizing the continuous editing of attribute strength. Based on continuous attribute strength, the decoder is designed accordingly, and context information can gradually realize high-quality facial image generation. Reconstructed images do not contain any attribute operations. We conducted in-depth mining of prior information to realize the target attribute synthesis. Similarly, we obtained a unique manifold distribution that can control the target attribute synthesis, as Fig. 2 shows.

We analyzed the relationship between attributes and face attribute changes in the face feature space. Combined with the homologous continuity principle of attribute change and based on different face attribute images and feature extraction modules, we gradually mapped the face attribute features to a high-dimensional feature space, where the distribution model of attribute features can be constructed. Based on this model, face attribute features can be edited quantitatively to realize the transformation of specific attributes and the control of attributes to different degrees.

The configuration and implementation details of each of the above components are described in the following sections.

3.2. Covering learning

Coverage learning is the existence of optimal coverage in a high-dimensional space. After the encoder processed the input image, it was sent to a high-dimensional space. However, this is difficult to visualize in a high-dimensional space, and data properties are challenging to mine. In addition, the distribution of different categories is more complex, and the category space must be effectively fitted using a single neuron. Therefore, it is necessary to combine existing simple geometries to form more complex geometries as basic units to cover relevant face attributes. Theorem 1 shows that coverage learning determines the best coverage for different facial attributes.

Covering learning is a proposed bionic network training method based on homology continuity. Most traditional pattern-recognition methods perform the best segmentation in the feature space where the classified samples are located, which is linear and prone to error. This can confuse face attributes, ages, and expressions, leading to a significant difference between the generated and authentic face images. However, this coverage learning method determines the best coverage of different sample classes after determining the distribution of samples by preprocessing (coding) the sample set and mapping it to the corresponding high-dimensional

space. Prior knowledge makes the entire training process faster and requires fewer parameters while achieving a higher correct rate and a more coordinated generation effect. The method of face attribute editing can be summarized as follows. First, the input face attribute images are mapped into the high-dimensional feature space, to find the stream shape distribution of the corresponding face attribute dataset in the high-dimensional feature space. Overlay learning is subsequently performed in the streaming subspace. Finally, by manipulating the expression and age semantic vectors, face attribute-edited images with consistent background and identity information before and after are generated by the generator.

Theorem 1 *(Covering learning). In the feature space R^n, among all samples of the same kind, there is the best coverage.*

Proof. In the feature space R^n, there are similar sample point sets $A = x_1, x_2, \ldots, x_n$
(1) When $n = 1$, the conclusion holds, and there is a unique optimal coverage.
(2) When $n \geq 2$, assume that all sample sets of the same type as A are M, according to the continuity of homology:
$\forall x_i, x_j \in A$ (where $ineqj$), $\varepsilon > 0$.
$\exists I_{ij} = a_1, a_2, \ldots, a_n < M$, where
$m \to \infty$;
$x_i = a_1, x_j = a_m$;
$\overset{p}{(} a_n, a_{n+1}) < \varepsilon, 1 \leq n \leq m-1$
For the closure of I_{ij}, it can be seen that $\vec{I_{ij}}$ is connected.
Take the union of all $\vec{I_{ij}}$, record $A' = \bigcup_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n \\ i \neq j}} \bar{I}_{ij}$, we know that A' is connected.
So it can be seen that $\exists k > 0$, making
$\left\{ x \mid \rho(x, y) \leq k, y \in A', x \in R^n \right\}$. is the best point set A overlay.
Where $\left\{ x \mid \rho(x, y) = k, y \in A', x \in R^n \right\}$ is the covering hypersurface of point set A.

The homology continuity encoder updates its parameters through covering learning. Covering learning indirectly enables stable attribute control by determining the parameters of the encoder. We have designed a novel encoder based on homology continuity, which maps different face images into the facial feature space. This encoder extracts information from input face images, generates effective vector representations, and models attribute relationships (modeling the relationship between instances and representation vectors in the facial feature space) to ensure accurate attribute synthesis and preservation of irrelevant regions. Furthermore, we modify the intensity of attributes by adjusting certain dimensions of the representation vectors and adding specific constraints. The identity consistency of attribute-generated face images is ensured through an identity loss function and a reconstruction loss function. Homology continuity means that for any mode, different states of existence are strongly correlated and exhibit a clear continuum through which the switching of different states is achieved. Homology continuity is a fundamental property that is reflected in many aspects and has experienced rapid development in recent years, finding applications in various fields. For example, a newborn lamb, after several years of development, becomes an adult sheep. The adult sheep has significant differences in appearance compared to the previous lamb, and the entire process is continuous and gradual as the lamb ages. Similarly, for humans, facial expressions are variable, and the change in facial expressions is achieved through the movement of facial muscles. Traditional pattern recognition methods are based on achieving classification by best segmenting the classified samples in the feature space. In contrast, homology continuum-based bionic pattern recognition achieves classification by mapping the preprocessed sample points into a high-dimensional feature space and achieving the classification effect through the best coverage of the sample distribution. Compared to traditional methods, covering learning can obtain more accurate results with fewer involved parameters, making the network more lightweight for further model deployment and practical applications. Coverage learning optimizes the encoder based on homology continuity and improves the quality of generated results and the stability of attributes. The encoder maps the input facial images to their corresponding feature space, obtaining the corresponding feature representations. These feature vectors are then used in the subsequent coverage learning optimization process. Coverage learning helps the encoder based on homology continuity find the optimal coverage. Through iterative refinement of the encoder, at each iteration, the feature vectors are fine-tuned to better match the requirements of the target attribute.

3.3. Editable feature extraction encoder based on homologous continuity principle

In most attribute synthesis networks, the generator uses the encoder-decoder architecture to calculate the pixel difference between the input image and the reconstructed image through cyclic generation, keeping the original identity and background information in the synthesized image. Although this method is effective and easy to operate, the down-sampling in the encoder will inevitably lead to the loss of image information, which can only guarantee consistency in a large range. Therefore, there are apparent differences in the face identity, some details, and background. In ICGNet, inspired by FPN [37], we embedded the range multiple attributes, and attribute intensity information based on the original attribute generation algorithm and fused the features extracted in different spaces to obtain the complete representation of the input image in the feature space.

Our method aims to generate images of face attributes with different intensities. When designing the encoder, the homologous continuity of face attributes is the design basis. Therefore, after the feature extraction operation of the encoder, images with different intensities of the same attribute of the same person can output Feat with consistent identity and editable attributes. Feat is an editable

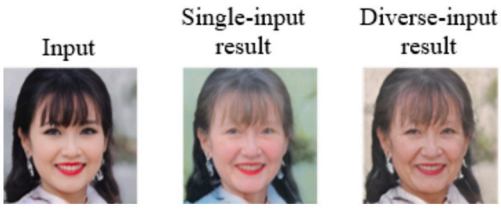

Fig. 3. Comparison of the effects of input images with different factors on the results (for example, age).

feature, which can be edited to generate different Feats, which can be input to the decoder to generate face images with different attribute intensity. Therefore, we believe that such Feat is effective and complete, and it is also the basis for realizing the continuous change of attribute strength.

We empirically configure the image resolution and different attribute strengths, and take them as inputs. When they are input into the designed encoder, the encoder will output different Feats. For these Feats, we use a certain feature weight to merge them to generate high-quality synthetic face images. Fig. 3 compares the effects of attribute intensity and image resolution settings on the synthetic image. The first column represents the original image, the second column represents the image synthesized by single resolution input without intensity change of the same attribute (such as age), and the last column represents the image synthesized by input with both resolution and attribute intensity change. From the comparison result in Fig. 3, we can see that the more diversified factor of the input image, the smaller the difference in information.

The process of diversified-input features extraction can be formulated as follows:

$$L_{Feat} = \sum_{i=1,2,3} A_i F(X_i) + \alpha * I \tag{1}$$

where X_i represents the input image of i*th* layer $i = 1, 2, 3$. F represents the feature extraction operation on different size, A_i represents the feature normalization operation to limit the feature in an appropriate range, the product of identity matrix I and minimum constant α is the reminder to avoid the singular matrix. Regarding the implementation details of the encoder, specifically, we employ VGGNet-19 as the network for feature extraction. It possesses powerful feature extraction capabilities, effectively capturing the relevant information from face images. Additionally, the network consists of multiple convolutional and fully connected layers, enabling the learning of high-dimensional feature representations that encompass complex relationships between facial attributes. Since VGGNet-19 has been pretrained on the ImageNet dataset, it has acquired a general facial feature representation, enhancing the performance of face attribute tasks and exhibiting strong transferability. To further enhance the processing capability of face attribute images, we have also trained the network on mainstream face datasets. During the feature extraction process, the input face features are mapped to feature representations using the attribute encoder based on homology continuity. Facial image features are extracted through a combination of pooling and convolutional layers. In the feature space, the extracted facial features are mapped to a specific region, ensuring continuity of facial attributes by incorporating fully connected layers. In modeling homology continuity, k-means clustering is employed to establish the homology continuity relationships among similar attribute faces. During attribute synthesis, based on the learned homology continuity representations, linear interpolation and adjustment of the facial feature vector direction are performed in the feature space to achieve attribute synthesis with varying intensities.

The encoder based on homology continuity achieves attribute control primarily through: Homology continuity modeling of attributes: The encoder based on homology continuity captures the variations and correlations between attributes by modeling the continuous relationships among samples in the feature space. It captures the patterns of attribute changes and their interdependencies. Learning feature representations: The encoder based on homology continuity learns feature representations of facial images, mapping them into the feature space. In the feature space, similar attribute faces have similar representations. By extracting meaningful and rich feature representations, the encoder accurately expresses the attribute information of the face, enabling stable control over attributes. Synthesis and interpolation: The encoder based on homology continuity utilizes the attribute relationships in the feature space to achieve attribute synthesis and interpolation. By performing interpolation or adjusting the facial attribute vectors in the feature space, facial images with different attribute intensities can be synthesized. This allows for precise control over attributes and ensures stability in attribute control. Facial attribute consistency: During attribute editing, the encoder based on homology continuity ensures that the edited facial images maintain similarity to the original images, particularly in preserving identity information. This is achieved by incorporating an identity loss function. This approach maintains the consistency and stability of facial images while editing attributes. The advantages over other feature extraction methods are mainly manifested in:

Capture of attribute correlations: The encoder based on homology continuity captures the correlations and patterns of attribute variations. By modeling the continuity relationships among attributes in the feature space, the encoder can more accurately represent and control the attributes. This leads to more accurate and stable relationships between attributes, enhancing the precision and reliability of facial attribute extraction.

Stable attribute control: The encoder based on homology continuity achieves stable control over attributes. By leveraging and utilizing homology continuity, the encoder ensures that facial images with similar attributes have similar representations in the feature space, enabling precise attribute synthesis and stable control. This contributes to more accurate and stable results in facial attribute editing and synthesis.

Powerful feature extraction capability: The encoder based on homology continuity utilizes VGGNet-19 as the feature extraction

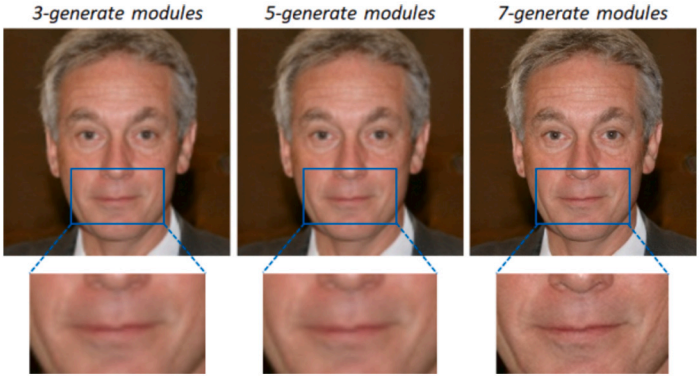

Fig. 4. Comparison of synthetic faces with different number of generation modules. Note that we used seven generation modules in ICGNet, achieving the best sharpness.

model. These models can extract rich and meaningful feature representations from facial images. Compared to traditional feature extraction methods, the encoder based on deep learning can better capture the details and variations of facial attributes, thereby improving the accuracy and robustness of attribute extraction.

Benefits of transfer learning: The encoder based on homology continuity often utilizes pretrained models for initialization and fine-tunes them on task-specific data. This leverages the general feature representations learned by the pretrained models on large-scale datasets. Through transfer learning, the encoder applies these general feature representations to the task of facial attribute control, enhancing the stability of attribute control. This advantage of transfer learning enables the encoder to perform well in tasks with small samples and specific domains.

3.4. Decoder based on homologous continuity principle with progressive generation

Feat output by the encoder is editable. The distribution model constructed in the feature space enables us to edit Feat effectively, so that we can generate face images with different attributes and different intensities. In this part, we will introduce the design and implementation of decoder based on different Feats to generate high-quality face images.

In general, the face attribute synthesis network can only synthesize low-resolution faces (64*64 or 128*128 pixels), which is mainly determined by the configuration of the generator. We increased the resolution of the synthetic face to 256*256, and theoretically, the output resolution could be increased to the target size by adding generation modules (mostly learned from StyleGAN). Specifically, unlike the traditional methods [5,12,13] which directly synthesize the target size, we adopted a method that progressively increases the size of the image. As shown in Fig. 1, an encoder can be used to obtain feat, which is the feature representation of the input data. It synthesizes the result of specific size by passing the data on the specified dimension into the corresponding generation module and superposes the previous output after up-sampling with the current output. Notably, we calculated the pixel loss and perceptual loss at each resolution to provide gradients for updating network parameters. As shown in Fig. 4, the higher the number of generated modules, the higher the facial resolution and the better the visual quality. The process of layer-by-layer image synthesis can be formulated as follows:

$$S_i = G_i(Feat_i) + \lambda_i * Up(G_{i-1}(Feat_{i-1})) \tag{2}$$

where S_i represents the output of *ith* layer $i = 1, 2, 3$, $Feat_i$ represents the eigenvector passed to the corresponding generation module, $Up(\bullet)$ represents the up-sampling operation, λ_i is a coefficient to balance the influence of output from last layer on current output (in the experiment, $\lambda_i = 0.5$). Through the construction of encoders and decoders, we realize the construction of skeleton network, and the update of network parameters can be expressed as:

$$\Delta W = lG\Delta W^{BP}, \tag{3}$$

where l represents the learning rate, ΔW^{BP} is calculated based on standard gradient back propagation.

The feature space is determined by the outputs of the encoder, where each feature vector corresponds to a feature representation. The feature space captures the relationships between different attributes and features in the facial images and forms a certain distribution within the feature space. Specifically, the feature distribution is influenced by the specific encoder used and the training data. Different attributes may form specific distributions, and the feature vectors of different attributes in the feature space may exhibit different clusters or distributions. For example, the feature vectors of gender attribute may form two distinct clusters in the feature space, corresponding to males and females. Other attributes such as age, expression, and skin color may also exhibit different distribution patterns in the feature space. The feature distribution in the feature space is learned by the encoder during the training process and is influenced by the training data. Therefore, the specific feature distribution in the feature space may vary depending on factors such as the architecture of the encoder and the diversity and scale of the training data. Within the feature distribution of each attribute, the feature space may exhibit certain patterns of variation. For example, for the expression attribute, the feature

vectors of similar expressions may be closer to each other in the feature space, while there may be some separation between the feature vectors of different expressions. The feature space is determined by the outputs of the encoder, where each feature vector corresponds to a feature representation. The feature space captures the relationships between different attributes and features in the facial images and forms a certain distribution within the feature space. Specifically, the feature distribution is influenced by the specific encoder used and the training data. Different attributes may form specific distributions, and the feature vectors of different attributes in the feature space may exhibit different clusters or distributions. For example, the feature vectors of gender attribute may form two distinct clusters in the feature space, corresponding to males and females. Other attributes such as age, expression, and skin color may also exhibit different distribution patterns in the feature space. The feature distribution in the feature space is learned by the encoder during the training process and is influenced by the training data. Therefore, the specific feature distribution in the feature space may vary depending on factors such as the architecture of the encoder and the diversity and scale of the training data. Within the feature distribution of each attribute, the feature space may exhibit certain patterns of variation. For example, for the expression attribute, the feature vectors of similar expressions may be closer to each other in the feature space, while there may be some separation between the feature vectors of different expressions.

3.5. Intensity controllable feature editing

To realize the directional synthesis of target attributes, we proposed to mine attribute of precedent information in feature expression. If the feature obtained from the encoder is sufficient and the change of attributes in the image space is continuous, the distribution of attributes in the feature space should be a distribution with a continuous manifold homologous continuity [38]. That is, we can use an n-dimensional space to fit the spatial distribution of any attribute (n represents the dimension of the input feature), as stated in Theorem 1.

Theorem 2. *A single attribute is continuous in the feature space.*

Assumption. We assume that the image acquisition process is affected by n conditions (such as illumination, angle, and distance). The collection of images can be expressed as $I(x_1, x_2, \ldots, x_n)$, where each dimension represents a collection condition, and the range of each dimension is $[v_i^s, v_i^l]$, $i \in [1, n]$. Using the Cartesian product for optimization, the previous process can be expressed as $D = (v_1^s, v_1^l)(v_2^s, v_2^l) \cdots (v_n^s, v_n^l)$, where D is the domain. In addition, the sample with attribute l in D is marked as D^l. For any two samples, D_1^l and D_2^l, there is $||l^1 - l^2|| \to 0$, when $|x_i^1 - x_i^2| \to 0$. In other words, when certain acquisition conditions of two images are consistent, their representations in specific dimensions are also consistent.

Proof.

$I^1(x_1^1, x_2^1, \ldots, x_3^1)$, $I^2(x_2^2, x_2^2, \ldots, x_3^2) \in D$,

$$
\begin{aligned}
T &= ||I^1(*) - I^2(*)|| \\
&\leqslant ||I^1(x_1^1|*) - I^2(x_1^2|*)|| + ||I^1(x_2^1|*) - I^2(x_2^2|*)|| \\
&\quad + \ldots + ||I^1(x_n^1|*) - I^2(x_n^2|*)||
\end{aligned}
$$

According to the basic assumption: $\forall x_i^1, x_i^2 \in [v_i^s, v_i^l], i = 1, 2, 3, \ldots, n$, $\exists \delta_i > 0$, makes $||I^1(*) - I^2(*)|| < \varepsilon/n$, when $|x_i^1 - x_i^2| < \delta_i$, $\forall \varepsilon > 0$.

Therefore, $||I^1(*) - I^2(*)|| < n * \varepsilon/n = \sigma$, when $|I^1(x_1^1, x_2^1, \ldots, x_n^1) - I^2(x_1^2, x_2^2, \ldots, x_n^2)| < \sigma$.

Hence, attribute is continuous in the feature space. To quantitatively describe this relationship, we used Support Vector Machine (SVM) to construct a hyperplane that completely distinguishes between two opposing classes of attributes in the following way,

$$n^T z = 0, \text{ for } \forall z \in R^d. \tag{4}$$

We can achieve continuous intensity control of a single attribute by moving in the direction of vector. Further, we will superimpose with different intensities, as shown in Equation (5).

$$S_t = G_t(C_t(X) + \lambda * n) \tag{5}$$

where S_t represents the final output result, X represents the original input, λ represents the current composition strength, n represents the target attribute axis, it is worth noting that C_t and G_t here does not represent the operation of any layer, but the whole encoder and decoder integrated result.

3.6. Loss function

To make the network output and input images as consistent as possible, we also calculated the differences in pixels and perceptual losses between the two based on the GAN and used the weighted results as the final loss function to update the parameters. The architecture of the GAN simplifies the configuration of the whole network, perceptual loss ensures the semantic similarity between the two, and pixel loss enhances the image details. The overall loss function can be expressed as follows:

$$L_{total} = \lambda_1 * L_{pix} + \lambda_2 * L_{percep} + \lambda_3 * L_{GAN}, \tag{6}$$

where λ_1, λ_2, and λ_3 are used to balance the influence of different loss functions. Pixel loss L_{pixel} is defined as:

$$L_{pixel} = \frac{1}{w \times H} \sum_{x=1}^{w} \sum_{y=1}^{h} \left| I_{x,y}^{output} - I_{x,y}^{input} \right|, \tag{7}$$

where w and h represent the width and height of the image, I^{output} represents the output image, and I^{input} represents the input image.

The perceptual loss can be expressed as follows:

$$L_{percep} = \sum_{i=1}^{4} \frac{1}{C_i H_i W_i} * ||E_i(x_{in}) - E_i(x_{out})||, \tag{8}$$

where x_{in} and x_{out} denote input and output images, $E_i(\cdot)$ denote output of ith layer of the convolutional network, C_i, H_i, W_i represents output size of current layer. We calculate the L2 distance of input and output on VGG16 network Conv1_1, Conv1_2, Conv3_2 and Conv4_2 respectively.

The Perceptual Loss is a commonly used loss function in image synthesis tasks. It measures the perceptual difference between the generated image G and the ground truth image T by comparing their feature representations extracted from a pre-trained deep neural network, such as VGGNet. The perceptual loss The Generative Adversarial Network (GAN) Loss is a two-player adversarial training scheme used in image synthesis. It consists of a generator network G and a discriminator network D. The GAN loss is defined as a minimax game, where the generator aims to produce realistic images that can deceive the discriminator, while the discriminator aims to distinguish between real and generated images. The GAN loss encourages the generator to produce high-quality and realistic images.

4. Experiments

4.1. Datasets

We selected multiple datasets from the face attribute community for the experiment, including RaFD [22], EmotionNet [29], CACD2000 [34], and Adience [23]. These covered 6 attributes: age, happy, sad, angry, surprised, and fear, each of which randomly selected approximately 5,000+ images and manually eliminated poor quality images. In the experiment, all the data were detected and aligned with faces and uniformly scaled to 256*256.

4.2. Implementation details

Our method was implemented using the TensorFlow machine learning library and executed on a computer using the NVIDIA Quadro RTX 5000 GPU (16G). The resolution of the output image is 256*256, with 16 as the basic batch size. For every half reduction in resolution, the batch size doubles. Adam [39] optimizer was used to optimize the model, and the learning rate was set as 0.0001. The loss function included $Cycle_{loss}$, Att_{loss}, GAN_{loss}.

Selecting 5700 facial attribute images from the four mainstream datasets, ensuring dataset diversity and representativeness by including variations in age, gender, skin color, pose, and lighting conditions. In this step, the dataset is cleaned and labeled, removing low-quality or blurry images and non-facial images, and ensuring that each facial image has corresponding labels. Then, the facial image dataset is scaled to 256*256, normalized, and cropped to focus on the main facial regions. This ensures that all facial images have the same size and format. To further increase the diversity of the dataset, data augmentation techniques such as random rotation, translation, scaling, and noise addition are applied.

4.3. Experiments of ICGNet

4.3.1. Attribute synthesis

In Section 3, we propose to use multilevel feature extraction to solve the problem of a large difference in image information and use progressive generation to improve the quality of the generated image. In this section we will evaluate the influence of ICGNet on attribute synthesis task from both quantitative and qualitative perspectives. And the user study is used to reflect the real effect of the edited image with the edited face attributes. Our model has better image quality, lower inference time, and smaller number of parameters compared to the current SOTA approach.

Qualitative evaluation. To evaluate the performance of ICGNet in different attribute synthesis tasks, we designed experiments to compare several methods based on style transfer and feature interpolation. Fig. 7 shows the attribute synthesis results of all methods. According to the results, all methods obtain synthesis results of target attributes. Among them, CycleGAN, StarGAN, AttGAN have low-quality results and poor image details, especially for the mouth area, which is prone to image artifacts. Although PSP and LIA have good-quality results, the background and identity and other nontarget area information also vary considerably, resulting in a large difference between results and inputs. For our method, most of the information of the original image is retained, whereas the target attribute is modified, and the quality of the generated image is further improved.

Table 3
Comparison of quantitative results of different methods.

Methods	Mask-SSIM [40]	Mask-PSNR [41]	FID	Inference Time(s)	Parameters(M)
CycleGAN [13]	0.63	27.91	65.77	10.54	**150.77**
StarGAN [12]	0.57	26.84	27.71	5.67	210.56
AttGAN [5]	0.59	27.33	35.99	0.37	345.88
DualG-GAN [42]	0.55	28.43	30.87	0.95	480.92
PSP [43]	0.54	29.91	25.97	0.28	457.69
LIA [17]	0.61	30.05	27.52	0.13	435.21
Feature Alignment [44]	0.62	31.44	21.91	0.16	600.55
Subspace Diffusion [45]	0.63	32.23	19.56	0.34	823.80
Xin et al. [46]	0.59	28.72	24.36	0.59	322.91
Hou et al. [47]	0.57	30.94	18.36	0.17	210.34
Xu et al. [48]	0.51	26.38	18.96	0.28	134.97
ICGNet	**0.65**	**33.27**	**15.77**	**0.08**	190.67

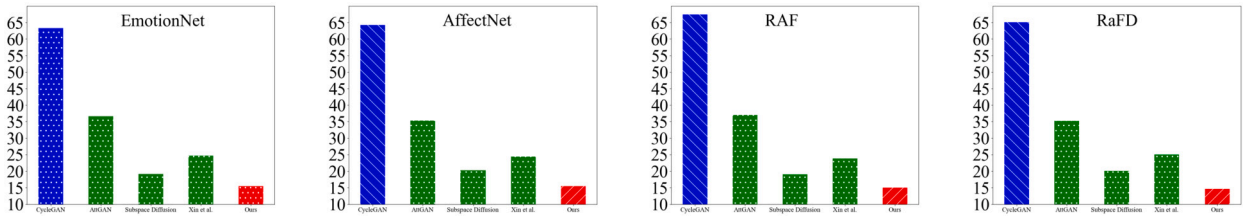

Fig. 5. FID in different expression datasets.

Quantitative evaluation. In addition to evaluating ICGNet from the perspective of comprehensive visual performance, we aim to make a quantitative evaluation of the visual quality of synthetic attributes in ICGNet and compare it with previous methods. Ideally, we only need to calculate the image quality of the face region before and after attribute synthesis. Unfortunately, no algorithm can automatically filter the face from the non-face region. In our work, we used Mask-SSIM [40] as a quantitative evaluation index for the visual quality of attribute synthesis results by referring to the work done on face replacement and face repair. In addition, we also expanded on this and designed Mask-PSNR [41] for quantitative evaluation. These two quantitative indicators, respectively, calculate the SSIM [49] and PSNR [50] scores of the synthetic image and the original image in the face region, without considering the background information. Because the calculation of Mask-SSIM and Mask-PSNR needs to accurately separate the face and background region of each image, to accelerate the calculation process, we used lightweight pre-trained BisenetV2 [8]. The Mask-SSIM score range is $[0, 1]$, and the higher the value, the better the image quality. The value range of Mask-PSNR is generally referred to [35]. Similarly, the larger the value, the better the quality of the image. Table. 3 shows the average results of all methods; among them, ICGNet has the highest score, which also confirms the observation in Fig. 7. As shown in Figs. 5 and 6, we compare the quality of the synthesized images on different data with the current SOTA method.

The Structural Similarity Index (SSIM) is a widely used metric to assess the similarity between two images. It measures the structural information, luminance, and contrast similarities between the reference image R and the generated image G. The SSIM score ranges from 0 to 1, with 1 indicating a perfect match:

$$\mathrm{SSIM}(R,G) = \frac{2\mu_R\mu_G + C_1}{\mu_R^2 + \mu_G^2 + C_1} \cdot \frac{2\sigma_{RG} + C_2}{\sigma_R^2 + \sigma_G^2 + C_2} \tag{9}$$

The Peak Signal-to-Noise Ratio (PSNR) is a commonly used metric to evaluate the image quality. It measures the ratio between the maximum possible pixel value and the mean squared error (MSE) between the reference image R and the generated image G. Higher PSNR values indicate better image quality:

$$\mathrm{PSNR}(R,G) = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}(R,G)}\right) \tag{10}$$

The Fréchet Inception Distance (FID) is a metric commonly used to evaluate the quality and diversity of generated images. It measures the distance between the feature representations of real images and generated images, as extracted by an Inception network. Lower FID values indicate better image quality and diversity:

$$\mathrm{FID}(R,G) = \|\mu_R - \mu_G\|^2 + \mathrm{Tr}(C_R + C_G - 2(C_R \cdot C_G)^{1/2}) \tag{11}$$

User study. For further evaluation of our proposed method, we invited 100 experts in the field of image generation to score the images generated by different models. The evaluation is based on two main criteria: similarity and image quality. Similarity reflects how well the model extracts the original background and identity information. On the other hand, image quality reflects how sharp the model generates images, which is a critical evaluation criterion. Table 4 shows the final User Study results. According to the results, it can be seen that our proposed model has a clear advantage in terms of similarity and image quality. This is because the

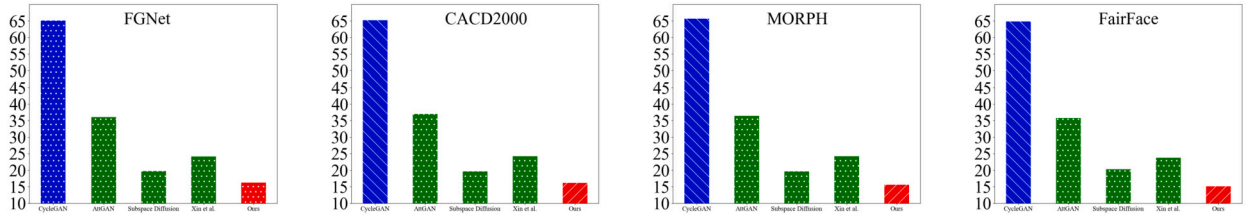

Fig. 6. FID in different age datasets.

Table 4
Results of user study. One hundred experts in image generation rated similarity and image quality.

Method	Similarity	Quality
CycleGAN [13]	80.12	78.55
StarGAN [12]	81.34	75.22
AttGAN [5]	82.34	80.55
LIA [17]	83.45	81.62
Feature Alignment [44]	82.15	85.22
Subspace Diffusion [45]	83.77	83.59
Xin et al. [46]	80.54	79.23
Ours	**85.44**	**87.39**

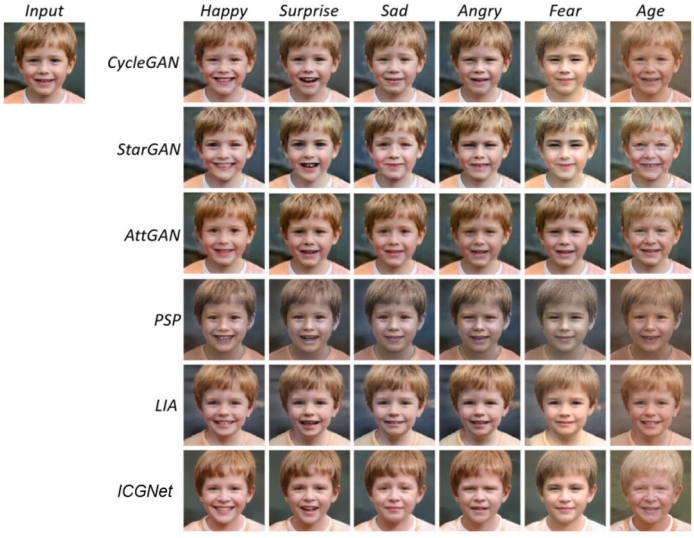

Fig. 7. Comparison of results from different methods. From left to right, each column represents "happy," "surprise," "sad," "angry," "fear," and "age." The experimental results of CycleGAN, StarGAN, AttGAN, PSP, LIA, ICGNet are shown in order from the first to the last line.

cover learning-based generation method can better retain the identity and background information of the Source Image and have richer texture details by directly manipulating the feature vector with reduced loss.

4.3.2. Intensity control

As described in Section 3, the ICGNet generator adopts the encoder–decoder architecture, which means that we can control the intensity of the target attribute by weighting the encoded features. To achieve this goal, SVM regression was performed on the labeled attribute features to obtain the change directions of different attributes, as shown in Fig. 10. We conducted an attribute intensity control experiment on age, happiness, sadness, anger, surprise, and fear. In the experiment, the intensity changes of attributes were realized by controlling the weight intensity of the attribute axis; the results are shown in Fig. 8 and Fig. 9. As the results show, we realized the intensity control of different attributes on a random identity, which is more in line with the actual biological law of development, especially for the control of age. The results not only realized the precise modification of facial area, including areas for skin and hair, but also implement the semantic changes, which are more realistic.

In addition, in order to realize the intensity-controllable face editing based on the principle of homology continuity, we propose a new feather that can be edited. By editing the feather, the intensity can be reasonably controlled compared to the previous ones. Using the coding module based on homology continuity, the face is mapped into the face feature space, and different input conditions

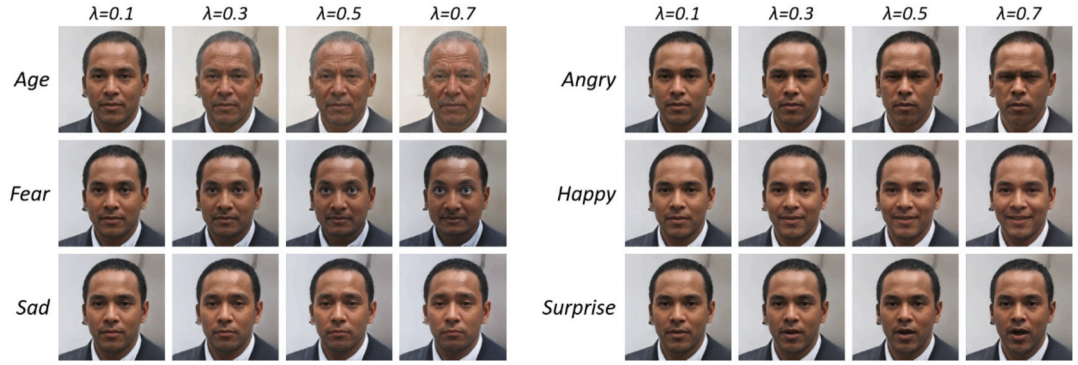

Fig. 8. Results of intensity control. λ means the weight intensity of the attribute axis.

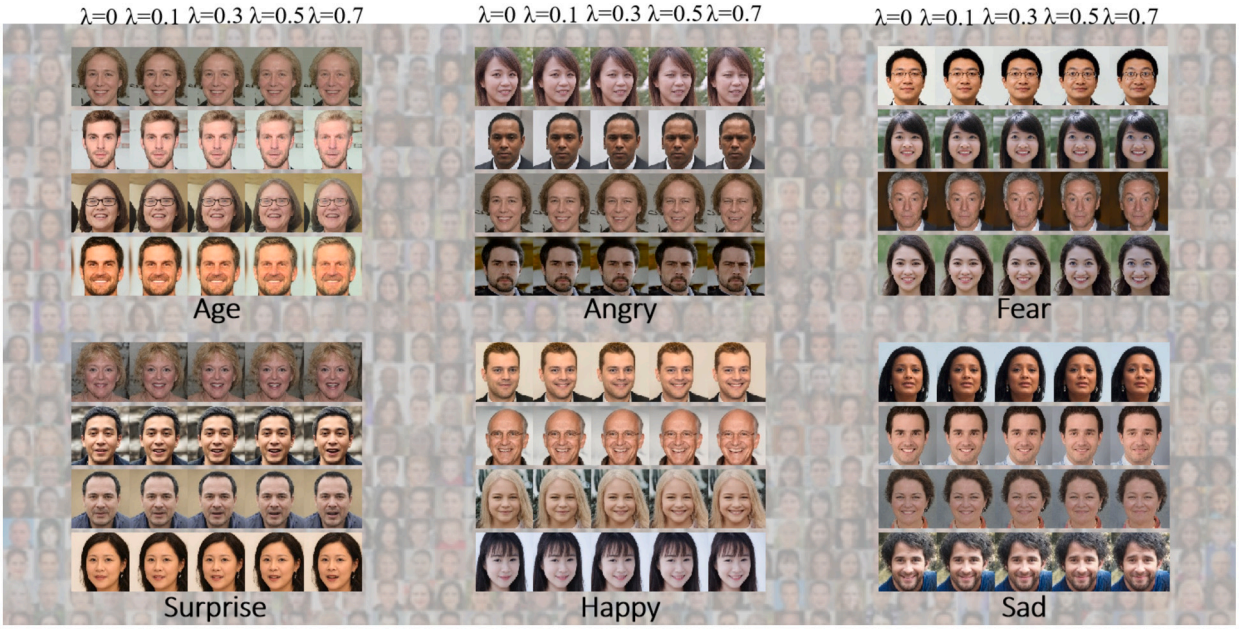

Fig. 9. Results of different image intensity control. λ means the weight intensity of the attribute axis.

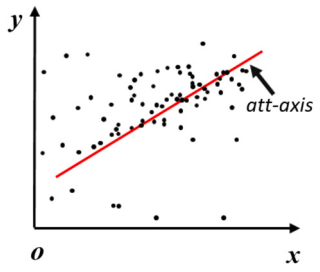

Fig. 10. Attribute regression example diagram.

are used to construct a new representation vector, and then model is based on the face attributes and the representation vector to ensure accurate synthesis and modeling of irrelevant regions, after editing the conditions, achieves control over the strength of face editing (Table 5).

5. Conclusion and discussion

In recent years, many methods have been proposed for facial attribute synthesis. Owing to the availability of large-scale attribute labels, most of these methods are based on deep-generation models. However, owing to different standards, lack of diversity within

Table 5
The results of ablation experiments on Covering Learning on four mainstream face attribute editing datasets.

Method	Datasets											
	RaFD			EmotionNet			CACD2000			Adience		
	PSNR	FID	Inference Time (ms)	PSNR	FID	Inference Time (ms)	PSNR	FID	Inference Time (ms)	PSNR	FID	Inference Time (ms)
ICGNet +only Adam	28.64	13.32	99.5	29.88	12.31	96	26.95	12.13	100	28.64	14.75	106
ICGnet +CL (Ours)	33.14	15.28	78	34.21	15.11	76	32.97	14.85	81	32.76	17.84	85

the class, and discretization of labels, existing methods have problems, such as low quality and large differences in synthetic results. This study systematically investigated the different methods currently used for face attribute synthesis. Moreover, we propose a multilevel network, ICGNet, that utilizes a fusion of different scale features to ensure the effectiveness and sufficiency of information extraction and applies progressive synthesis, depending on the context information, to obtain high-quality synthetic results. Most importantly, we used high-order neurons and the homology continuity principle to realize continuous control of the intensity of face attributes and simultaneously maintain face identity information at the same time. We verified the effectiveness of the proposed method through qualitative and quantitative index analyses of related tasks.

Although ICGNet can significantly improve face editing tasks, certain problems cannot be ignored. The shortcomings of ICGNet, due to the design principle, are mainly. The method's limitations primarily involve the following aspects, and we have discussed potential solutions. Imbalanced facial attribute data in the dataset: the encoder may be more sensitive to frequently occurring attributes while having weaker recognition and synthesis capabilities for rare attributes. Potential solutions include data augmentation techniques such as sample augmentation and data synthesis to balance the distribution of each attribute's samples. A weighted strategy can also be applied to the loss function, giving larger weights to the training samples of rare attributes to strengthen their learning. In some scenarios, facial attributes, such as facial expressions and lighting conditions, may exhibit diversity and varying intensities. The encoder may face limitations in handling these situations because the principle of homology continuity may struggle to maintain continuity for excessively large attribute variations. Potential solutions include introducing more sophisticated attribute modeling methods, such as incorporating multiple encoders to handle different attribute subspaces or adopting more flexible attribute interpolation algorithms to accommodate attributes with different intensity variations. Furthermore, in certain cases, conflicts may exist between different attributes, such as the relationship between wearing glasses and facial expressions. The encoder may struggle to satisfy all the requirements of all attributes, resulting in suboptimal synthesis results. Potential solutions include introducing constraints or prior knowledge to guide the encoder to make more reasonable choices when dealing with conflicting attributes. Additionally, joint training or multi-objective optimization can be employed to consider multiple attributes and seek a balanced solution simultaneously. There is limited research on multimodal attribute editing, and relying solely on visual guidance may not fully satisfy users' needs. Therefore, we aim to explore attribute editing based on multimodal data by combining images and semantic descriptions for editing purposes. By integrating multiple sources of information into the editing process, more comprehensive and diverse attribute editing effects can be achieved. In the future, we intend to pay more attention to facial editing techniques' societal and privacy aspects. We will investigate the ethical considerations and protective measures to strike a balance between facial-editing technology and societal privacy concerns.

CRediT authorship contribution statement

Xin Ning: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Feng He:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Xiaoli Dong:** Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Weijun Li:** Conceptualization, Data curation, Investigation, Software, Supervision, Writing – original draft, Writing – review & editing. **Fayadh Alenezi:** Conceptualization, Data curation, Funding acquisition, Methodology, Validation, Visualization. **Prayag Tiwari:** Conceptualization, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare no conflict of interests.

Data availability

Github link is shared in the paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (no. 62373343) and Beijing Natural Science Foundation (no. L233036).

References

[1] C. Chen, D. Carlson, Z. Gan, C. Li, L. Carin, Bridging the GAP between stochastic gradient MCMC and stochastic optimization, in: Artificial Intelligence and Statistics, PMLR, 2016, pp. 1051–1060.
[2] L. Chen, Z. You, N. Zhang, J. Xi, X. Le, UTRAD: anomaly detection and localization with U-transformer, Neural Netw. 147 (2022) 53–62.
[3] T. Muralidharan, N. Nissim, Improving malicious email detection through novel designated deep-learning architectures utilizing entire email, Neural Netw. (2022).
[4] P. Dong, L. Wu, L. Meng, X. Meng Hr-prgan, High-resolution story visualization with progressive generative adversarial networks, Inf. Sci. 614 (2022) 548–562.
[5] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, AttGAN: facial attribute editing by only changing what you want, IEEE Trans. Image Process. 28 (2019) 5464–5478.
[6] Q. Jiao, J. Zhong, C. Liu, S. Wu, H.-S. Wong, Perturbation-insensitive cross-domain image enhancement for low-quality face verification, Inf. Sci. 608 (2022) 1183–1201.
[7] R. Abdal, Y. Qin, P. Wonka, Image2StyleGAN++: how to edit the embedded images?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8296–8305.
[8] J. Zhu, D. Zhao, B. Zhang, B. Zhou, Disentangled inference for GANs with latently invertible autoencoder, Int. J. Comput. Vis. 130 (2022) 1259–1276.
[9] Q. Huang, C. Huang, X. Wang, F. Jiang, Facial expression recognition with grid-wise attention and visual transformer, Inf. Sci. 580 (2021) 35–54.
[10] B.-C. Chen, Y.-Y. Chen, Y.-H. Kuo, W.H. Hsu, Scalable face image retrieval using attribute-enhanced sparse codewords, IEEE Trans. Multimed. 15 (2013) 1163–1173.
[11] D.P. Kingma, J. Ba Adam, A method for stochastic optimization, arXiv preprint, arXiv:1412.6980, 2014.
[12] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, StarGAN: unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[13] X. Xia, X. He, L. Feng, X. Pan, N. Li, J. Zhang, X. Pang, F. Yu, N. Ding, Semantic translation of face image with limited pixels for simulated prosthetic vision, Inf. Sci. 609 (2022) 507–532.
[14] Y. Jo, J. Park, SC-FEGAN: face editing generative adversarial network with user's sketch and color, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1745–1753.
[15] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, Adv. Neural Inf. Process. Syst. 30 (2017).
[16] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, arXiv preprint, arXiv:1710.10196, 2017.
[17] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, A. Torralba, Inverting layers of a large generator, in: ICLR Workshop, vol. 2, 2019, p. 4.
[18] S. Wang, J. Lai, First Step to Multi-Dimensional Space Biomimetic Infomatics, National Defense Industry Press, Beijing, China, 2008, pp. 2–25.
[19] M.J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, J. Budynek, The Japanese female facial expression (JAFFE) database, in: Proceedings of Third International Conference on Automatic Face and Gesture Recognition, 1998, pp. 14–16.
[20] D. Lundqvist, A. Flykt, A. Öhman, Karolinska directed emotional faces, Cogn. Emot. (1998).
[21] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, M. Bartlett, The computer expression recognition toolbox (CERT), in: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), IEEE, 2011, pp. 298–305.
[22] O. Langner, R. Dotsch, G. Bijlstra, D.H. Wigboldus, S.T. Hawk, A. Van Knippenberg, Presentation and validation of the Radboud Faces Database, Cogn. Emot. 24 (2010) 1377–1388.
[23] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensics Secur. 9 (2014) 2170–2179.
[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[25] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), IEEE, 2000, pp. 46–53.
[26] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE, 2010, pp. 94–101.
[27] H. Anas, B. Rehman, W.H. Ong, Deep convolutional neural network based facial expression recognition in the wild, arXiv preprint, arXiv:2010.01301, 2020.
[28] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2852–2861.
[29] C. Fabian Benitez-Quiroz, R. Srinivasan, A.M. Martinez, EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5562–5570.
[30] D. Gera, S. Balasubramanian, Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition, Pattern Recognit. Lett. 145 (2021) 58–66.
[31] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, DISFA: a spontaneous facial action intensity database, IEEE Trans. Affect. Comput. 4 (2013) 151–160.
[32] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, W. Zhang, FERV39K: a large-scale multi-scene dataset for facial expression recognition in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20922–20931.
[33] A. Lanitis, C.J. Taylor, T.F. Cootes, Toward automatic simulation of aging effects on face images, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 442–455.
[34] B.-C. Chen, C.-S. Chen, W.H. Hsu, Cross-age reference coding for age-invariant face recognition and retrieval, in: European Conference on Computer Vision, Springer, 2014, pp. 768–783.
[35] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, Int. J. Comput. Vis. 126 (2018) 144–157.
[36] K. Karkkainen, J. Joo, FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1548–1558.
[37] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.
[38] X. Ning, W. Tian, F. He, X. Bai, L. Sun, W. Li, Hyper-sausage coverage function neuron model and learning algorithm for image classification, Pattern Recognit. 136 (2023) 109216.
[39] B.D. MacArthur, A. Lachmann, I.R. Lemischka, A. Ma'ayan, GATE: software for the analysis and visualization of high-dimensional time series expression data, Bioinformatics 26 (2010) 143–144.
[40] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, Adv. Neural Inf. Process. Syst. 30 (2017).
[41] Q. Huynh-Thu, M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electron. Lett. 44 (2008) 800–801.

[42] X. Luo, X. He, X. Chen, L. Qing, J. Zhang, DualG-GAN, a dual-channel generator based generative adversarial network for text-to-face synthesis, Neural Netw. 155 (2022) 155–167.
[43] A. Creswell, A.A. Bharath, Inverting the generator of a generative adversarial network, IEEE Trans. Neural Netw. Learn. Syst. 30 (2018) 1967–1974.
[44] T. d, S. Farias, J. Maziero, Feature alignment for approximated reversibility in neural networks, arXiv preprint, arXiv:2106.12562, 2021.
[45] B. Jing, G. Corso, R. Berlinghieri, T. Jaakkola, Subspace diffusion generative models, arXiv preprint, arXiv:2205.01490, 2022.
[46] X. Ning, W. Tian, Z. Yu, W. Li, X. Bai, Y. Wang, HCFNN: high-order coverage function neural network for image classification, Pattern Recognit. 131 (2022) 108873.
[47] X. Hou, X. Zhang, H. Liang, L. Shen, Z. Lai, J. Wan, GuidedStyle: attribute knowledge guided style manipulation for semantic face editing, Neural Netw. 145 (2022) 209–220.
[48] Y. Xu, Y. Yin, L. Jiang, Q. Wu, C. Zheng, C.C. Loy, B. Dai, W. Wu, Transeditor: transformer-based dual-space GAN for highly controllable facial editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7683–7692.
[49] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[50] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.