3D-CariGAN: An End-to-End Solution to 3D Caricature Generation From Normal Face Photos

Zipeng Ye [ID], Mengfei Xia, Yanan Sun [ID], Ran Yi [ID], Minjing Yu, Juyong Zhang [ID], *Member, IEEE*, Yu-Kun Lai [ID], *Member, IEEE*, and Yong-Jin Liu [ID], *Senior Member, IEEE*

Abstract—Caricature is a type of artistic style of human faces that attracts considerable attention in the entertainment industry. So far a few 3D caricature generation methods exist and all of them require some caricature information (e.g., a caricature sketch or 2D caricature) as input. This kind of input, however, is difficult to provide by non-professional users. In this paper, we propose an end-to-end deep neural network model that generates high-quality 3D caricatures directly from a normal 2D face photo. The most challenging issue for our system is that the source domain of face photos (characterized by normal 2D faces) is significantly different from the target domain of 3D caricatures (characterized by 3D exaggerated face shapes and textures). To address this challenge, we: (1) build a large dataset of 5,343 3D caricature meshes and use it to establish a PCA model in the 3D caricature shape space; (2) reconstruct a normal full 3D head from the input face photo and use its PCA representation in the 3D caricature shape space to establish correspondences between the input photo and 3D caricature shape; and (3) propose a novel character loss and a novel caricature loss based on previous psychological studies on caricatures. Experiments including a novel two-level user study show that our system can generate high-quality 3D caricatures directly from normal face photos.

Index Terms—Face reconstruction, 3D caricature, PCA representation, caricature shape space

---◆---

1 INTRODUCTION

TRADTIONAL 2D caricature is a type of rendered image that uses exaggeration, simplification and abstraction to express the most distinctive characteristics of people [1]. They are also used to express sarcasm and humor for political and social problems. Traditional caricatures drawn by artists are 2D images. Although widely used, they are insufficient for many graphics applications, such as 3D printing, 3D special effects and animation in feature movies, etc. (Fig. 1). 3D caricatures are suitable for these applications, but it is costly and time consuming for artists to create them with professional 3D modeling tools.

In this paper, we study the problem of automatic generation of 3D caricatures from *normal face photos*. It is an extreme cross-domain task where the input is a normal 2D images and the output is an exaggerated 3D meshes whose space and style are both distinct. In the literature, efforts have been made to address problems that solve part of this task. Some recent works aim to translate photos into 2D caricatures [2], [3], [4], [5]. However, image translation mainly focuses on warping and stylization of textures, which only works for 2D images. Wu *et al.* [6] propose a method to reconstruct 3D caricatures from 2D caricatures. However, its input and output are both caricature styles and the focus is mapping from 2D to 3D. Han *et al.* [7] propose a method for generating 3D caricatures from sketches, but the exaggerated deformation is dominated by sketches, which cannot be applied to normal face photos. Both methods [6], [7] require some caricature information as input, which is not easy for non-professional users to provide.

In summary, none of the existing methods can automatically generate 3D caricatures directly from normal face photos. For our extreme cross-domain problem, a straightforward baseline approach is to combine several existing methods [5], [6], [8], i.e., first automatically generate 2D caricatures from photos and then generate 3D caricatures from 2D caricatures. However, this approach is time-consuming and tends to lose information in intermediate steps — because the output 3D caricature cannot directly make use of the information in the input photos — which results in the volatility between the input and the output. In this paper, we present an end-to-end deep learning method for the described problem, which is much more efficient and suitably regularized by making use of information in input photos as much as possible and allows intuitive control.

Another related area to ours is 3D face reconstruction from photos, which has been widely studied in recent years. It is popular to use parametric models to represent faces/heads due to the regular structure of faces. Some parametric models such as the well-known 3D Morphable Model (3DMM) [9] are linear models based on principal component analysis (PCA) of normal 3D faces, which are useful and effective. However, such models do not work for caricature faces due to their limited capability of extrapolation [6]. In this paper, we build a PCA model for 3D caricature meshes, and generate 3D caricature models that can be regarded as interpolation in our PCA space, making the problem more tractable.

Training datasets are indispensable for learning to transform photos to 3D caricatures. For the domain of photos, we use CelebA-Mask-HQ dataset [10] which contains 30,000 portrait photos. For the domain of 3D caricatures, we are not aware of any existing large-scale 3D caricature datasets, so we create our own 3DCari dataset which contains 5,343 3D caricature meshes with the same connectivity. The two datasets are unpaired because it is difficult to obtain the corresponding 3D caricature for a photo. In this paper, we present an end-to-end method named 3D-CariGAN for the automatic generation of 3D caricatures from photos. To train 3D-CariGAN using unpaired training data, we propose a novel character loss and a novel caricature loss, both of which are based on previous psychological studies on caricatures [11], [12], [13]. 3D-CariGAN achieves real-time performance and allows users to interactively adjust the caricature facial shapes, with simple and effective user controls. Experiments including a novel two-level user study shows that our method produces high-quality 3D caricatures from 2D photos in real-time, which is significantly faster and of better quality than the baseline method.

In particular, the contributions of this paper include[1]:

- We create a large dataset of 3D caricatures, and based on this, build a novel PCA-based 3D linear morphable model for 3D caricature shapes.

- Zipeng Ye, Mengfei Xia, Yanan Sun, and Yong-Jin Liu are with BNRist, MOE-Key Laboratory of Pervasive Computing, the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: {yezp17, xmf20, sunyn20}@mails.tsinghua.edu.cn, liuyongjin@tsinghua.edu.cn.
- Ran Yi is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: ranyi@sjtu.edu.cn.
- Minjing Yu is with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China. E-mail: minjingyu@tju.edu.cn.
- Juyong Zhang is with the School of Mathematical Sciences, University of Science and Technology of China, Hefei, Anhui 230026, China. E-mail: juyong@ustc.edu.cn.
- Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF10 3AT, U.K. E-mail: laiy4@cardiff.ac.uk.

1. The dataset, PCA model and source code are available on https://github.com/qq775193759/3D-CariGAN

Fig. 1. Physical prototypes of 3D caricatures with texture.

- We propose the first method to automatically generate 3D caricatures *directly* from normal face photos. Our end-to-end solution addresses cross-domain and cross-style challenges (2D to 3D, and normal photo to caricature) by utilizing a caricature morphable model and introducing novel cross-domain character loss and caricature loss.

2 RELATED WORK

2D Caricature. Many works have studied generating 2D caricatures from photos. The main differences between photos and caricatures are the 2D geometric shape and image style. Some methods [2], [3] focus on geometric exaggeration while other works [4], [14], [15] focus on stylization. The work CariGANs [5] proposes a method that combines these two aspects to generate 2D caricatures using two networks: CariGeoGAN for geometric exaggeration and CariStyGAN for stylization. CariStyGAN disentangles a photo into the style component and the content component, and then replaces the style component by that of a reference or a sample. CariGeoGAN translates the facial landmarks of a photo from a normal shape to those of an exaggerated shape, which are used to warp the image. WarpGAN [16] generates caricatures by warping and stylization. It extracts the content component from the photo, takes a sample in the style latent space, and then transfers the style by combining the content component and sampled style component, which is similar to CariStyGAN. It warps a photo into a caricature while preserving its identity by predicting a set of control points. The stylization in these methods can be adapted to stylize textures for 3D caricatures, but geometric exaggeration for 3D caricatures is more complicated, which is a major focus of our paper.

3D Face Reconstruction. Generating normal 3D faces from photos is well studied in computer graphics. The reader is referred to [17] for a comprehensive survey and the references therein. Due to the regular structure of faces, it is popular to use parametric models to represent faces/heads. 3DMM [9], [18], [19], [20] and multi-linear models [21], [22] are two major types of parametric models. 3DMM is a PCA representation of faces including shapes and textures, and multi-linear models utilize a multi-linear tensor decomposition on attributes such as identity and expression. Parametric models provide a strong constraint to ensure the plausibility of reconstructed 3D face shapes, while substantially reducing the dimensionality of the generation space by regressing the parameters. For this reason, they are widely used for face reconstruction. Recent works [23], [24], [25] use convolutional neural networks (CNNs) to regress the parameters for face reconstruction. However, these methods mainly work for normal photos and generate normal 3D faces. Likewise, existing parametric models do not have enough extrapolation capability to represent 3D caricature faces [6]. This motivates us to build a new 3D caricature parametric model and a new neural network for unpaired cross-domain translation.

3D Caricatures. Although generating 3D caricatures from 2D caricatures or normal photos is similar to 3D face reconstruction, only
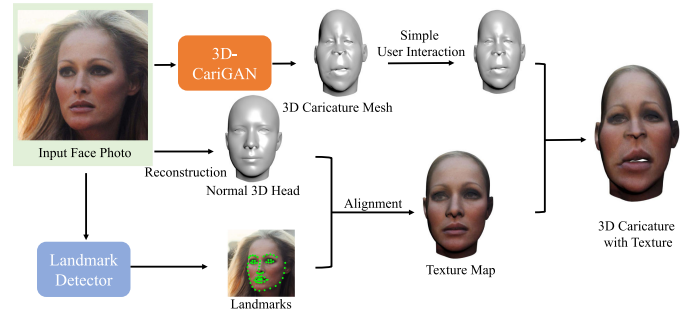

Fig. 2. The pipeline of our method, including 3D-CariGAN for transforming photos to 3D caricature meshes, and landmark based texture mapping for generating textured 3D caricatures. Our system also supports a simple and intuitive user interaction for 3D caricature shapes.

a few works tackle the problem of automatically generating caricatures. Sela *et al.* [26] present a method for directly exaggerating 3D face models, which locally amplifies the area of a given 3D face model based on Gaussian curvature. A deep learning based sketching system [7] is proposed for interactive modeling of 3D caricature faces by drawing facial contours. A method by Clarke *et al.* [27] generates a 3D caricature from a facial photograph and a corresponding 2D hand-drawn caricature which captures the artistic deformation style. However, the method requires paired data as input which is difficult to obtain. An optimization-based method [6] is proposed to reconstruct 3D caricatures from 2D caricatures. This method formulates 3D caricatures as deformed 3D faces. To support exaggeration, their method uses an intrinsic deformation representation which is capable of extrapolation. Therefore, 3D caricature reconstruction is turned into an optimization problem with facial landmark constraints. However, all of these methods rely on 2D sketches or 2D caricatures which contain the information of how to exaggerate and deform the 3D surface, but normal 2D face photos do not have such information. Our work addresses a new challenge of automatically transforming normal 2D face photos to 3D caricatures, without any caricature information used as input.

3 METHOD

The pipeline of our method is illustrated in Fig. 2. A 3D caricature consists of three components: a 3D mesh, a 2D texture image and a texture mapping (represented by texture coordinates). Automatically generating 3D caricatures from photos is decoupled into three steps: (1) we develop 3D-CariGAN to infer 3D caricature meshes from photos; (2) to map the texture to 3D caricature mesh, we reconstruct a normal 3D head model with the same mesh connectivity as the 3D caricature mesh and consistent landmark positions when projected onto the photo; and (3) we use the input photo as the texture image and the projection matrix for texture mapping (by transferring texture coordinates directly from the normal 3D face to the generated 3D caricature mesh).

3.1 Representation of 3D Caricatures

Usually a head mesh has tens of thousands of vertices, e.g., 11,510 vertices used in this paper. A straightforward way is to directly use a neural network for predicting the coordinates of each mesh vertex [28]. However, this requires a very large training set and may produce noisy meshes due to insufficient constraints, especially in our cross-domain scenario. Three examples are shown in Fig. 3 where we adapt our pipeline to directly generate mesh vertices of 3D caricatures (see Section 4.1 for more details). The resulting meshes are rather noisy, indicating insufficient constraints due to the amount of training data available.

In normal 3D face reconstruction, to address similar issues, PCA models such as 3DMM provide a strong prior and reduce the

Fig. 3. 3D caricatures obtained by changing our pipeline to directly generate mesh vertex coordinates rather than 3DCariPCA vectors. This alternative approach produces noisy output meshes, due to the higher dimensional space and lack of constraint.

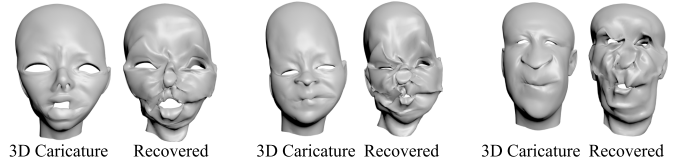3D Caricature   Recovered   3D Caricature Recovered   3D Caricature Recovered

Fig. 4. Using the PCA representation of normal faces to represent 3D caricatures. For each group, the left shows the input 3D caricature model and the right shows its corresponding model represented by the normal PCA representation. These examples show that the representation does not have sufficient extrapolation capability to faithfully reconstruct 3D caricatures. This motivates us to create a PCA model for 3D caricatures.

dimensionality of the generation space. However, all the existing PCA models are only suitable for *interpolation* in the shape space of normal 3D faces and they do not work well for *extrapolation* in 3D caricature shape space. To show this, we use a 100-dimensional PCA representation of normal 3D faces from FaceWarehouse [22] to represent caricature faces, but the recovered caricatures have substantial distortions, as illustrated in Fig. 4. Therefore it is necessary to build a parametric model specifically for 3D caricature shape space.

In our study, we build a PCA model for 3D caricature meshes in the following steps, as illustrated in Fig. 5. Since we are not aware of any existing large 3D caricature datasets, we first collect 5,343 hand-drawn portrait caricature images from Pinterest.com and WebCaricature dataset [29], [30] with facial landmarks extracted by a landmark detector [31], followed by human interaction for correction if needed. An optimization-based method [6] is then used to generate 3D caricature meshes using facial landmarks of 2D caricatures. We use this method to generate $N = 5,343$ 3D caricature meshes of the same topology. We align the pose of the generated 3D caricature meshes with the pose of a template 3D head using an ICP method, where we use 5 key landmarks on the eyes, nose and mouth as the key point constraints. We normalize the coordinates of the 3D caricature mesh vertices by translating the center of meshes to the origin and scaling them to the same size. A 3D caricature mesh, denoted as M_i ($1 \leq i \leq N$), can be represented as a long vector containing the coordinates of all the vertices. We apply PCA to all the caricature meshes $\{M_i\}$, and obtain the mean head $\bar{h}$ (which is a $3n_v$-vector containing the coordinates of n_v mesh vertices), and d dominant components α_i ($1 \leq i \leq d$). Then, a new caricature mesh can be represented as

$$h = \bar{h} + \sum_{i=1}^{d} h_i \alpha_i = \bar{h} + \mathbf{H}\boldsymbol{\alpha}, \qquad (1)$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)^T$ is a collection of d components, and $\mathbf{H} = [h_1, h_2, \ldots, h_d]$ is a d-dimensional vector that compactly represents the 3D caricature. The PCA model has only one hyperparameter, i.e, the number of components. Balancing the amount of information against the number of components, we set $d = 200$, where the sum of explained variance ratios is 0.9997, which means it contains almost all the information from the 5,343 meshes. We refer to the above representation as 3D Caricature PCA (3DCariPCA), and use it in our translation network.

In Eq. (1), the representation is a linear combination of principal components and only involves matrix multiplication, which can be efficiently implemented on the GPU. The gradients of these operations are also linear so the losses related to the mesh can be computed and their gradients can be back-propagated. Therefore, the PCA representation is ideal for our neural-network-based method.

3.2 Normal Head Mesh Reconstruction From Photos

In our pipeline, to generate the 3D caricature, we reconstruct a normal 3D head mesh (which has the same connectivity as our caricature meshes) from the input photo. To do so, we first use the

method [32][2] to reconstruct a normal 3D face mesh (i.e., only a front face mesh without ears, neck or the back of the head) from the input photo. Then we use the NICP method [33] to register a template 3D head mesh (which is taken from FaceWarehouse [22]) to the resulting face mesh, and simultaneously use the method [34][3] to register the face PCA to the head PCA, i.e., building a correspondence between the two PCA models. The template head mesh registration is performed as follows: (1) we use facial landmarks as the landmarks for NICP and (2) in the outer loop of NICP, we decrease the stiffness from 50 to 0.2 and decrease the landmark weights from 5 to 0.

The registered head model (including ears, neck and the back of the head) is used as a bridge to transfer the information from the input photo to the output 3D caricature in two parts: (1) we define a perceptual contrast between a normal 3D head mesh and 3D caricature, to measure the character similarity and caricature style (Section 3.3), and (2) we use the texture mapping on the normal 3D head model as the texture mapping on the 3D caricature (Section 3.4).

3.3 Generating Caricature Meshes From Photos

We now describe our network architecture for translation from a 2D photo to a 3D caricature. It is an extreme cross-domain task where the input is a normal face image and the output is an exaggerated 3D mesh whose forms and styles are both totally distinct. We use CelebAMask-HQ dataset [10] which contains 30,000 portrait photos and our 3DCari dataset containing 5,343 full-head 3D caricatures as the training datasets, which are naturally unpaired.

Our network tries to learn the PCA parameters from the photos using a GAN (generative adversarial network) structure, so that it can generate 3D caricature meshes automatically. Denote by $\mathcal{P}$ the domain of photos and $\mathcal{C}$ the domain of PCA representation of 3D caricature meshes. The input to our network is a normal face photo $p \in \mathcal{P}$ and the output is a 3D caricature mesh $c \in \mathcal{C}$, represented by our 3DCariPCA representation, to make learning more efficient and incorporate 3D caricature constraint to improve generation results. Our network consists of a generator G and a discriminator D. G generates a 3DCariPCA representation $G(p)$ from an input photo p, while D discriminates whether an input 3DCariPCA representation is real or synthesized.

Network Architecture. Since our network deals with both 2D images (for which CNNs and residual blocks [35] are effective) and 3D caricature meshes in the 3DCariPCA space as a d-dimensional vector (for which fully connected layers are suitable), both structures are used in our architecture. The network maps a 2D face photo to a 3D caricature mesh. This involves a sequence of downsampling convolutional layers, residual blocks, reshaping and fully connected layers. We use 1D batch normalization for fully connected layers and 2D batch normalization for convolutional layers and residual blocks.

Perception of Caricatures. A good caricature optimally selects and exaggerates the most representative facial regions in an artistic
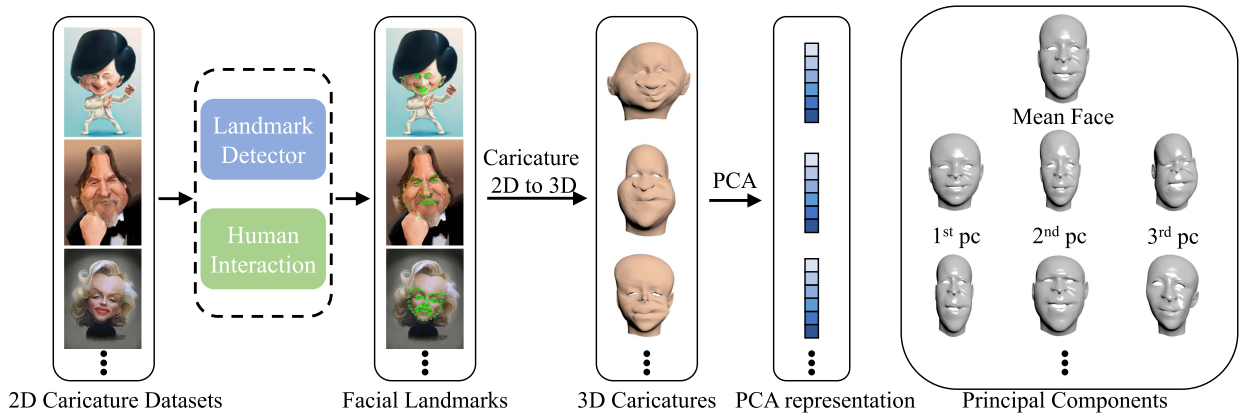
Fig. 5. An illustration of PCA-based 3D caricature representation. To build this representation, we have collected 5,343 3D caricature meshes by detecting facial landmarks on 2D caricature photos (with manual correction if needed) and reconstructing 3D caricature meshes from them. We represent them as a 200-dimensional vector by applying PCA.

way, while humans can still perceive the same identity of the input normal face and the output caricature. To offer a good measure of caricature perception and define good loss terms, we follow the recent neuroscience and psychology studies on caricatures [11], [12], [13]. These studies show that in a face space $\mathcal{F}$, the difference between a face $f \in \mathcal{F}$ and the neutral face (i.e., the mean face in 3D normal face space) represent the identity of f, and we use this difference as the feature vector of f. Then the face identity consistency of two faces in $\mathcal{F}$ can be defined as the cosine of the angle between the feature vectors of them, i.e., the dot product of two normalized feature vectors, we call this the *cosine measure*. A good caricature increases the perceptual contrast while maintaining the face identity. In our application, we define an exaggeration equation:

$$C(p, u) = H_{\mathrm{mean}} + u(H(p) - H_{\mathrm{mean}}), \qquad (2)$$

where H_{mean} is a vector containing coordinates of mean mesh vertices in the normal head space, u is the exaggeration coefficient and $H(p)$ is the reconstructed normal 3D head from the input photo p. Note that (1) $H(p) - H_{\mathrm{mean}}$ is the feature vector of $H(p)$; and (2) H_{mean} is obtained by registering a template head mesh to the mean normal face in BFM model [9], which has the same connectivity as caricature meshes. All the caricature head models $C(p, u)$ have the same face identity as $H(p)$, and the larger the value u (s.t. $u > 1$), the more exaggerated the style is. An example is shown in Fig. 6. Below we use the cosine measure to define two novel loss terms.

Loss Terms. Adversarial loss L_{adv} is useful, however, it is insufficient to ensure that the input and the output are the same character, leading to the generation of random persons. We introduce two novel perceptual losses, i.e., character loss L_{cha} and caricature loss

L_{cari}, to constrain the identity of generated caricatures. All of three loss terms L_{adv}, L_{cha} and L_{cari} are used for training 3D-CariGAN.

Adversarial loss. L_{adv} is the adversarial loss which ensures the distribution of $c' = G(p)$ is the same as that of c. We adapt the adversarial loss of LSGAN [36] as

$$\begin{aligned} L_{\mathrm{adv}}(G, D, \mathcal{P}, \mathcal{C}) &= \mathbb{E}_{p \sim \mathcal{P}}(\|D(G(p))\|_2^2) \\ &\quad + \mathbb{E}_{c \sim \mathcal{C}}(\|1 - D(c)\|_2^2). \end{aligned} \qquad (3)$$

Character loss. We propose L_{cha} which aims to measure character similarity between the input photo and the generated 3D caricature, penalizing the identity change. As the input and output domains are rather different, we first reconstruct the 3D head of the input photo using the method introduced in Section 3.2. Based on the cosine measure, L_{cha} is defined as

$$L_{\mathrm{cha}}(G, \mathcal{P}) = \mathbb{E}_{p \sim \mathcal{P}}[1 - \overline{d_G} \cdot \overline{d_P}], \qquad (4)$$

where $d_G = v(G(p)) - H_{\mathrm{mean}}$ and $d_P = H(p) - H_{\mathrm{mean}}$ are the feature vectors of $G(p)$ and $H(p)$, respectively. $v(G(p))$ is the vector containing the vertices' coordinates of the mesh represented by caricature PCA parameter $G(p)$, where $\bar{v}$ is the normalized vector of v and $\cdot$ is the dot product.

Caricature loss. The cosine of the angle between the two feature vectors measures the face identity consistency of two faces, while the length of a feature vector v measures the caricature style, i.e., the larger the magnitude $\|v\|$ ($\|v\| > 1$), the more exaggerated the style is [11]. Therefore, we propose a novel caricature loss L_{cari} to constrain the caricature style

$$L_{\mathrm{cari}}(G, \mathcal{P}) = \mathbb{E}_{p \sim \mathcal{P}}\left[\exp\left(-(\overline{d_G} \cdot \overline{d_P})\frac{\|d_G\|}{\|d_P\|}\right)\right], \qquad (5)$$

which penalizes identity inconsistency and insufficient exaggerations. The exponential form helps the magnitude to converge to a proper value: to decrease the loss term, $\|d_G\|$ will tend to be large, meanwhile since the exponential coefficient is negative, its gradient will decay exponentially. Furthermore, the adversarial L_{adv} can constrain the magnitude of d_G and make it converge to a proper value.

Overall Loss Function. The overall loss function for 3D-CariGAN is

$$L_{\mathrm{obj}} = L_{\mathrm{adv}} + \lambda_{\mathrm{cha}} L_{\mathrm{cha}} + \lambda_{\mathrm{cari}} L_{\mathrm{cari}},$$

where λ_{cha} and λ_{cari} are the weights for balancing the multiple objectives. For all experiments, we set $\lambda_{\mathrm{cha}} = 2$, $\lambda_{\mathrm{cari}} = 20$. We use the Adam solver to optimize the objective function for training the neural network.
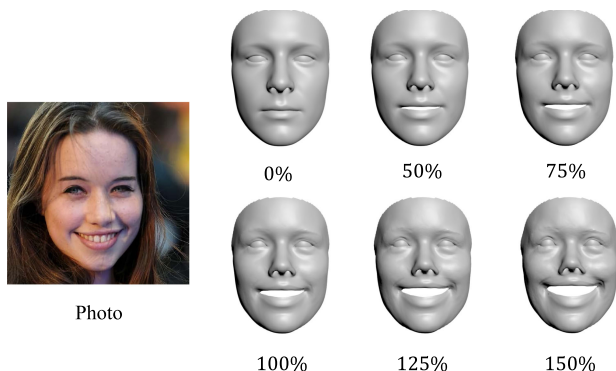

Fig. 6. Exaggeration between a reconstructed face (100%) and the mean face (0%). The larger the exaggeration coefficient (125% and 150%), the more exaggerated the style is.

Fig. 7. The second option provided by our pipeline, which assigns color to each vertex of the full 3D caricature mesh.

Fig. 8. Examples of user control with different combinations of u_1 and u_2 in Eq. (6), i.e., $u_1 = 0.5, 1, 1.5$ and $u_2 = 0.5, 1, 1.5$.

3.4 Caricature Texture Generation

Texture is essential for the appearance of 3D caricatures. In our pipeline, we provide two options so that a user can choose the best for their application.

Based on the normal 3D head $H(p)$ reconstructed from input photo p, we detect common facial landmarks on both $H(p)$ and p. Then we compute a projection matrix according to the correspondence between these two sets of facial landmarks. This projection matrix enables us to build a texture mapping that maps the photo p to the 3D head model $H(p)$. Since we ensure that the 3D caricature mesh $G(p)$ has the same connectivity as $H(p)$, there is a one-to-one correspondence between vertices of $G(p)$ and $H(p)$; therefore, the same texture mapping can be applied to $G(p)$. This texture mapped version of the 3D caricature is our first option. This option has a high texture resolution, but the texture only exists in the (non-occluded) front of the face.

The first option does not have texture on the back of the head. In some applications such as 3D printing, the color on the entire head is needed. To achieve this, in our second option, we store the color for each vertex at the front of the face of the 3D caricature (using the first option), and make up the color for remaining vertices. To do so, we use a smooth interpolation scheme which minimizes the Dirichlet energy by solving a linear system [37]. This interpolation is too smooth and lacks high-frequency texture information. To obtain a more natural texture, we calculate the variance of the front face texture and add random noise of the same variance to the remaining vertices. Two examples are shown in Fig. 7. This second option specifies the color at each vertex, and is therefore of a lower resolution than the first option; however, colors are assigned over the full head.

3.5 Simple User Interaction

A fully automatic method is convenient for users. However, sometimes they may not be satisfied with the results of fully automatic methods. Therefore, allowing users to control the output in a simple way is desired. In our pipeline, given a 3D caricature $G(p)$ (automatically generated by 3D-CariGAN) and a normal 3D head model $H(p)$ (automatically reconstructed from input photo p), we extend the exaggeration equation in Eq. (2) to provide two simple-to-use and intuitive parameters (i.e., u_1 and u_2) for user control

$$C(p, u_1, u_2) = H_{\text{mean}} + u_1 d_G + u_2 d_P, \tag{6}$$

where $d_G = v(G(p)) - H_{\text{mean}}$ and $d_P = H(p) - H_{\text{mean}}$ are feature vectors of $G(p)$ and $H(p)$, respectively. Two parameters u_1 and u_2 have clear geometric meaning: (1) u_2 controls the exaggeration degree of $H(p)$ with the strict face identity; (2) u_1 controls the exaggeration degree of $G(p)$ with diverse exaggeration style; and (3) if $u_1 + u_2 = 1$, the result is an interpolation between $G(p)$ and $H(p)$. An example of user control with different combinations of u_1 and u_2 is shown in Fig. 8.

4 EXPERIMENTS

We have implemented the proposed 3D-CariGAN, CariGeoGAN [5] and normal 3D head reconstruction in PyTorch. We have also implemented 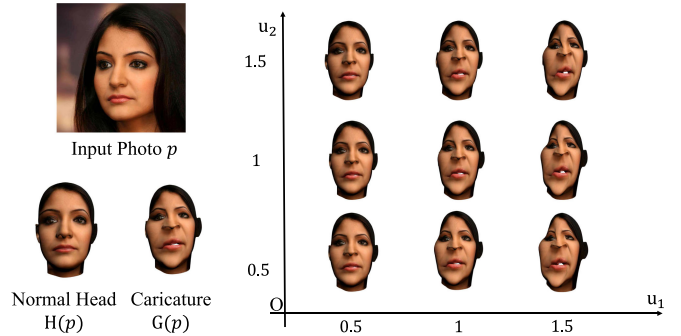the optimization-based method [6] in C++, as a baseline for comparison. We tested them on a PC with an Intel E5-2640v4 CPU (2.40 GHz) and an NVIDIA GeForce RTX 2080Ti GPU. The resolution of face photos in the CelebAMask-HQ dataset is 1024×1024. They are resized to 256×256 as input to 3D-CariGAN.

In addition to the results presented in this section, the dataset details (Appendix A), which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2021.3126659, evaluation of 3DCariPCA (Appendix B), available in the online supplemental material, implementation details (Appendix C, available in the online supplemental material) and more experimental results (Appendix D, available in the online supplemental material) are summarized in the appendix, available in the online supplemental material.

4.1 Ablation Study

We first show the benefits and necessity of using our PCA representation for 3D caricatures. As an alternative, we show the results of a variant of our method that instead uses the mesh vertex coordinates to represent 3D caricatures. Some results are shown in Fig. 3, which are very noisy and visually unacceptable, due to the higher dimensional space and lack of the facial constraint.

We then perform an ablation study to demonstrate the effectiveness of each loss term and some results are shown in Fig. 9. We successively add adversarial loss, character loss and caricature loss to the objective function. The adversarial loss only ensures that the generation results are 3D caricatures, but the results are highly random. The character loss constrains the generation results such that the identity and expression match those of input photos, which make the generation results look like the input photos. The caricature loss strengthens the exaggeration style of the generation results. Character loss and
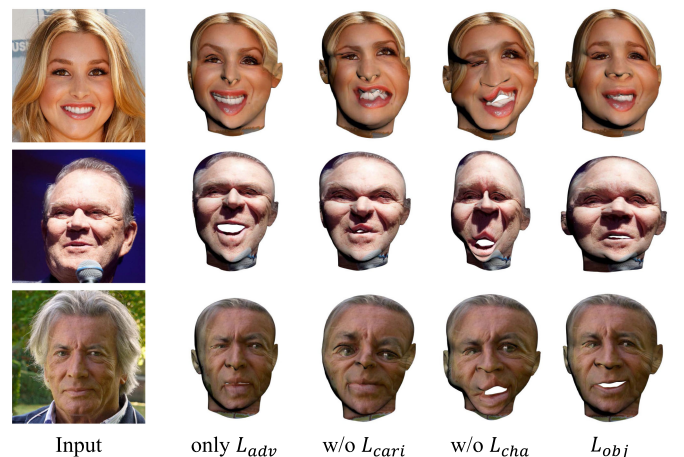

| Input | only L_{adv} | w/o L_{cari} | w/o L_{cha} | L_{obj} |

Fig. 9. Results for ablation study. The results with only L_{adv} are highly random. The results without L_{cari} are not exaggerated properly. The results without L_{cha} have low quality. Only whole method can generate ideal results.
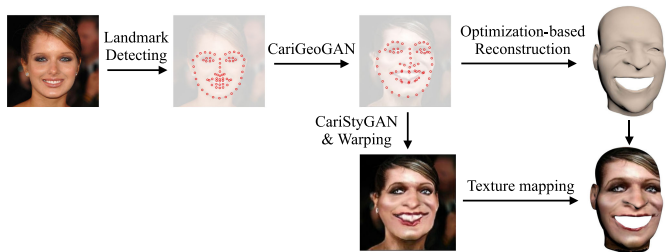
Fig. 10. The baseline method consists of several steps: landmark detection, landmark exaggeration using CariGeoGAN, image warping and generating 3D caricature meshes using facial landmarks.

caricature loss work well together. They ensure the generation results have good exaggeration styles and maintain good face identities. However, using only one of them cannot obtain good results.

4.2 Comparison With Baseline Method

To the best of the authors' knowledge, there is no existing method for generating 3D caricatures directly from photos. We build a baseline method by concatenating three methods. So far the optimization-based method [6] is the only method for generating 3D caricature meshes from 2D caricature images. However, it needs facial landmarks of caricature images as input, which are difficult to obtain automatically. CariGeoGAN [5] is a method that exaggerates the facial landmarks extracted from face photos. We can warp the photo with the guidance of exaggerated landmarks using differentiable spline interpolation [8]. Therefore, one possible way for generating 3D caricature meshes is to use a sequence of steps as shown in Fig. 10. Obviously, the baseline method cannot allow users to adjust generation results interactively, while our method offers a simple tool for users to interactively adjust the generation results. For fair comparison, we use the same texture for the baseline method and our method.

Some comparison results of the proposed 3D-CariGAN and the baseline are shown in Fig. 11. More results are shown in the appendix, available in the online supplemental material. For the baseline method, the exaggerated 2D facial landmarks are not directly designed for 3D caricatures, so the results of the baseline method can be too common or too strange. As a comparison, our method has a good exaggeration effect. In Section 4.3, we further conduct a user study for comparing the baseline and our method, which further demonstrates the advantage of our method. We also summarize the running times of both methods in Table 1, demonstrating that our method is much faster than the baseline.

4.3 User Study

Caricature is a kind of artistic style and so far there is still a lack of suitable objective evaluation methods. Therefore we design a two-level user study to compare our method with other methods.
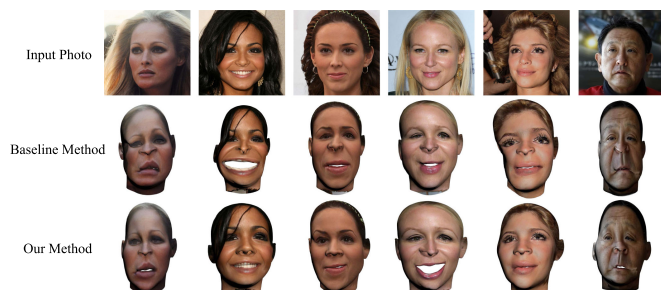

Fig. 11. Visual comparison between our method and baseline. Our user study in Section 4.3 shows that our method has the best performance on RS, Sim and Dist, while baseline has the best performance on Weirdness.

TABLE 1
Running Time Comparison of Baseline and our Method,
Averaged Over 100 Results

Methods	Step	Time (s)
Baseline	Landmark Detection	0.086
	Warping	7.431
	CariGeoGAN	0.001
	Optimization-based reconstruction	14.510
3D-CariGAN	Face Reconstruction (CPU)	1.109
	Other parts (CPU)	0.091
3D-CariGAN	Face Reconstruction (GPU)	0.215
	Other parts (GPU)	0.011

The input is a 256×256 face photo and the output mesh has 12,124 vertices and 24,092 triangles. The running times of 3D-CariGAN on both CPU and GPU platforms are presented.

First we searched the Internet for caricature-related characteristics, and collected eleven terms: Uniform style, Reasonable structure, Clear theme, Color richness, Cultural connotation, Art skill, Creativeness, Similarity, Distinctiveness, Weirdness/Grotesque, and Exaggeration. We invited seven artists and asked them to select 3-5 of the most important characteristics for 3D caricatures. The selection results showed that the following four characteristics — Reasonable structure, Similarity, Distinctiveness and Weirdness — received more than half of the votes, and they characterized different aspects of 3D caricature. Then in the second-level user study, we invited participants to evaluate them by presenting the following criteria and explanations:

- *Reasonable Structure (RS).* The 3D caricature has reasonable structure looking like a human head. Please select the one which is most similar to a human head's structure.
- *Similarity (Sim).* The 3D caricature has the same or similar identity to the input face photo. Please select the one which is most similar to the face photo.
- *Distinctiveness (Dist).* The 3D caricature highlights the most significant part in the input face photo, instead of random deformation. Please select the one that catches the most significant characteristics of the face photo.
- *Weirdness.* The 3D caricature has weird and eerie aesthetic feeling. Please select the one which is most weird but makes you feel novel and beautiful.

Five methods were compared: baseline, three methods in the ablation study (only L_{adv}, w/o L_{cari} and w/o L_{cha}) and our method 3D-CariGAN. 20 participants were recruited to conduct this user study. 10 face photos were used, and each photo has five 3D caricatures corresponding to five methods. For each photo, we randomly selected 5 pairs of 3D caricatures — e.g., (A_p, B_p), (A_p, C_p), (A_p, D_p), (A_p, E_p) and (B_p, C_p) from five caricatures $\{A_p, B_p, C_p, D_p, E_p\}$ corresponding to the photo p — and presented each pair to 10 participants. Finally 100 pairs of 3D caricatures were used and each participant watched 50 pairs. For each pair, the better one was selected for each of the four characteristics. Then we convert the votes into a score using the ranking equation [38]:

$$score_p(I_i) = \sum_{I_j \in \Omega_p \setminus I_i} \frac{s_i - s_j}{s_{\max}}, \qquad (7)$$

where $\Omega_p = \{A_p, B_p, C_p, D_p, E_p\} = \{I_i\}_{i=1}^5$ is a set of 5 caricature results for each photo p, s_i is the number of votes for the result $I_i \in \Omega_p$, and $s_{\max} = 40$ is the maximum votes that each result can get. The range of score in Eq. (7) is $[-4, 4]$. The larger the score, the better quality the 3D caricature has. Finally for each characteristic, the scores were averaged over 10 photos. The averaged scores for four characteristics are summarized in Table 2. These results show that

TABLE 2
The Average Scores of Four Caricature Characteristics in User Study

Method	RS ↑	Sim ↑	Dist ↑	Weirdness ↑
Baseline	0.20	0.23	0.34	**0.53**
only L_{adv}	0.80	0.69	0.43	0.44
w/o L_{cari}	-0.28	-0.29	-0.03	0.11
w/o L_{cha}	-2.39	-2.38	-2.19	-1.49
Ours	**1.66**	**1.75**	**1.45**	0.41

The larger the score, the better the method is.

our method achieves the best quality on RS, Sim and Dist, while the baseline has the best quality on Weirdness. A possible reason is that the baseline explicitly uses 2D caricature information to guide the generation of 3D caricature, without imposing the constraint on face identity to the input photo, so it can generate a better weirdness effect but is worse on the other three characteristics. As a comparison, our method achieves a good balance on all four characteristics.

5 DISCUSSIONS & CONCLUSIONS

Identity and Expression Preservation Ability. We design a quantitative evaluation for identity and expression preservation ability of our method. The details and results are presented in the appendix, available in the online supplemental material.

Limitations. In this paper, we only generate exaggerated geometry and directly use the texture of a photo. It is difficult to generate texture for a 3D caricature because 2D caricature is not a suitable reference. To address this challenge, 3D caricatures with textures created by artists are helpful. We could simultaneously model geometry and texture using neural networks if we had this kind of dataset. Another limitation is that the complete textures is defined as colors at mesh vertices. Generating complete texture in terms of texture mapping needs to be explored in the future.

In this paper, we propose an end-to-end deep neural network model that transforms a normal face photo into a 3D caricature, which is an extreme cross-domain task. To accomplish this task, we build a 3D caricature dataset, establish a PCA model, and propose two novel loss terms based on previous psychological studies. Our method is fast and thus makes further interaction control possible. We also propose a simple and intuitive method that allows a user to interactively adjust the results. Experiments and a user study demonstrate the effectiveness of our method.

REFERENCES

[1] S. B. Sadimon, M. S. Sunar, D. Mohamad, and H. Haron, "Computer generated caricature: A survey," in *Proc. Int. Conf. Cyberworlds*, 2010, pp. 383–390.
[2] S. E. Brennan, "Caricature generator: The dynamic exaggeration of faces by computer," *Leonardo*, vol. 40, no. 4, pp. 392–400, 2007.
[3] X. Han et al., "CaricatureShop: Personalized and photorealistic caricature sketching," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 7, pp. 2349–2361, Jul. 2020.
[4] W. Li, W. Xiong, H. Liao, J. Huo, Y. Gao, and J. Luo, "CariGAN: Caricature generation through weakly paired adversarial learning," 2018, *arXiv:1811.00445*.
[5] K. Cao, J. Liao, and L. Yuan, "CariGANs: Unpaired photo-to-caricature translation," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 244:1–244:14, Dec. 2018.
[6] Q. Wu, J. Zhang, Y.-K. Lai, J. Zheng, and J. Cai, "Alive caricature from 2D to 3D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7336–7345.
[7] X. Han, C. Gao, and Y. Yu, "DeepSketch2Face: A deep learning based sketching system for 3D face and caricature modeling," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, 2017.
[8] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing normalized faces from facial identity features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3386–3395.
[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. ACM SIGGRAPH*, 1999, pp. 187–194.
[10] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," 2019, *arXiv:1907.11922*.
[11] M. Q. Hill et al., "Deep convolutional neural networks in the face of caricature," *Nat. Mach. Intell.*, vol. 1, no. 11, 2019, pp. 522–529.
[12] P. J. Benson and D. I. Perrett, "Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images," *Eur. J. Cogn. Psychol.*, vol. 3, no. 1, pp. 105–135, 1991.
[13] G. Rhodes, G. Byatt, T. Tremewan, and A. Kennedy, "Facial distinctiveness and the power of caricatures," *Perception*, vol. 26, pp. 207–223, 1997.
[14] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.
[15] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–15, 2017.
[16] Y. Shi, D. Deb, and A. K. Jain, "WarpGAN: Automatic caricature generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 762–10 771.
[17] M. Zollhöfer et al., "State of the art on monocular 3D face reconstruction, tracking, and applications," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 523–550, 2018.
[18] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2009, pp. 296–301.
[19] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," *Intl. J. Comp. Vis.*, vol. 126, no. 2–4, pp. 233–254, 2018.
[20] H. Dai, N. Pears, W. A. Smith, and C. Duncan, "A 3D morphable model of craniofacial shape and texture variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3085–3093.
[21] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 426–433, 2005.
[22] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Trans. Vis. Comp. Graph.*, vol. 20, no. 3, pp. 413–425, Mar. 2014.
[23] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1031–1039.
[24] A. Tewari et al., "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1274–1283.
[25] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu, "3D face reconstruction with geometry details from a single image," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4756–4770, Oct. 2018.
[26] M. Sela, Y. Aflalo, and R. Kimmel, "Computational caricaturization of surfaces," *Comput. Vis. Image Understanding*, vol. 141, pp. 1–17, 2015.
[27] L. Clarke, M. Chen, and B. Mora, "Automatic generation of 3D caricatures based on artistic deformation styles," *IEEE Trans. Vis. Comp. Graph.*, vol. 17, no. 6, pp. 808–821, Jun. 2011.
[28] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 704–720.
[29] J. Huo, Y. Gao, Y. Shi, and H. Yin, "Variation robust cross-modal metric learning for caricature recognition," in *Proc. ACM Multimedia Conf.*, 2017, pp. 340–348.
[30] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin, "WebCaricature: A benchmark for caricature recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–12.
[31] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.
[32] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 285–295.
[33] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[34] S. Ploumpis, H. Wang, N. Pears, W. A. Smith, and S. Zafeiriou, "Combining 3D morphable models: A large scale face-and-head model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10934–10943.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[36] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley , "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.
[37] Z. Ye, R. Yi, W. Gong, Y. He, and Y.-J. Liu, "Dirichlet energy of Delaunay meshes and intrinsic Delaunay triangulations," *Comput. Aided Des.*, vol. 126, 2020, Art. no. 102851.
[38] Y. Chen, Y.-J. Liu, and Y.-K. Lai, "Learning to rank retargeted images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3994–4002.

Zipeng Ye received the BEng degree in 2017 from Tsinghua University, China, where he is currently working toward the PhD degree with the Department of Computer Science and Technology. His research interests include computational geometry, computer vision, and computer graphics.

Mengfei Xia received the BS degree in 2020 from Tsinghua University, China, where he is currently working toward the PhD degree with the Department of Computer Science and Technology. His research interests include mathematical foundation in deep learning, image processing and computer vision. He was the recipient of Silver Medal twice in 30th and 31st National Mathematical Olympiad of China.
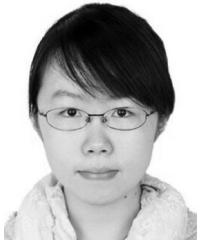
Yanan Sun received the BEng degree in 2020 from Tsinghua University, China, where she is currently working toward the master's degree with Department of Computer Science and Technology, Tsinghua University. Her research interests include computer vision and computer graphics.

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

Minjing Yu received the BE degree from Wuhan University, Wuhan, China, in 2014, and the PhD degree from Tsinghua University, Beijing, China, in 2019. She is currently an assistant professor with the College of Intelligence and Computing, Tianjin University, China. Her research interests include computer graphics, artificial intelligence, and cognitive computation.

Juyong Zhang received the BS degree from the University of Science and Technology of China in 2006, and the PhD degree from Nanyang Technological University, Singapore. He is currently an associate professor with the School of Mathematical Sciences University of Science and Technology of China. His research interests include computer graphics, computer vision, and numerical optimization. He is an associate editor for the *IEEE Transactions on Multimedia and The Visual Computer*.

Yu-Kun Lai received the BS and PhD degrees in computer science from Tsinghua University, in 2003 and 2008 respectively. He is currently a professor with the School of Computer Science and Informatics, Cardiff University, U.K. His research interests include computer graphics, computer vision, geometric modeling, and image processing. For more information, visit https://users.cs.cf.ac.uk/Yukun.Lai/

Yong-Jin Liu (Senior Member, IEEE) received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, in 2004. He is currently a professor with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include computational geometry, computer graphics, and computer vision. For more information, visit https://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.html

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.