The Visual Computer
https://doi.org/10.1007/s00371-024-03461-9

RESEARCH

Free editing of Shape and Texture with Deformable Net for 3D Caricature Generation

Yuanyuan Iin[1] · Ju Dai[2] · Junjun Pan[1] · Feng Zhou[3] · Junxuan Bai[4,5]

Accepted: 4 May 2024

Abstract
2D caricature editing has shown superior performance. However, 3D exaggerated caricature face (ECF) modeling with flexible shape and texture editing capabilities is far from achieving satisfactory high-quality results. This paper aims to model shape and texture variations of 3D caricatures in a learnable parameter space. To achieve this goal, we propose a novel framework for highly controllable editing of 3D caricatures. Our model mainly consists of the texture and shape hyper-networks, texture and shape Sirens, and a projection module. Specifically, two hyper-networks take the texture and shape latent codes as inputs to learn the compact parameter spaces of the two Siren modules. The texture and shape Sirens are leveraged to model the deformation variations of textural styles and geometric shapes. We further incorporate precise control of the camera parameters in the projection module to enhance the quality of generated ECF results. Our method allows flexible editing online and swapping textural features between 3D caricatures. For this purpose, we contribute a 3D caricature face dataset with textures for training and testing. Experiments and user evaluations demonstrate that our method is capable of generating diverse high-fidelity caricatures and achieves better editing capabilities than state-of-the-art methods.

Keywords 3D exaggerated caricature face · Texture modeling · Shape reconstruction · Latent code

1 Introduction

ECF modeling is an essential aspect of emerging applications such as Metaverse [3], role-playing [5], and animation filmmaking [17]. Allowing the free and flexible editing of geometric shape and texture styles of caricature faces can significantly improve the user stickiness of these applications.

Thanks to the advancement of the generative adversarial network (GAN) [30, 40], conspicuous works [25, 33] of 2D ECF modeling have been achieved, which can execute exaggerated shape and textural styles reconstruction. However, 2D ECF generation lacks depth and geometric shape information and maintains restricted expressive and modeling capabilities. Unfortunately, the 3D ECF is not well explored. The primary challenge lies in producing high-quality 3D caricatures with exaggerated shapes and diverse textures while maintaining controllability over shape and texture.

To tackle the above shortages, earlier endeavors [19, 28] shift their attention to producing 3D ECF through manual manipulation using traditional tools such as OpenGL and UV mapping. Despite the impressive performance, the manual manner is a grueling and labor-intensive process. Recent progress to reconstruct 3D ECF is mainly based on 3D face data [19, 31]. However, those approaches do not allow flexible customization, such as shape exaggeration, texture editing, and style variations. The deep-deformable model [19] is capable of generating exaggerated 3D caricature shapes through an Multilayer Perceptron (MLP)-based framework for building a deformable surface model, but it can not perform multiple styles of cross-caricature editing. Although 3D-CariNet [34] can edit caricatures with a fixed number of styles, it must prepare the styles in advance and can only edit offline. Toward these challenges, this work aims

✉ Ju Dai
daij@pcl.ac.cn

✉ Junjun Pan
pan_junjun@buaa.edu.cn

1 Beihang University, Beijing, China

2 Peng Cheng Laboratory, Shenzhen, China

3 North China University of Technology, Beijing, China

4 Institute of Artificial Intelligence in Sports, Capital University of Physical Education and Sports, Beijing, China

5 Emerging Interdisciplinary Platform for Medicine and Engineering in Sports (EIPMES), Beijing, China

Published online: 23 May 2024

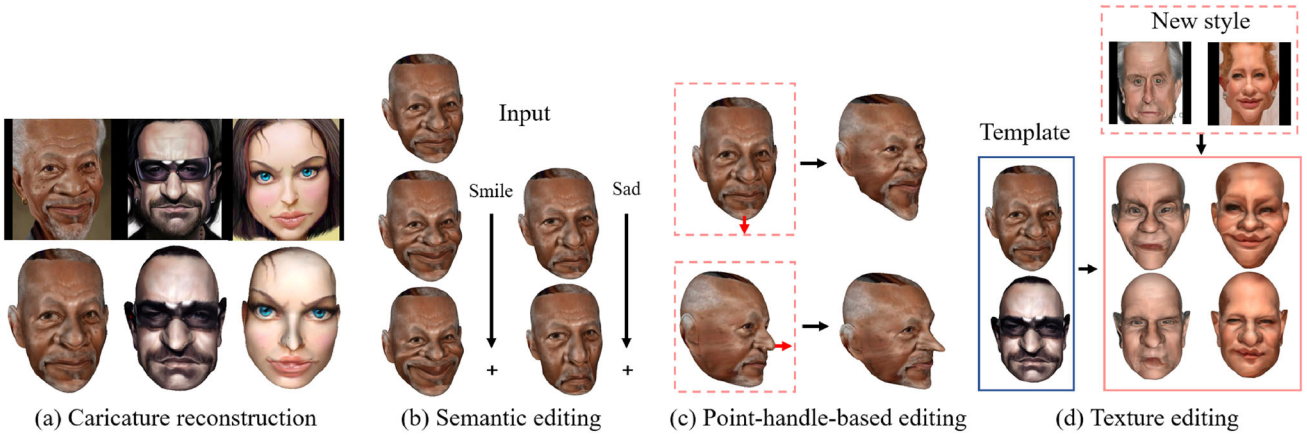

Fig. 1 We present a novel approach to flexible editing shapes and textures of 3D caricatures. In testing, our network can reconstruct high-quality 3D caricatures (a), enable varying degrees of semantic editing (b), conduct point-handle-based editing (c), and implement cross-character texture editing with learned latent codes (d)

to provide an agile manner for 3D ECF modeling: the ability to manipulate 3D face shapes flexibly and the ability to perform stylistic variations and detail transport among different caricatures.

In this paper, we propose a novel framework for 3D ECF generation, which leverages a pair of learnable latent codes to represent the shape and texture of caricature faces and flexibly achieve shape and texture manipulation. Our network mainly consists of the texture hyper-network, shape hyper-network, texture siren, shape siren, and projection module. The two hyper-networks take the latent codes as input and generate transformation parameters for the siren module. The two siren networks concentrate on learning deformable values of 3D caricature texture and shape. The projection module aims to improve generation quality with reverse mapping using camera parameters. We optimize the editable latent codes and the generation network through the 3D modeling loss and 2D projection loss. The 3D caricature face modeling loss enables disentangling geometric shapes and textural style information, and the 2D projection loss facilitates improved perceptual quality. Our model not only maintains the topological structure consistency of shape and texture but also can decompose the texture and shape during decoding. The powerful editing abilities of our work are demonstrated in Fig. 1. Furthermore, we construct a 3D training dataset to research high-quality 3D caricatures. We improve the mesh density of the 3DCaricShop [31] dataset and re-divide the original meshes of eyes and mouth. Experiments on the 3DCaricShop dataset and user evaluations from 20 animation professionals validate the superiority of our method against existing advanced methods. In summary, the contributions of our method are mainly threefold:

- We propose a novel end-to-end network to generate high-quality 3D caricatures from 2D images with flexible, editable, and controllable capacities.
- Our method allows for diversified editing, consisting of semantic editing, point-handle-based editing, and textural style editing across different characters.
- We contribute a new high-quality dataset of 3D caricature faces with corresponding camera parameters to validate the superiority of the proposed method against cutting-edge methods.

2 Related Work

2.1 2D Caricature Editing

Caricature is drawing significant attention for artistic expressions such as satire and humor through exaggerated geometric shapes and stylistically diverse textures [1, 8, 13]. Earlier endeavors have investigated learning the shaped deformation and appearance texture of caricature faces [11, 12, 17, 24]. However, those methods still have room for improvement in the quality of caricature production and the exaggerations.

With the emergence of GAN and its superiority in content generation tasks [22, 30, 40], significant progress has been made for ECF with different exaggerated shapes and texture styles [23, 38, 41]. For instance, MW-GAN [16] consists of a stylistic network and a shaped network designed to perform stylistic transfer and shaped exaggeration, respectively. It is capable of generating caricatures with arbitrary stylistic and shaped exaggeration, which can be specified as samples by random sampling of the latent code or from a given caricature. StyleCariGAN [18] enables the automatic creation of realistic and detailed caricatures with optional

control over the degree of shape exaggeration and the type of color stylization. CariGANs [5] is compromised of CariGeoGAN and CariStyGAN modules. The CariGeoGAN models geometry-to-geometry transformations from facial photographs to caricatures, while the CariStyGAN transfers stylized appearance from caricatures to facial photographs without any geometric distortion. Nevertheless, those methods focus on 2D ECF generation from standard face images and cannot perform its 3D counterpart generation. Our work aims to produce 3D caricature faces with geometric shapes and texture styles that can be flexibly edited.

2.2 3D Deformable Shape Model

3D face deformable model [10, 26, 29] is divided into 3D standard face deformable model [6, 9, 20] and 3D exaggerated face deformable model. 3DMM [2] represents a standard face as a morphable model that includes shape and texture parameters. It has been widely used in face reconstruction and face editing [10] and has achieved outstanding performance. According to the evolving demands in industries such as Metaverse and anime, exaggerated face deformation has been gradually emphasized. Researchers divert their attention to learning deformation spaces from standard faces to generate exaggerated face models [37]. However, 3DMM-based deformation spaces lack scalability and have limited expressive capabilities in the 3D caricature field. Constructing a dataset with 2D caricature images and 3D counterpart shapes for creating 3D models of ECF has become a crucial research direction [3, 4, 31]. Meanwhile, some approaches explore the generation of 3D exaggerated faces by portraying facial feature contours to enhance the user interaction experience [14, 15]. However, due to the mesh's severe stretching, the effectiveness of an expression is inadequate.

Another important branch focuses on studying 3D face deformation representations. The primary 3D face deformation can be achieved through the signed distance function (SDF) [7, 28] and the surface deformation function [19, 36]. Usually, the former category learns the continuous SDF representation through MLP and then represents the whole 3D face shape. The latter reconstructs a 3D caricature mesh by modeling a template surface's continuous surface deformation function through MLP-based networks. However, those methods can not simultaneously perform texture rendering and character style editing for 3D ECF generation. In contrast, our work can generate more natural and exaggerated 3D face shapes with diversified texture styles.

2.3 3D Caricature Reconstruction

In the early days, although combining stereo with class-based knowledge could reconstruct 3D faces [35], and manipulate faces' 3D shape and 2D surface reflectance components to achieve 3D face editing [27], these methods struggle to achieve diversified shape editing and flexible cross-character style variations.. With significant progress in 2D caricature editing and 3D deformable shape modeling, 3D ECF generation and editing have attracted much attention.

The most impressive and related to ours works are 3D-CariGAN [39], 3DCarishop [31], 3D-CariNet [34], 3DMagicMirror [42]. 3D-CariGAN [39] proposes an end-to-end network that transforms a normal face photo into a 3D caricature. However, it only generates exaggerated geometric shapes without texture effects and cannot edit shapes and textures. 3DCarishop [31] is an important baseline work, which collects the first large-scale 3D caricature dataset containing 2000 high-quality diversified 3D caricatures manually crafted by professional artists. 3DCarishop also proposes a novel view-collaborative graph convolution network (VC-GCN) to generate high-fidelity 3D caricature. 3D-CariNet [34] learns the shape and texture of 3D caricatures by designing a graph convolutional autoencoder to build a non-linear colored mesh model. 3DMagicMirror [42] reconstructs 3D face shapes from photographs and every frame of video and then converts 3D face shapes from regular style to caricature style. All of these efforts require the preparation of datasets of different caricature styles before realizing the caricature face style transformation. In contrast, our method can achieve shape and cross-character style editing with high-quality flexibility by disentangling the shape and texture.

3 Methodology

3.1 Overview of the proposed method

Our work aims to generate 3D exaggerated caricatures automatically with high-fidelity performance and online editing capabilities. To achieve this goal, we propose a novel network that combines the advantages of 3D caricature mesh representation power (flexibility and the ability to edit easily) with the high-fidelity texture of 2D images. An overview of the proposed framework is illustrated in Fig. 2. Our model mainly consists of five components: texture hyper-network, shape hyper-network, texture Siren, shape Siren, and projection block. The two hyper-networks receive the latent codes of texture and shape as inputs and produce intermediate latent representations, which will be transformed into the parameters for texture Siren and shape Siren. The two Siren modules concentrate on modeling each shape and texture as a deformation of a fixed template surface. The projection block projects the generated 3D caricature faces into corresponding 2D counterparts to leverage the 2D ground truth caricature faces to supervise the network training and enhance the 3D caricature face generation quality.

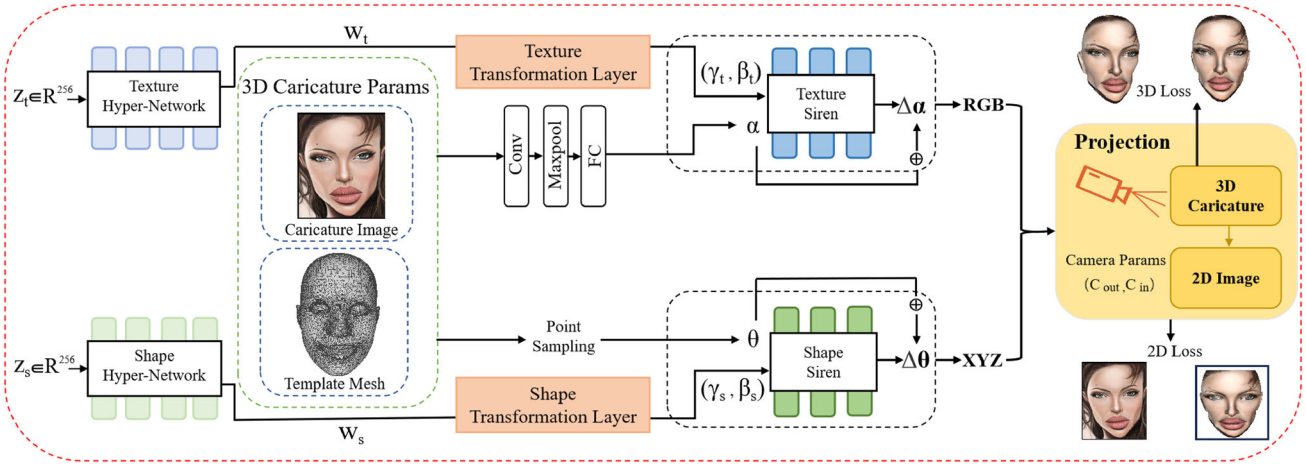

Fig. 2 Overview of the proposed framework. Our model consists of texture and shape hyper-networks, texture and shape Sirens, and a projection model. The network maintains two latent codes for the shape and texture of 3D caricature faces, produces the parameters of the two Siren modules using the two hyper-networks and the corresponding transformation layers, predicts shape and texture deformation values leveraging two Sirens, and aligns the shape and texture based on the pixel alignment of each light given by the camera parameter in the projection module. We optimize the model through a 3D reconstruction loss and a 2D projection loss

Choosing appropriate 3D face representation is vital for exaggerated caricature generation. As validated in [19], face mesh and surface demonstrate promising results. Therefore, we leverage a standard template mesh as the geometry shape representation. Our constructed face mesh consists of 35, 200 vertexes and 70, 383 surfaces, and the coordinate of each vertex is denoted as $\theta_i = (x, y, z) \in \mathbb{R}^3$. Besides the original vertexes, we also uniformly sample 17, 600 points on the surface. We concatenate the two vertexes and obtain 52, 800 vertexes $\theta = \{\theta_i\} \in \mathbb{R}^{52800\times 3}$ to serve as inputs for the shape Siren. Textures of 3D caricatures come from corresponding caricature images. Since the input dimension of the image is $512 \times 512 \times 3$, to ensure that the dimension of the caricature image is consistent with the geometry shape coordinate, we first implement convolution, max pooling, and linear transformation operations for the input image to obtain harmonious input $\alpha \in \mathbb{R}^{52800\times 3}$ for texture Siren. To manipulate the texture and shape, We construct a pair of latent codes $\mathbf{z}_t \in \mathbb{R}^{256}$ and $\mathbf{z}_s \in \mathbb{R}^{256}$ to store texture and shape information, respectively. Since we represent 3D caricature texture and shape as 3D vectors separately and model their variations, we can obtain 3D ECF using an efficient and uniform network structure.

3.2 Network Structure

In this section, we elaborate on the key components, consisting of the texture hyper-network, shape hyper-network, texture Siren, shape Siren, and projection modules. The two hyper-networks possess the same architecture but do not share parameters, so as to the two Siren modules. Thus, we take the texture hyper-network and texture Siren as examples to give details.

Texture Hyper-Network. The presented texture hyper-network aims to model the function variations of 3D caricature textures. Inspired by StyleGAN [21], the texture hyper-network takes the latent code as input and produces the intermediate latent codes. Since the input latent space must follow the probability density of the 3D ECF data, which leads to some degree of unavoidable entanglement. As observed in [21], the intermediate latent space is capable of being free from the restriction mentioned above. Thus, we first utilize the hyper-network to learn intermediate representation and then transform it into the desired parameters required by the texture Siren.

Our texture hyper-network is made up of four hyper-network blocks, and we illustrate one hyper-network block in Fig. 3, which consists of a fully connected (FC) layer and a LeakyReLu activation layer. Precisely, the texture hyper-network receives the texture latent code $\mathbf{z}_t$ as input to learning the intermediate representation $\mathbf{w}_t$, which will be transformed into frequency γ_t and phase shift β_t through texture transformation layer [21]. γ_t and β_t are leveraged to condition each layer of the subsequent texture Siren. The whole learning process can be expressed as follows:

$$\mathbf{w}_t = F_{th}(\mathbf{z}_t; \phi_{th}), \tag{1}$$

$$\gamma_t, \beta_t = F_{ta}(\mathbf{w}_t; \phi_{ta}), \tag{2}$$

where F_{th} and F_{ta} denote the texture hyper-network and texture transformation layer. ϕ_{th} and ϕ_{ta} are corresponding parameters that should be optimized.

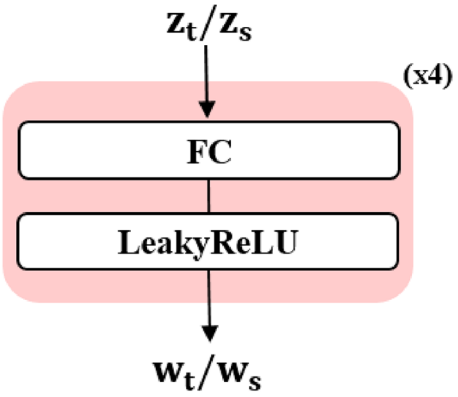

Fig. 3 The hyper-network block architecture. ×4 means stacking four hyper-network blocks

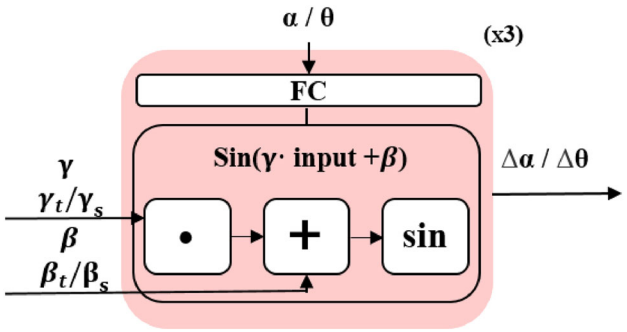

Fig. 4 The siren block architecture. ×3 means stacking three Siren blocks

Shape Hyper-Network. The shape hyper-network and transformation layer concentrate on producing parameters required by the shape Siren. Both have the same structure as the texture hypernetwork and transformation layer without sharing parameters. Thus, the data processing flows have exactly the same forms, which transforms shape latent code $\mathbf{z}_s$ to γ_s and β_s through shape transformation layer [21]:

$$\mathbf{w}_s = F_{sh}(\mathbf{z}_s; \phi_{sh}), \tag{3}$$

$$\gamma_s, \beta_s = F_{sa}(\mathbf{w}_s; \phi_{sa}), \tag{4}$$

where F_{sh} and F_{sa} signify the shape hyper-network and shape transformation layer, $\mathbf{w}_s$ is the intermediate representation for caricature shape. ϕ_{sh} and ϕ_{sa} refer to the corresponding parameters.

Texture Siren. In our framework, we model the deformation values of texture for learning the latent space of 3D caricature instead of directly modeling texture values. We find it is more effective to learn the deformation values and model them as continuous functions. We leverage the texture Siren to represent the constant function defined on the surface of the template.

Our texture Siren is inspired by the FiLM layer [43], which enables conditional adjustment of texture features. The FiLM layer works by first giving a texture input feature and then scaling and offsetting the input texture feature through the sinusoidal function. As shown in Fig. 4, Texture Siren contains a linear FC layer and an affine transformation layer with its frequency γ_t and phase shift β_t coming from the texture hyper-network. The FC layer applies a linear transform defined by a weight matrix $\mathbf{W}_\alpha$ and the biases $\mathbf{b}_\alpha$ to the texture input α. The Texture Siren constructs a learnable Sine function, making the texture of the generated 3D ECF more explicit and enables modeling continuous exaggerated texture variations $\Delta\alpha \in \mathbb{R}^3$.

$$F(\alpha) = \mathbf{W}_\alpha \cdot \alpha + \mathbf{b}_\alpha, \tag{5}$$

$$\Delta\alpha = \sin(\gamma_t \cdot F(\alpha) + \beta_t). \tag{6}$$

With the learned texture deformations $\Delta\alpha$, the final texture representation $\mathbf{p}_\alpha$ of 3D ECF is obtained:

$$\mathbf{p}_\alpha = \alpha + \Delta\alpha. \tag{7}$$

Shape Siren. Shape Siren is used to model the shape deformation values $\Delta\theta$ for 3D ECF reconstruction. The shape Siren structure and shape data processing flow are similar to those of texture Siren.

$$F(\theta) = \mathbf{W}_\theta \cdot \theta + \mathbf{b}_\theta, \tag{8}$$

$$\Delta\theta = \sin(\gamma_s \cdot F(\theta) + \beta_s), \tag{9}$$

$$\mathbf{p}_\theta = \theta + \Delta\theta. \tag{10}$$

where $F(\theta)$ represents the linear transformation result of the input θ defined by the weight matrix $\mathbf{W}_\theta$ and the biases $\mathbf{b}_\theta$. $\mathbf{p}_\theta$ is the geometry shape values.

Projection module. The primary function of the projection module is to find the corresponding grid point of a 3D ECF on the 2D image. Then we can obtain the texture value of the 2D image corresponding point. The 3D face template divides the face into the front face and back face based on the mesh topology. Our work projects the front face of the 3D face mesh onto the 2D image along the light from the center of the camera using the camera's internal parameters C_{int} and external parameters C_{ext}. The projection method and obtaining texture values method are as follows:

$$\theta_{2D} = C_{int} \cdot C_{ext} \cdot \theta, \tag{11}$$

$$\mathbf{p}_{\alpha 2D} = F_P(\theta_{2D}, \mathbf{X}_{2D}), \tag{12}$$

where θ_{2D} is the corresponding 2D coordinate positions of the sampled mesh points projected onto the 2D caricature image. F_P represents an indexing function, which is used to obtain the texture values $\mathbf{p}_{\alpha 2D}$ of pixel positions θ_{2D} on the input 2D image $\mathbf{X}_{2D}$.

3.3 Network Training

Our model aims to generate 3D caricature faces with exaggerated geometry shapes and diversified texture styles given the editable latent codes of shape and texture. To achieve the goal, we leverage the 3D modeling loss $\mathcal{L}_{mod}$ of the corresponding 3D caricatures and 2D projection loss $\mathcal{L}_{pro}$ to optimize the whole network:

$$\mathcal{L} = \mathcal{L}_{mod} + \mathcal{L}_{pro}. \tag{13}$$

3D modeling loss. 3D ECF modeling of our work involves shape and texture learning. Regarding shape modeling, we compute the vertex coordinate distances between the ground truth $\hat{\mathbf{p}}_\theta$ and the generated result $\mathbf{p}_\theta$. The same computation process is adopted for texture modeling. We calculate the texture distance between the ground truth $\hat{\mathbf{p}}_\alpha$ and the deformed value $\mathbf{p}_\alpha$. Thus, the 3D modeling loss consists of the texture term $\mathcal{L}_t$ and shape term $\mathcal{L}_s$, and can be formulated as follows:

$$\mathcal{L}_{mod} = \mathcal{L}_t + \mathcal{L}_s, \tag{14}$$

$$\mathcal{L}_t = \frac{\lambda_1}{N}\sum_i^N \|\mathbf{p}_{\alpha,i} - \hat{\mathbf{p}}_{\alpha,i}\|_2^2 + \frac{\lambda_2}{d}\|\mathbf{z}_t\|_2^2, \tag{15}$$

$$\mathcal{L}_s = \frac{\lambda_1}{N}\sum_i^N \|\mathbf{p}_{\theta,i} - \hat{\mathbf{p}}_{\theta,i}\|_2^2 + \frac{\lambda_2}{d}\|\mathbf{z}_s\|_2^2, \tag{16}$$

where N refers to the total number of sample points, d denotes latent code dimension and is set to 256, λ_1 signifies the reconstruction weight, λ_2 controls the regularization on the latent codes.

2D projection loss. Generating high-quality textured 3D caricatures requires 2D images with caricature styles. Thus, we can project the 3D caricatures into their 2D counterparts. We obtain corresponding coordinate points of 3D ECF mesh on a 2D image by the projection module. We supervise the learning process by calculating the distance between mesh point texture values and the corresponding projected results texture values. We utilize the 2D projection loss $\mathcal{L}_{pro}$ to optimize the whole network:

$$\mathcal{L}_{pro} = \frac{1}{N}\sum_i^N \|\mathbf{p}_{\alpha,i} - \mathbf{p}_{\alpha 2D,i}\|_2^2, \tag{17}$$

where $\mathbf{p}_{\alpha,i}$ is the i-th mesh point texture values of 3D ECF, and $\mathbf{p}_{\alpha 2D,i}$ is the i-th texture values of the corresponding mesh point projected onto the 2D image.

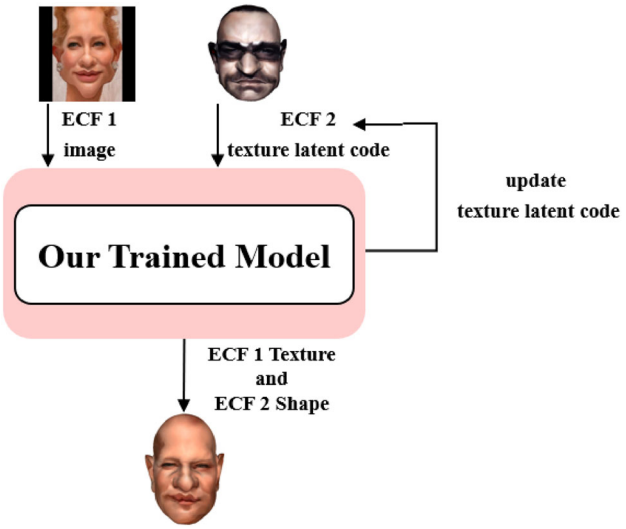

Fig. 5 The workflow of cross caricature texture editing

3.4 Face Editing and Style Customization

We next describe how to apply our network for face shape editing and texture style editing. The generation results are illustrated in Fig. 1.

Texture editing. Since the learned shape and texture feature spaces are separate, we can update the texture features while keeping the shaped features unchanged. After finishing optimizing the proposed network, each sampled texture latent code in the texture latent space represents a 3D ECF texture feature. Thus, we can transfer texture features from one caricature to another with the learned texture latent space. We transfer the reference 2D image texture to a 3D ECF by optimizing the new texture latent space $\mathbf{z}_{t,new}$. We exhibit the workflow of cross caricature texture editing in Fig. 5. Given the reference 2D caricature image and the target 3D ECF, the relationship between the texture latent code before transformation$\mathbf{z}_t$ and the texture latent code after transformation $\mathbf{z}_{t,new}$ is calculated by calculating the 3D ECF texture reconstruction loss:

$$\mathbf{z}_{t,new} = argmin_{\mathbf{z}_t}\mathcal{L}_t. \tag{18}$$

Semantic editing. Regarding shape editing, our work allows us to manipulate the learned shape latent space of 3D caricature faces directly for semantic editing. We resort to the single attribute manipulation technique of InterFaceGAN [32] for semantic editing. Each caricature in our training set has a set of semantic labels, and our work can enable semantic editing on the optimized latent shape space.

Point-handle-based editing. Our work also allows us to manipulate the shape latent space of 3D caricatures directly for point-handle-based editing. We refer to the deep deformable [19] in the deformation editing of 3D caricature

faces, where local deformation and local expansion necessary are considered.

4 Experiments

In this section, we first describe experiment settings. Next, we visualize experimental results and compare them with cutting-edge methods [7, 19, 34] to demonstrate the superiority. Finally, we conduct ablation studies to verify the effectiveness of our network design.

4.1 Experiment Settings

Dataset. Empirical evaluations are conducted on the widely used benchmark 3DCaricShop [31] dataset which owns 2K meshes sculpted by 3D artists. 1,409 registered 3D caricature meshes have been collected by the authors. The registered meshes have vertex connectivity similar to FaceWarehouse [4], with the neck area of the dataset removed and the holes in the eyes and mouth closed. Also, the dataset contains 2,000 sets of face images and corresponding 3D models (done manually by a modeler), labeled with camera parameters and 3D keypoint information.

We fully utilize this dataset's characteristics and select 531 caricature face meshes that are frontal rather than side. Meanwhile, the original meshes of eyes and mouth are deleted and re-divided. Then, the center of gravity of the full-face mesh is obtained, the center of gravity encrypts the mesh, and finally, the face images are inverted and mapped to the 3D model by the camera parameters, and the points on the 2D images are obtained to be the texture of the nearest points on the corresponding 3D model pixel values. In this way, a textured 3D exaggerated caricature face model is constructed. As shown in Fig. 6, the dataset of 3DCaricshop does not have grids for the eyes and nose, while our dataset has those grids. In the meantime, our dataset possesses a full face texture. Ultimately, a better 3D exaggerated caricature dataset with 531 faces is constructed, with 501 caricatures as the training set and the left samples as the testing set.

Training Details. Our model is optimized using the Adam optimizer for 2,000 epochs with a batch size of 4 and a learning rate of 0.0001. The texture and shape latent space dimension is 256. Both texture and shape hyper-networks contain four hyper-network blocks, and each block has an FC layer and a LeakyReLu activation layer. The hidden units for the four FC layers are all set to 256. Further, the texture and shape are constructed with three Siren layers. The first SIREN layer maps the 3D position and 3D texture to 256-dimensional features. The hidden and last SIREN layers have 256 hidden feature dimensions. The input 2D image is sequentially processed by a convolution (Conv) layer with $3 * 3$ kernel and output channels of 3, a $4 * 4$ max pooling layer, and an FC layer with 52800 hidden units. Both the shape and texture latent codes are randomly initialized with normal distribution.

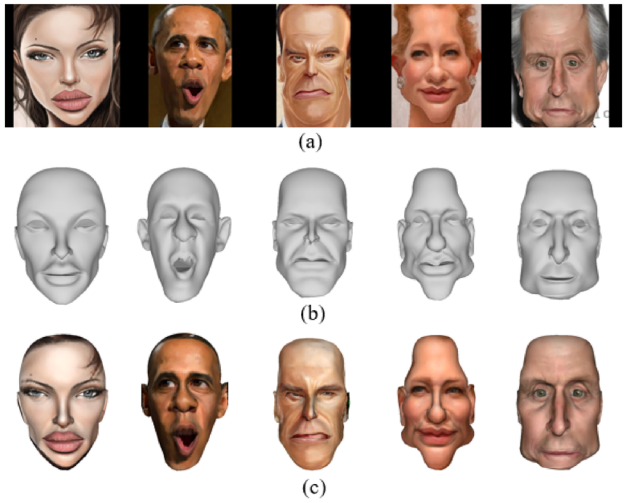

Fig. 6 Some samples of 3DCaricshop [31] dataset and its variants. (a) is the 2D images corresponding to the 3DCaricshop dataset, (b) is the 3DCaricshop dataset, and (c) is our dataset

Evaluation protocol. Our work focuses on flexible editing of exaggerated shapes and texture styles for 3D caricature faces with high-quality generation. We give the visualizations of the texture editing and shape exaggeration with extensive qualitative experiments. At the same time, we provide the numerical evaluation of the mean position loss, which calculates the shape distance between the ground truth and the generated 3D ECF to demonstrate the impact of the generated shapes quantitatively.

4.2 Evaluation

Texture Modeling. Our network enables high-quality texture reconstruction in generating 3D caricatures. To demonstrate its superiority, we compare it with 3D-CariNet [34] and illustrate the results in Fig. 7. 3D-CariNet [34] is a cutting-edge work for generating 3D caricature faces with textures, and its training data is also 3D caricature mesh data. From Fig. 7, it can be observed that the face images generated by 3D-CariNet present artifacts, blurring, and lack of eyes and mouthparts. In contrast, our method can reconstruct the whole 3D face with high-quality texture details. We attribute the promising results of our method to the fact that our texture modeling can learn better texture representation and improve the matching correspondence between the 3D caricature and 2D face image through back projection.

We engage 20 professionals in animation to score the generation of quality in Fig. 7, and ask them to choose the best method regarding texture modeling and caricature exaggeration. The highest and lowest scores are set to 10 and 0,

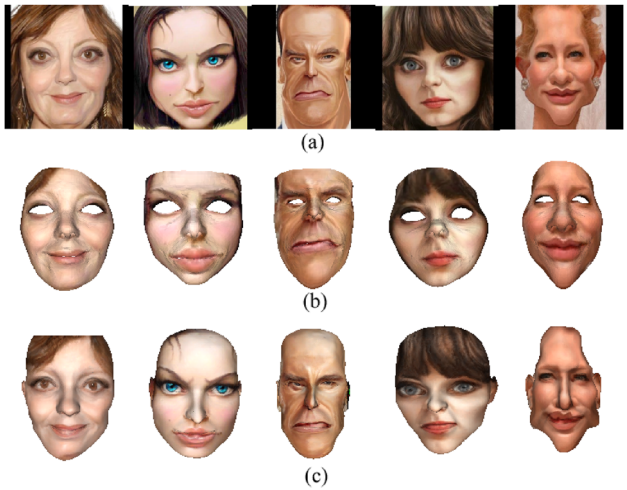

Fig. 7 Visual comparison with 3D-CariNet [34] regarding texture modeling. (a) signifies ground truth. (b) denotes the result generated by 3D-CariNet. (c) is the result generated by our method. Our method can reconstruct the whole 3D face with a higher-quality texture effect

Table 1 Statistical comparison regarding texture modeling. **PTM** denotes percentage about texture modeling and **PCE** means percentage about caricature exaggeration. The best results are denoted in bold font

Method	Average score	PTM	PCE
3D-CariNet [34]	6.250	0.200	0.300
Ours	**8.125**	**0.800**	**0.700**

respectively. The statistical results are reported in Table 1. It can be observed that our method has been widely recognized by experts regarding the three evaluation terms.

Shape Modeling. 3D caricature faces are characterized by exaggerated facial shapes. To identify the geometric shape modeling abilities, we visualize the comparisons with deep deformable [19], DIF-NET [7] in Fig. 8. Compared with deep deformable [19], our method can express more high-fidelity and natural 3D ECF shapes. In addition, we implement quantitative evaluation and compare with the deep deformable. Experimental results of the mean position loss for the 3D ECF testing set are reported in Table 2. It can be seen that our method has the smallest reconstruction error. Compared to DIF-NET, although DIF-NET could reconstruct the overall shape, it ignores important facial details. For example, the part of the eye, we can not even see the shape of the eye. Since the DIF-NET is used to express the face shape in the form of the point cloud, the specific coordinate points corresponding to the coordinate points of the ground truth cannot be found. So, we do not make quantitative comparisons.

Similar to the texture modeling, the 20 professionals also are requested to evaluate deep deformable [19], DIF-NET [7], and our method in Fig. 8 in view of shape modeling. The best shape modeling is asked to be selected. We record the

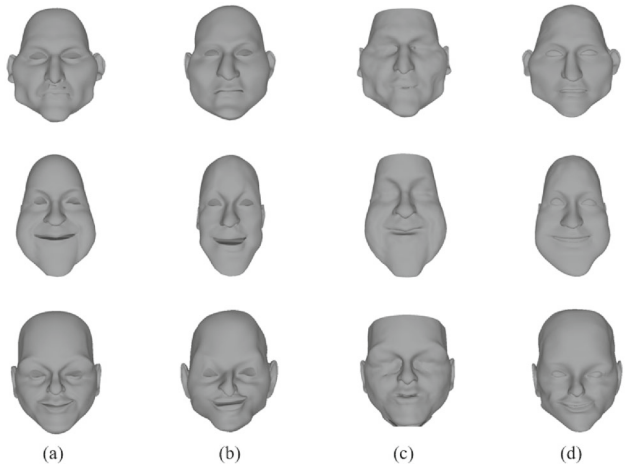

Fig. 8 Visual comparison with deep deformable [19] and DIF-NET [7] regarding shape modeling. (a) signifies the ground truth. (b) denotes the result generated by deep deformable. (c) denotes the result generated by DIF-NET. (d) is the result generated by our method

Table 2 Comparisons with deep deformable [19]. The best results are denoted in bold font

Method	Mean position loss
Deep deformable [19]	0.029
Ours	**0.015**

Table 3 Statistical comparison regarding shape modeling. The results are denoted in bold font

Method	Average score	Percentage
Deep deformable [19]	7.225	0.300
DIF-NET [7]	5.700	0.050
Ours	**7.875**	**0.650**

results in Table 3. From Table 3, we can see that our method is considered to be the best method for shape modeling.

Texture Style Transformation. Since we represent the textured feature of 3D caricature faces in texture latent spaces, we can easily transfer the textures of reference 2D images to target 3D ECF by editing the texture latent space. We summarize the generation effects of texture style editing in Fig. 9 and compare our method with 3D-CariNet [34]. 3D-CariNet requires preparing the data of four style textures in advance and training the texture encoder offline when performing style transformation. 3D-CariNet can only generate textured 3D caricatures within four predefined texture styles. While our work can perform texture editing across character styles online. Most importantly, our work can generate a wide variety of texture styles based on diversified input images, not limited to four styles. From Fig. 9, we can observe that our work can handle texture editing for different characters, flexibly generating a variety of texture styles and preserving

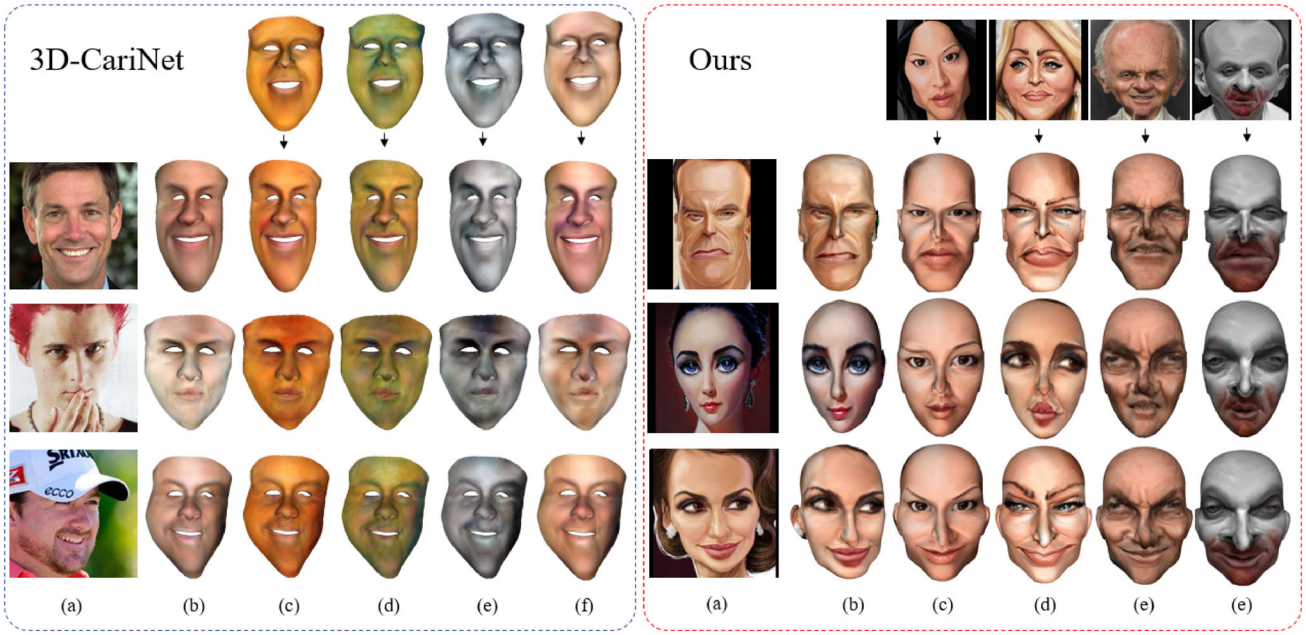

Fig. 9 Visual comparison with 3D-CariNet [34] regarding diversified texture style editing. The blue box shows the experiment results of 3D-CariNet, and the red box shows the experiment results of our method. (a) is the style of the ground truth picture. (b) is the original style from the ground truth picture. (c), (d), (e), and (f) are textural styles for editing

shape details under different 3D caricature faces with high quality and naturalness.

Multi-type Shape Editing. Our method supports multi-type shape editing: semantic-based editing and point-handle-based editing. The dataset we used provides semantic labels for each caricature. Through network structure learning, we obtain the orientation of each attribute and can directly perform semantic editing by manipulating the latent codes. Regarding point-handle-based editing, our work can perform online shape editing by generating reasonable deformations even if the grid points are sparsely located. We demonstrate the editing performance in Fig. 10. For semantic editing, our method can realize editing of 3D ECF with varying degrees through a data-driven editing space for 3D ECF shapes. Our model can make the input images more smiling or more sad. We can also edit 3D ECF using the learned latent space for point-handle-based editing. For the cheek, stretch the two points sideways after picking a point at each side of the cheek. For the nose, move the point to the front after selecting a point on the nose tip. For ears, stretch the two points after picking a point at each side of the ears.

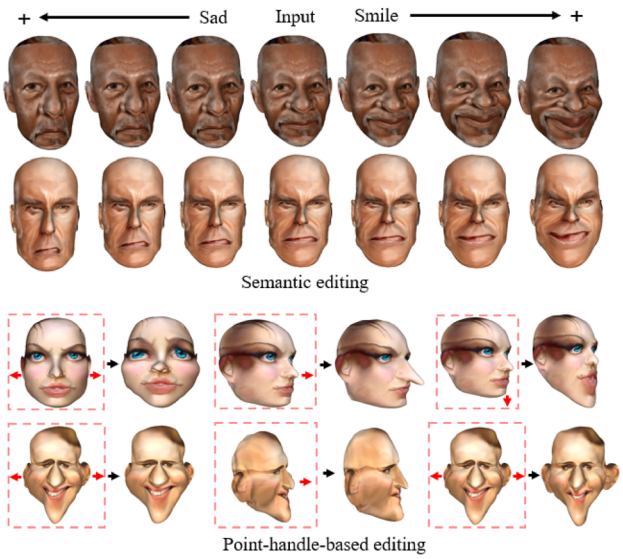

Fig. 10 Visualizations of our method for fine-grained semantic editing and multi-dimensional point-handle-based editing

4.3 Ablation Study

In our framework, we leverage two separate latent spaces to represent the texture and shape features of 3D caricature faces. Further, two individual Sirens are utilized to learn the deformation values of texture and shape. The latent codes of texture and shape can be shared in space. Meanwhile, the same goes for the deformation spaces of texture and shape. Nevertheless, we claim that the separate representations of latent spaces and deformation spaces for texture and shape can obtain better 3D ECF modeling performance. We implement ablation studies to validate our network design.

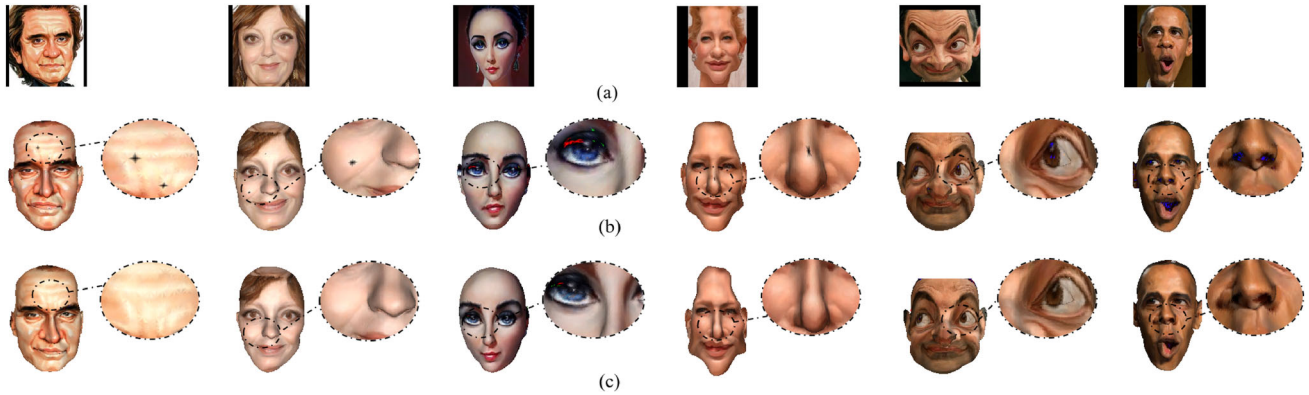

Fig. 11 Ablation study about whether utilizing separated shape and texture latent codes. (a) is the ground truth. (b) is the reconstructed result using merged latent codes. (c) is the reconstructed result using separated latent codes

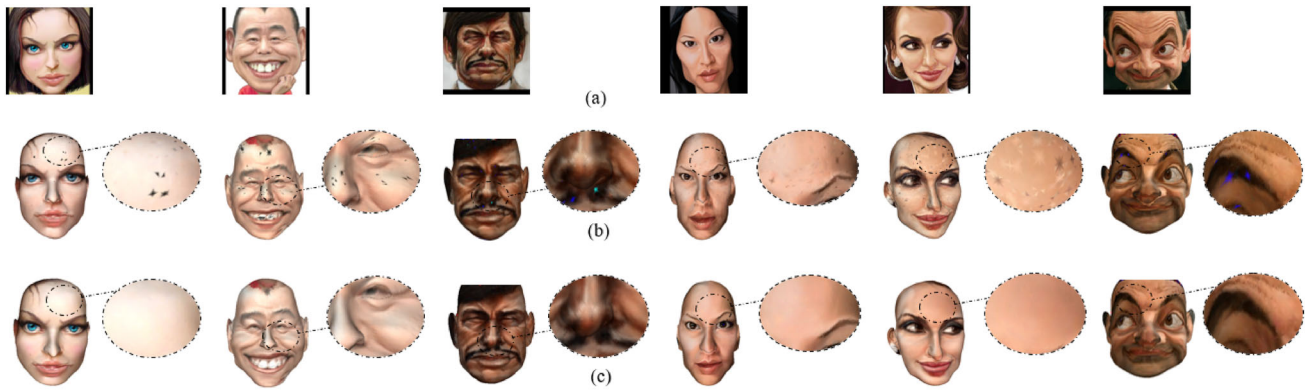

Fig. 12 Ablation study about whether decoupling shape and texture modeling. (a) signifies the ground truth. (b) is the reconstructed result of merging shape and texture. (c) is the reconstructed result of decoupling shape and texture

The importance of separate latent codes. To verify the superiority of separate latent codes, we conduct experiments using merged latent codes to represent caricature texture and shape. The contrast results of utilizing separated and merged shape and texture latent codes are exhibited in Fig. 11. We can observe that merged latent codes result in a lot of texture errors, such as red and black spots. In contrast, separating latent codes achieves high-quality modeling performance for both shape and texture. The possible reason may be that due to the inconsistency between the spatial distribution of shape features and texture features, combining the latent codes will lead to adverse mutual influences for feature learning, resulting in the deviation of both shape and texture from the original feature distributions.

The importance of shape and texture decoupling. We implement comparison experiments of merging or separating shape and texture siren networks to verify the importance of decoupling deformation spaces. The generated 3D caricature faces are illustrated in Fig. 12. It can be summarized from Fig. 12 that merging texture and shape deformation space yields considerable texture errors, such as the blue, green, and black patches. We justify that when integrating the shape and texture of the Siren network, the 3D caricature shape and texture are entangled, leading to the learning of a flawed, not good enough deformation space. On the contrary, decoupling shape and texture can effectively perform feature extraction and editing of shape and texture. In the meantime, we can also keep the consistency of shape and texture through the structural topology of 3D caricature faces.

Fig. 13 3D ECF effects with side 2D images as input. (a) is 2D images. (b) and (c) are 3D results of left and right sides

4.4 Limitation

Currently, we only leverage the frontal images to generate 3D ECF. Since the 3DCaricshop dataset provides only one 2D image for each character, as shown in Fig. 13, texture information of other sides of the generated 3D ECF are missing

when using non-frontal 2D images. If we have 2D data from multiple views, we can reconstruct 3D ECF with full texture through the network. In the future, we will explore using existing datasets to build parametric ECF with prior knowledge and reconstruct 3D ECF through a 2D image.

5 Conclusion

In this paper, we present a novel deformable framework for 3D ECF modeling and editing. Our work can model 3D ECF shapes and textures with high fidelity and high definition. The generated 3D exaggerated caricatures enable shape, semantic, and point-handle-based editing. At the same time, our work allows online texture style editing across multiple caricature characters. Qualitative comparisons with related works and expert evaluation results demonstrate that our approach has higher fitting accuracy and more flexible editing capabilities. Our work can open up more possibilities for accelerating content creation for emerging applications such as meta-universes and animation.

Acknowledgements This research is supported by National Key R&D Program of China (No. 2022ZD0115902), National Natural Science Foundation of China (No. 62102208), Beijing Natural Science Foundation (No. 4232023), Young Elite Scientists Sponsorship Program by BAST (No. BYESS2023382), Beijing Emerging Interdisciplinary Platform for Medicine and Engineering in Sports (EIPMES), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2024C06).

Data availability. Data is available on reasonable request from the corresponding author.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

1. Akleman, E., Palmer, J., Logan, R.: Making extreme caricatures with a new interactive 2d deformation technique with simplicial complexes. In: Proceedings of visual, vol. 1, p. 2000 (2000)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 157–164 (2023)
3. Cai, H., Guo, Y., Peng, Z., Zhang, J.: Landmark detection and 3d face reconstruction for caricature using a nonlinear parametric model. Graphical Models **115**, 101,103 (2021)
4. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2013)
5. Cao, K., Liao, J., Yuan, L.: Carigans: Unpaired photo-to-caricature translation. ACM Transactions on Graphics **37**(6), 244 (2018)
6. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 20,311–20,322 (2022)
7. Deng, Y., Yang, J., Tong, X.: Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10,286–10,296 (2021)
8. Ding, Y., Ma, X., Luo, M., Zheng, A., He, R.: Unsupervised contrastive photo-to-caricature translation based on auto-distortion. In: International Conference on Pattern Recognition, pp. 4520–4527 (2021)
9. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics **40**(4), 1–13 (2021)
10. Galanakis, S., Gecer, B., Lattas, A., Zafeiriou, S.: 3dmm-rf: Convolutional radiance fields for 3d face modeling. In: Winter Conference on Applications of Computer Vision, pp. 3536–3547 (2023)
11. Garg, J., Peri, S.V., Tolani, H., Krishnan, N.C.: Deep cross modal learning for caricature verification and identification (cavinet). In: ACM international conference on Multimedia, pp. 1101–1109 (2018)
12. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
13. Gong, J., Hold-Geoffroy, Y., Lu, J.: Autotoon: Automatic geometric warping for face cartoon generation. In: Winter Conference on Applications of Computer Vision, pp. 360–369 (2020)
14. Han, X., Gao, C., Yu, Y.: Deepsketch2face: a deep learning based sketching system for 3d face and caricature modeling. ACM Transactions on Graphics **36**(4), 1–12 (2017)
15. Han, X., Hou, K., Du, D., Qiu, Y., Cui, S., Zhou, K., Yu, Y.: Caricatureshop: Personalized and photorealistic caricature sketching. IEEE Transactions on Visualization and Computer Graphics **26**(7), 2349–2361 (2018)
16. Hou, H., Huo, J., Wu, J., Lai, Y.K., Gao, Y.: Mw-gan: multi-warping gan for caricature generation with multi-style geometric exaggeration. IEEE Transactions on Image Processing **30**, 8644–8657 (2021)
17. Huo, J., Li, W., Shi, Y., Gao, Y., Yin, H.: Webcaricature: a benchmark for caricature recognition. In: British Machine Vision Conference, p. 223 (2017)
18. Jang, W., Ju, G., Jung, Y., Yang, J., Tong, X., Lee, S.: Stylecarigan: caricature generation via stylegan feature map modulation. ACM Transactions on Graphics **40**(4), 1–16 (2021)
19. Jung, Y., Jang, W., Kim, S., Yang, J., Tong, X., Lee, S.: Deep deformable 3d caricatures with learned shape control. In: ACM SIGGRAPH Conference Proceedings, pp. 1–9 (2022)
20. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics **36**(4), 1–12 (2017)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
22. 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, pp. 8110–8119 (2020)
23. Kim, J., Kim, M., Kang, H., Lee, K.: U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: International Conference on Learning Representations (2020)
24. Laishram, L., Shaheryar, M., Lee, J.T., Jung, S.K.: A style-based caricature generator. In: International Workshop on Frontiers of Computer Vision, pp. 71–82 (2023)

25. Li, W., Xiong, W., Liao, H., Huo, J., Gao, Y., Luo, J.: Carigan: Caricature generation through weakly paired adversarial learning. Neural Networks **132**, 66–74 (2020)
26. Liu, Y., Shu, Z., Li, Y., Lin, Z., Zhang, R., Kung, S.: 3d-fm gan: Towards 3d-controllable face manipulation. In: European Conference on Computer Vision, pp. 107–125 (2022)
27. O'Toole, A.J., Vetter, T., Blanz, V.: Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing. Vision research **39**(18), 3145–3155 (1999)
28. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 165–174 (2019)
29. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: International Conference on Advanced Video and Signal based Surveillance, pp. 296–301 (2009)
30. Pinkney, J.N., Adler, D.: Resolution dependent gan interpolation for controllable image synthesis between domains. arXiv preprint arXiv:2010.05334 (2020)
31. Qiu, Y., Xu, X., Qiu, L., Pan, Y., Wu, Y., Chen, W., Han, X.: 3dcaricshop: A dataset and a baseline method for single-view 3d caricature face reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10,236–10,245 (2021)
32. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 2004–2018 (2020)
33. Shi, Y., Deb, D., Jain, A.K.: Warpgan: Automatic caricature generation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10,762–10,771 (2019)
34. Huang, Meijia and Dai, Ju and Pan, Junjun and Bai, Junxuan and Qin, Hong: 3D-CariNet: End-to-end 3D Caricature Generation from Natural Face Images with Differentiable Renderer. In: Pacific Graphics Short Papers, Posters, and Work-in-Progress Papers (2021)
35. Wallraven, C., Blanz, V., Vetter, T.: 3d-reconstruction of faces: Combining stereo with class-based knowledge. In: Mustererkennung 1999: 21. DAGM-Symposium Bonn, pp. 405–412 (1999)
36. Wang, W., Ceylan, D., Mech, R., Neumann, U.: 3dn: 3d deformation network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1038–1046 (2019)
37. Wu, Q., Zhang, J., Lai, Y.K., Zheng, J., Cai, J.: Alive caricature from 2d to 3d. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7336–7345 (2018)
38. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Pastiche master: Exemplar-based high-resolution portrait style transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7693–7702 (2022)
39. Ye, Z., Xia, M., Sun, Y., Yi, R., Yu, M., Zhang, J., Lai, Y.K., Liu, Y.J.: 3d-carigan: an end-to-end solution to 3d caricature generation from normal face photos. IEEE Transactions on Visualization and Computer Graphics **29**(4), 2203–2210 (2021)
40. Zhao, X., Chen, W., Xie, W., Shen, L.: Style attention based global-local aware gan for personalized facial caricature generation. Frontiers in Neuroscience **17**, 1136,416 (2023)
41. Zheng, Z., Wang, C., Yu, Z., Wang, N., Zheng, H., Zheng, B.: Unpaired photo-to-caricature translation on faces in the wild. Neurocomputing **355**, 71–81 (2019)
42. Zheng, Z., Zhu, J., Ji, W., Yang, Y., Chua, T.S.: 3d magic mirror: Clothing reconstruction from a single image via a causal perspective. arXiv preprint arXiv:2204.13096 (2022)
43. Zhou, P., Xie, L., Ni, B., Tian, Q.: Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. arXiv preprint arXiv:2110.09788 (2021)

Yuanyuan lin, currently a master's student at the School of Software, Beihang University , received her bachelor's degree from Fuzhou University in 2016. She is interested in motion prediction, three-dimensional face modeling, and human posture recognition.

Ju Dai is currently a research assistant fellow in Peng Cheng Laboratory (PCL), Shenzhen, China. She received both B.S. and M.S. degrees in electronic engineering, China University of Geosciences (CUG), Wuhan, China, in 2011 and 2014, respectively, and the Ph.D. degree in signal processing in Dalian University of Technology (DUT), Dalian, China, in 2020. She worked in PCL as postdoctoral research fellow from 2020 to 2022. Her research interests include motion analysis, motion control, character animation, person re-identification, and saliency detection.

Junjun Pan is currently a professor in School of Computer Science, Beihang University. He received both BS and MS degrees in School of Computer Science, Northwestern Polytechnical University, China. In 2006, he studied in National Centre for Computer Animation (NCCA), Bournemouth University, UK, as PhD candidate with full scholarship. In 2010, he received the PhD degree and worked in NCCA as postdoctoral research fellow. From 2012 to 2013, he worked as a research associate in Center for Modeling, Simulation and Imaging in Medicine, Rensselaer Polytechnic Institute, USA. In November 2013, he was appointed as associate professor in School of Computer Science, Beihang University, China. His research interests include virtual surgery and computer animation.

Feng Zhou received the B.S. degree in computer science from Beijing Union University in 2009 and the M.S. degree in computer science and application from Yun Nan University in 2014. He received the Ph.D. degree in computer science from the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China, in 2020. He is currently a lecturer with the School of Information Science and Technology, North China University of Technology, Beijing, China. His research interests include pattern recognition, image processing, virtual reality, computer graphics, and computer vision.

Junxuan Bai is a lecturer at the Institute of Artificial Intelligence in Sports (IAIS), Capital University of Physical Education and Sports (CUPES), Beijing, China. He received a BS degree in mathematics from Dalian Maritime University, Dalian, China, in 2012. He received both MS and PhD degrees in computer science from Beihang University, Beijing, China, in 2015 and 2021, respectively. Then, he worked at China Mobile Research Institute (CMRI) as a researcher from 2021 to 2022. His research interests include computer animation, motion synthesis, motion analysis, and virtual surgery.