Landmark Detection and 3D Face Reconstruction for Caricature using a Nonlinear Parametric Model

Hongrui Cai, Yudong Guo, Zhuang Peng, Juyong Zhang*

***Abstract*—Caricature is an artistic abstraction of the human face by distorting or exaggerating certain facial features, while still retains a likeness with the given face. Due to the large diversity of geometric and texture variations, automatic landmark detection and 3D face reconstruction for caricature is a challenging problem and has rarely been studied before. In this paper, we propose the first automatic method for this task by a novel 3D approach. To this end, we first build a dataset with various styles of 2D caricatures and their corresponding 3D shapes, and then build a parametric model on vertex based deformation space for 3D caricature face. Based on the constructed dataset and the nonlinear parametric model, we propose a neural network based method to regress the 3D face shape and orientation from the input 2D caricature image. Ablation studies and comparison with state-of-the-art methods demonstrate the effectiveness of our algorithm design. Extensive experimental results demonstrate that our method works well for various caricatures. Our constructed dataset, source code and trained model are available at https://github.com/Juyong/CaricatureFace.**

***Index Terms*—Landmark Detection, 3D Face Reconstruction, Caricatures, Nonlinear Representation**

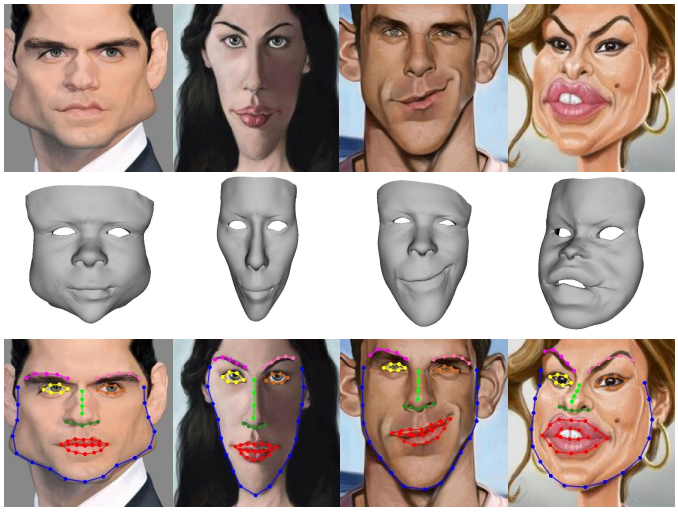

Fig. 1. Some examples of automatic landmark detection and 3D reconstruction on test set. Given a single caricature image (first row), our algorithm generates its 3D model with orientation (second row) and corresponding 68 landmarks (third row).

I. Introduction

As a vivid artistic form that represents human faces in abstract and exaggerated ways, caricature is mainly used to express satire and humor for political or social incidents. It also has many applications in our daily life, such as social network, animation and entertainment industry. Since Brennan developed the first caricature generator in 1985 [1], the studies of caricatures have mainly focused on some specific tasks, such as caricature generation [2], [3], [4], [5], caricature recognition [6], [7], [8], and caricature reconstruction [9], [10], [11], [12]. Most of these tasks need facial landmarks to help to preprocess the caricatures. As a fundamental process for various caricature processing tasks, automatic facial landmark detection and 3D face reconstruction can greatly improve the efficiency and accuracy of other caricature processing tasks. Although the state-of-the-art face alignment methods work well for normal facial images, they are not applicable to caricatures. For example, it still needs to manually refine the landmark positions after applying face alignment methods to caricature images, as reported in [13], [12], [4].

Compared with other tasks like caricature generation [4], [5], [14] and editing [15], there is little research on automatic landmark detection for caricatures. As far as we know, one related work is proposed by Sadimon and Haron [16], which adopted the neural network to predict a facial caricature configuration. However, it can not process a single 2D caricature without its original facial image because the training dataset is constructed by image pairs- one normal facial image and its corresponding caricature image. Besides, their training and testing caricatures are all from exactly one artist, and thus the trained model can not be adapted to other caricatures with different art styles. There exist two main difficulties of facial landmark detection for caricature. One difficulty is that caricatures have abstract and exaggerate patterns, and another is that caricatures have large representation varieties among different artists. As pointed out in [13], compared with landmark detection on normal facial images, it is much more challenging on landmark detection for caricatures.

In comparison to normal facial images, caricatures have two fundamental attributes- exaggeration and variety, and thus approaches for standard landmark detection can not be directly applied to solve this problem. One straightforward way is to regress the 2D landmarks' coordinates of caricature directly. However, 2D landmarks are controlled by facial shape, expression, orientation, and artistic style, which makes it a challenging problem to detect 2D landmarks. In order to alleviate the problem difficulty, we propose to decouple these factors. By regressing the 3D face model and orientation, 2D landmarks can be recovered by projecting the 3D landmarks

*Email: juyong@ustc.edu.cn.

J. Zhang, Y. Guo and Z. Peng are with School of Mathematical Sciences, University of Science and Technology of China.

H. Cai is with School of Data Science, University of Science and Technology of China.

with the orientation.

However, existing parametric 3D face models are mainly designed to represent normal face shapes, and thus they do not work well for caricature faces due to their limited capability of extrapolation. In this paper, to solve this challenging problem, we specifically design a parametric model for 3D caricature faces and propose a method for landmark detection and 3D reconstruction of caricature based on this model. To this end, we manually label landmarks of about 6K caricature images with different styles. We further automatically generate nearly 2K caricatures with labeled landmarks from normal facial images via the method described in [4]. Based on the labeled landmarks, we recover the corresponding 3D caricature shape and orientation using an optimization method. With the large scale training dataset, we propose a novel convolutional neural network based method to regress the 3D caricature shape and orientation from the input 2D caricature. To well represent the 3D exaggerated face, we propose to regress its deformation representation rather than the Euclidean coordinates, which helps to improve the landmark detection and 3D reconstruction ability. In summary, the main contributions of this paper include the following aspects:

- To the best of our knowledge, this is the first work for automatic landmark detection and 3D face reconstruction for general caricatures.
- Rather than directly regress the 2D landmarks, we regress the 3D caricature shape and orientation from input 2D caricature image. 3D caricature shape is represented by a nonlinear parametric model learned from our constructed 3D caricature dataset.

Comparisons with state-of-the-art methods and ablation studies demonstrate the effectiveness of our algorithm pipeline and each module of our proposed method. Extensive qualitative and quantitative experiments demonstrate that our method can automatically produce high accuracy results of 2D landmark detection and 3D shape reconstruction for caricature.

II. Related Work

This section briefly reviews some works related to this paper, with a special focus on face alignment and 3D face reconstruction for normal facial images, and face alignment and 3D face reconstruction for caricatures.

Face Alignment. Face alignment and landmark detection for normal facial images have achieved great success in the last few years with the power of convolution neural networks. Kazemi and Sullivan [18] used an Ensemble of Regression Trees to estimate the facial landmark positions, and their method has been integrated into the Dlib library [19], a modern C++ toolkit containing some machine learning algorithms. Wu *et al.* [20] proposed *vanilla* CNN, which is naturally hierarchical and requires no auxiliary labels beyond landmarks. Kowalski *et al.* [21] developed Deep Alignment Network (DAN), a robust deep neural network architecture that consists of multiple stages. By adopting a coarse-to-fine Ensemble of Regression Trees, Valle *et al.* [22] proposed a real-time facial landmark regression algorithm. Liu *et al.* [23] noticed that the semantic ambiguity degrades the detection performance and addressed this issue by latent variable optimization methods. Dong *et al.* [24] presented an unsupervised approach to improving facial landmark detectors, and Honari *et al.* [25] showed a new architecture and training procedure for semi-supervised landmark localization. To solve the occlusion problem, Zhu *et al.* [26] developed an occlusion-adaptive deep network, which contains a geometry-aware module, a distillation module, and a low-rank learning module. Merget *et al.* [27] proposed a novel network architecture that has an implicit kernel convolution between a local-context subnet and a global-context subnet composed of dilated convolutions.

3D Face Reconstruction from A Single Image. 3D face reconstruction algorithms can be divided into different categories on the ground of the input modality: single RGB image based [28], video based [29], depth image based [30], and so on. 3D face reconstruction from a single image is to recover 3D facial geometry from a given facial image, which has applications like face recognition [31], [32], face alignment [33], [34] and expression transfer [35], [36]. Since Blanz and Vetter proposed a 3D Morphable Model (3DMM) in 1999 [37], model-based methods have become popular in solving problems of 3D face reconstruction. Earlier, a large number of model-based algorithms considered some significant facial parts between 2D images and 3D templates, such as facial landmarks [38], [39], [40], [36], [41], latent representation [42] and so on. Cao *et al.* [43] utilized some RGBD sensors to create an extensive face database named FaceWareHouse, which contains 150 identities and 47 expressions of each identity. In recent years, deep learning based methods have shown promising results in terms of computation time, robustness to occlusions, and reconstruction accuracy [44]. Guo *et al.* [28] proposed a real-time dense face reconstruction method by constructing a large scale dataset augmented based on traditional optimization methods and adopting a coarse-to-fine CNN framework. During the same year, Tran *et al.* [45] demonstrated a nonlinear 3DMM, which is learned from a large set of unconstrained face images without collecting 3D face scans. Gecer *et al.* [46] harnessed Generative Adversarial Networks (GANs) for reconstructing facial texture and shape from single images by training a generator of facial texture in UV space. Feng *et al.* [47] presented a model-free method to rebuild the 3D facial geometry from a single light field image with a densely connected network. However, due to the diversity of style and geometry of caricatures, the approaches for normal face reconstruction can not be directly applied to general caricatures.

Face Alignment and Reconstruction of Caricature. Compared with researches on normal facial images, there are fewer works about caricatures [48], [49]. For face reconstruction, existing methods mainly focus on constructing a 3D caricature model from a normal 3D face model. Lewiner *et al.* [9] introduced a caricature tool that interactively emphasizes the differences between two 3D meshes by utilizing the manifold harmonic basis of a shape to control the deformation and scales intrinsically. Vieira *et al.* [10] proposed a method based on deformations by manipulation of moving spherical influence zones. Sela *et al.* [50] presented a framework to scale the gradient fields of the surface coordinates by a function of the

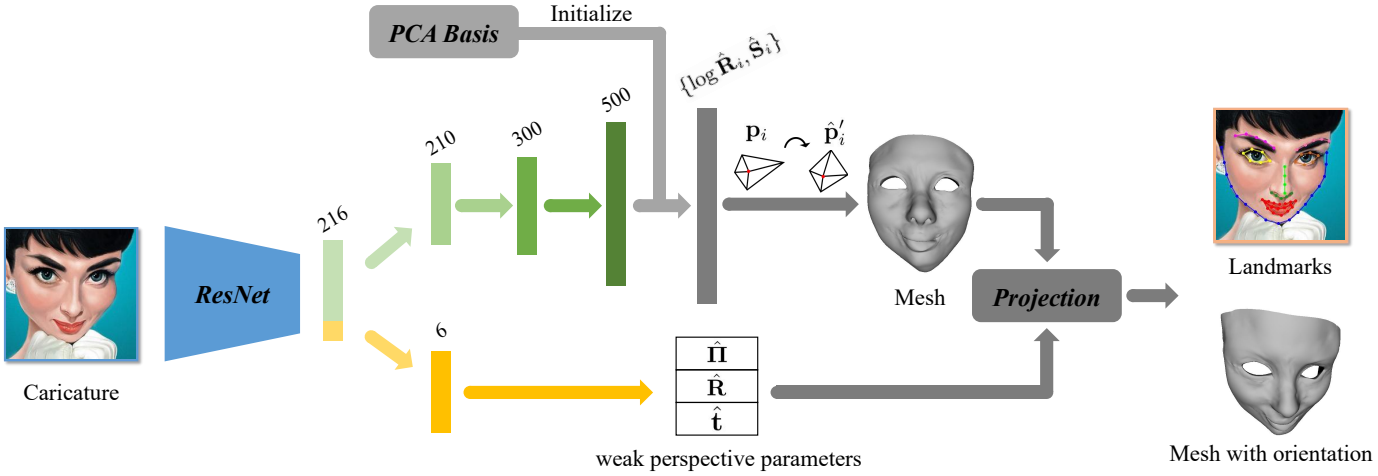

Fig. 2. Overview of our proposed Framework for Landmark Detection and 3D Reconstruction on general caricatures. Our network includes two parts, an encoder and a decoder. We use ResNet-34 [17] backbone as the encoder and 3 Fully Connected (FC) layers as the decoder to recover the 3D caricature shape. The PCA basis of vertex based deformation presentation $\{\log \mathbf{R}_i, \mathbf{S}_i\}$ is used to initialize the last FC layer.

Gaussian curvature of the surface and solve a corresponding Poisson equation to find the exaggerated shape. Besides, there are some works on modeling 3D caricatures from images. Liu *et al.* [51] chose a semi-supervised manifold regularization(MR) method to learn a regressive model between 2D normal faces and enlarged 3D caricatures. With the power of deep learning, Han *et al.* [11] developed a CNN based sketching system that allows users to draw freehand imprecise yet expressive 2D lines representing the contours of facial features. With an intrinsic deformation representation that enables considerable face exaggeration, Wu *et al.* [12] introduced an optimization framework to address this issue. However, [12] needs labeled landmarks as input, which are not easy to get and always need manually labeling. Different from [12], we propose a learning based approach to automatically detect landmarks and regress 3D shape from the input 2D caricature. Furthermore, our method constructs a nonlinear parametric model based on deformation representation, which greatly improves the reconstruction accuracy. As [12] is the state-of-the-art caricature reconstruction method, we adopt it to construct the 3D caricature shape set. Landmark detection on caricature images is also a fundamental problem of caricature perception, but there exist few works on this topic. As a related research direction, manga images have aroused Stricker *et al.*'s [52] interest. Based on DAN [21] framework, they proposed a new landmark annotation model for manga images and a deep learning approach to detect them. Huo *et al.* [13] shows that caricature landmark detection is of great interest, but researches on this topic are still far from saturated. Besides, most studies on caricature generation need facial landmarks as control points [2], [4], [5], which demonstrate that facial landmarks play an essential role in caricature related researches.

III. Algorithm

Given a 2D caricature, we aim to automatically reconstruct its 3D face shape and obtain landmarks around its eyes, nose, mouse, and so on, as shown in Fig. 1. To this end, we construct a 2D caricature dataset with around 8K 2D images and their corresponding labeled landmarks. The dataset contains both artists-designed caricatures and machine-generated caricatures. With their corresponding 68 landmarks, we build a 3D caricature dataset via an optimization based method [12]. Then, based on a deformation representation, we propose an encoder-decoder framework to directly recover the 3D face shape and weak perspective parameters from the input 2D caricature image. Notably, we use the principal component analysis (PCA) basis to initialize the weight of the last fully connected layer. The algorithm pipeline is shown in Fig. 2. In the following, we give the algorithm details for each component.

A. Dataset Construction and Augmentation

Currently, there exist some public available caricature datasets. For the study of caricature recognition, Huo *et al.* [13] constructed a WebCaricature database including 6042 caricatures and 5974 photographs from 252 persons with 17 labeled facial landmarks for each image. Mishra *et al.* [53] built IIIT-CFW database for face classification and caricature generation, which contains 8928 cartoon faces of 100 public figures with annotation of various attributes, e.g., face bounding box, age group, facial expression, and so on. However, these datasets can not be directly used for our task as they do not supply enough labeled landmarks for 3D reconstruction.

By searching and selecting nearly 6K various caricatures from different artists on the Internet, we construct a caricature dataset in which each caricature has 68 labeled landmarks. The landmark positions are initialized via the Dlib library [19], and then manually refined. To further increase the diversity of our dataset, we design a data augmentation method based on CariGANs [4]. CariGANs is able to translate normal facial images to caricatures with two generative adversarial networks (GANs), namely CariGeoGAN and CariStyGAN.

CariGeoGAN learns a mapping to exaggerate the shape by adjusting facial landmarks, while CariStyGAN learns another mapping to translate the appearance from normal facial image style to caricature style. With trained CariGANs, we can generate a caricature and its corresponding 68 landmarks from a given normal facial image. In this way, we generate around 2K caricatures and add them to our dataset. Some examples of our collected data and augmented data are shown in Fig. 3.

For each 2D caricature, based on the labeled 68 facial landmarks, we adopt an optimization based method [12] to recover its 3D shape. In this way, we build a caricature dataset containing around 8K 2D labeled images and their corresponding 3D shapes. Notably, this dataset contains different geometry shapes and image styles since the 2D caricatures are drawn by different artists or generated by the data augmentation method.

B. Deformation Representation of 3D Caricatures

Some statistical parametric models, like 3DMM [37] and FaceWareHouse [43], are popularly used in 3D face reconstruction to represent a complex face shape with a low dimensional parametric vector. This kind of representation makes optimization and learning based 3D face reconstruction easier. However, linear parametric models are only good for interpolation in the shape space of 3D normal faces but do not work well for extrapolation in 3D caricature shape space. Therefore, to recover exaggerated 3D shapes of various caricatures, we adopt a vertex based deformation representation. Compared with 3D Euclidean coordinates, this deformation representation is suitable to represent local and large deformation in a natural way, which makes the reconstructed exaggerated meshes more natural and match the input 2D caricature quite well.

To make our paper self-contained, we first introduce the deformation representation between two meshes with the same topology.

Suppose there are two meshes with the same topology, which means the number of vertices, the order of vertices, and the connectivity of vertices of them are all identical. We treat one mesh as a reference mesh and another as a target deformed mesh. We denote the position of the i^{th} vertex v_i on the reference as $\mathbf{p}_i$ and the i^{th} vertex v_i on the target as $\mathbf{p}'_i$. We can define the deformation matrix in the one-ring neighborhood of v_i from the reference to the target as an affine transformation matrix $\mathbf{T}_i$ by minimizing

$$\min_{\mathbf{T}_i} \sum_{j \in \mathcal{N}_i} c_{ij} \|(\mathbf{p}'_i - \mathbf{p}'_j) - \mathbf{T}_i(\mathbf{p}_i - \mathbf{p}_j)\|_2^2, \tag{1}$$

where $\mathcal{N}_i$ is the neighborhood index set of v_i, and c_{ij} is the cotangent weight [54] to avoid discretization bias in deformation. With polar decomposition, the deformation matrix $\mathbf{T}_i$ can be decomposed into a rigid component represented by a rotation matrix $\mathbf{R}_i$ and a non-rigid component represented by a real symmetry matrix $\mathbf{S}_i$, as $\mathbf{T}_i = \mathbf{R}_i\mathbf{S}_i$.

To obtain efficient linear combination, we use the axis-angle representation [55] to replace the rotation matrix $\mathbf{R}_i$. Following Rodrigues' rotation formula, for the i^{th} vertex v_i, we denote the cross-product matrix and rotation angle by $\mathbf{K}_i$, θ_i. We can convert $\mathbf{R}_i$ to a matrix logarithm notation:

$$\log \mathbf{R}_i = \theta_i \mathbf{K}_i, \tag{2}$$

$$\mathbf{K}_i = \begin{bmatrix} 0 & -k_{i,z} & k_{i,y} \\ k_{i,z} & 0 & -k_{i,x} \\ -k_{i,y} & k_{i,x} & 0 \end{bmatrix}, \tag{3}$$

where $\mathbf{k}_i = (k_{i,x}, k_{i,y}, k_{i,z}) \in \mathbb{R}^3$ and $\|\mathbf{k}_i\|_2 = 1$. Then, the logarithm rotation matrix $\log \mathbf{R}_i$ can be represented by a vector $\mathbf{r}_i = \theta_i \mathbf{k}_i \in \mathbb{R}^3$ and the scalar matrix $\mathbf{S}_i$ can be represented by a vector $\mathbf{s}_i \in \mathbb{R}^6$. To handle the ambiguity of axis-angle representation, Gao *et al.* [56] propose an integer programming approach to make all $\mathbf{r}_i$ as consistent as possible globally. In this paper, we define $[\mathbf{r}_i, \mathbf{s}_i] \in \mathbb{R}^9$ as the deformation representation/gradient of the i^{th} vertex v_i on a target mesh, correspondingly, $\{\log \mathbf{R}_i, \mathbf{S}_i\}$ is its matrix form.

The deformation representation has many advantages, especially for our method, it can be used for linear combination [57] of two rotation matrices $\mathbf{R}_i^0$ and $\mathbf{R}_i^1$ by $\exp(\log \mathbf{R}_i^0 + \log \mathbf{R}_i^1)$. To build a deformation space, we choose a reference mesh and n deformed meshes which have the same topology with the reference model. For the i^{th} vertex of the l^{th} deformed mesh, we obtain its deformation representation $\{\log \mathbf{R}_i^l, \mathbf{S}_i^l\} (l = 1, 2, \ldots, n)$. Then, corresponding to an essential deformation representation, a target mesh can be approximately reconstructed by a linear combination of several known deformation gradients. In detail, based on a reference mesh and n deformed meshes, we build a linear combination of deformation gradients for the i^{th} vertex v_i as

$$\mathbf{T}_i(\mathbf{w}) = \exp(\sum_{l=1}^{n} w_{R,l} \log \mathbf{R}_i^l)(\mathbf{I} + \sum_{l=1}^{n} w_{S,l}(\mathbf{S}_i^l - \mathbf{I})), \tag{4}$$

where $\mathbf{w} = (\mathbf{w}_R, \mathbf{w}_S)$ is the combination weight vector, consisting of weights of rotation $\mathbf{w}_R = \{w_{R,l} | l = 1, \cdots, n\}$ and weights of scaling/shear $\mathbf{w}_S = \{w_{S,l} | l = 1, \cdots, n\}$.

Given a target mesh, we can calculate its optimal weight $\mathbf{w}$ by minimizing the following energy:

$$\min_{\mathbf{w}} \sum_{v_i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} c_{ij} \|(\mathbf{p}'_i - \mathbf{p}'_j) - \mathbf{T}_i(\mathbf{w})(\mathbf{p}_i - \mathbf{p}_j)\|^2, \tag{5}$$

Fig. 3. The first row shows some examples of our collected images with manually labeled landmarks, while the second row shows some examples of our augmented images and corresponding landmarks generated by [4].

where the definitions of $\mathcal{N}_i$ and c_{ij} are same as Eq. (1), $\mathcal{V}$ is the vertex set of the mesh. Since Eq. (5) is a non-linear least squares problem, we first compute the Jacobian matrix $\partial \mathbf{T}_i(\mathbf{w})/\partial \mathbf{w}$, and then use the Levenberg-Marquardt algorithm [58] to solve it.

For each 3D caricature from the constructed dataset in Sec. III-A, we receive its corresponding optimal weight by solving Eq. (5). Then we calculate the deformation representation of its every vertex via Eq. (4). In our experiments, we select the reference mesh and n deformed meshes from FaceWareHouse dataset [43] ($n = 99$). In detail, we choose the mean face as the reference mesh, 24 expressions and 75 identities with the neutral expression, which have large differences to the mean face, as the deformed meshes.

As the deformation representation for each vertex $[\mathbf{r}_i, \mathbf{s}_i]$ contains 9 variables, the deformation representation of a whole 3D caricature mesh with n_v vertices can be represented as a $9n_v$ vector $\{[\mathbf{r}_i, \mathbf{s}_i], i = 1, \ldots, n_v\}$. Based on the constructed 3D caricature dataset, we build a deformation space for 3D caricatures, where each 3D caricature is represented as a deformation form. Therefore, we formulate automatic caricature reconstruction as a geometric deformation problem, where the deformation representation of a 3D caricature is learned from a data-driven approach. In detail, given a 2D caricature, we aim to train an encoder-decoder framework that ends with several fully connected layers to regress its corresponding $9n_v$ deformation representation vector directly. Owing to the representation's robust expression ability, the translation from 2D caricature domain to 3D deformation field is quite natural.

C. Landmark Detection and 3D Reconstruction

Although the deformation representation of 3D caricatures is well constructed, the large number of variables (each representation is a $9n_v$ vector) makes it hard for a convolutional neural network to regress the vector directly. To reduce the prediction difficulty of the network regressing the 3D shape, we make dimensionality reduction based on deformation representation. A similar approach is adopted in [59], which constructs the sparse localized basis of a triangle based deformation representation. However, different from [59], we aim to estimate the deformation from the reference face to an arbitrary caricatured form, which need to capture the global shape deformation. Taking this into consideration, we adopt PCA model to assist network learning. Specifically, we propose an encoder-decoder framework to recover the 3D face shape and weak perspective parameters from the input 2D caricature image. We utilize the constructed PCA basis to initialize the last fully connected (FC) layer's weight. Based on the learnable statistical model, we propose a fully data-driven approach to this problem.

In detail, we propose a CNN-based approach to directly regress the intrinsic deformation representation and the weak perspective projection parameters with a single 2D caricature image. As shown in Fig. 2, we utilize ResNet-34 backbone [17] to encode the input 2D caricature into a latent vector $\boldsymbol{\chi} \in \mathbb{R}^{216}$. The latent vector contains two parts, where $\boldsymbol{\chi}_s \in \mathbb{R}^{210}$ resolves the 3D shape and $\boldsymbol{\chi}_p = (\hat{\mathbf{s}}, \hat{\mathbf{R}}, \hat{\mathbf{t}}) \in \mathbb{R}^6$ represents the parameters of weak perspective projection, where the meanings of $\hat{\mathbf{s}}$, $\hat{\mathbf{R}}$, $\hat{\mathbf{t}}$ will be discussed later. We construct a decoder composed of 3 fully connected layers to convert $\boldsymbol{\chi}_s$ to the estimated latent deformation representation $\{[\hat{\mathbf{r}}_i, \hat{\mathbf{s}}_i], i = 1, \ldots, n_v\}$, where n_v is the number of mesh vertices. The deformation gradients $\{(\log \hat{\mathbf{R}}_i, \hat{\mathbf{S}}_i), i = 1, \ldots, n_v\}$ and the deformation matrixes $\{\hat{\mathbf{T}}_i, i = 1, \ldots, n_v\}$ then can be recovered according to the derivation process in Sec. III-B. To help the model training, we use the first 500 principal components of a PCA basis extracted from the training dataset to initialize the weight of the last fully connected (FC) layer.

Loss for Caricature Shape. As before, the estimated vertex coordinate $\{\hat{\mathbf{p}}'_i\}$ of target mesh can be obtained by solving

$$\arg\min_{\{\hat{\mathbf{p}}'_i\}} \sum_{j \in \mathcal{N}_i} c_{ij} \|(\hat{\mathbf{p}}'_i - \hat{\mathbf{p}}'_j) - \hat{\mathbf{T}}_i(\mathbf{p}_i - \mathbf{p}_j)\|_2^2, \tag{6}$$

which is equivalent to solve the following linear system:

$$2 \sum_{j \in \mathcal{N}_i} c_{ij}(\hat{\mathbf{p}}'_i - \hat{\mathbf{p}}'_j) = \sum_{j \in \mathcal{N}_i} c_{ij}(\hat{\mathbf{T}}_i + \hat{\mathbf{T}}_j)(\mathbf{p}_i - \mathbf{p}_j). \tag{7}$$

As the deformation representation is translation independent, and thus we need to specify the position of mesh center or exactly one vertex. As the ground truth 3D caricature meshes are under the same specification, we construct a loss term to constrain the coordinate difference between the reconstructed mesh and the ground truth mesh as

$$\mathbf{E}_{ver}(\boldsymbol{\chi}_s) = \sum_{v_i \in \mathcal{V}} \|\hat{\mathbf{p}}'_i - \mathbf{p}'_i\|_2^2, \tag{8}$$

where $\mathbf{p}'_i$ presents the ground truth coordinate of the i^{th} vertex of the corresponding 3D mesh from the dataset, and $\mathcal{V}$ represents the vertex set.

Loss for Landmarks. Reconstructing the 3D mesh from a 2D image is an inverse process of observing a 3D object by projecting it to 2D visual space. As before, we assume that the projection plane is the z-plane and thus the scaled projection matrix can be written as $\boldsymbol{\Pi} = s\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, where $\mathbf{s}$ is the scale factor. To better recover the landmark positions, we construct a landmark loss term to measure the difference between the projected landmarks and the ground truth landmarks:

$$\mathbf{E}_{lan}(\boldsymbol{\chi}_s, \boldsymbol{\chi}_p) = \sum_{v_i \in \mathcal{L}'} \|\hat{\boldsymbol{\Pi}}\hat{\mathbf{R}}\hat{\mathbf{p}}'_i + \hat{\mathbf{t}} - \mathbf{q}'_i\|_2^2, \tag{9}$$

where $\mathcal{L}'$ and $\mathcal{Q}' = \{\mathbf{q}'_i, v_i \in \mathcal{L}'\}$ are the set of 3D landmarks and 2D landmarks separately, $\hat{\boldsymbol{\Pi}}$ is the estimated scaled projection matrix, $\hat{\mathbf{R}}$ is the estimated rotation matrix, and $\hat{\mathbf{t}}$ is the estimated translation vector. As our 3D caricature meshes have the same connectivities, the indices of 3D landmarks are the same for different caricature shapes.

Compared with normal face, the positions of caricature silhouette landmarks have large variance, and thus it is quite challenging to detect their positions accurately. Moreover, the 3D vertices corresponding with these silhouette landmarks

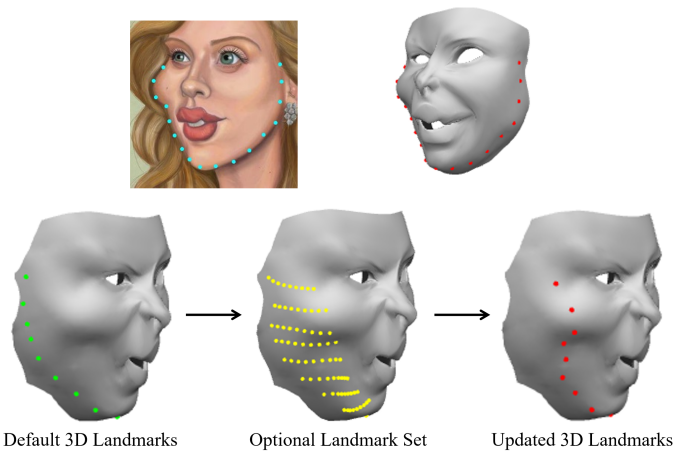

Fig. 4. For non-frontal face caricatures, we need to update the indices of silhouette landmarks on the 3D face shape to better match the corresponding 2D landmarks (shown in cyan in upper-left). The default 3D silhouette landmarks are shown in green in the lower-left. We construct an optional landmark set from each horizontal line (shown in yellow in lower-middle) that has a vertex lying on the silhouette and select among them a set of the updated silhouette landmarks according to the estimated rotation matrix $\hat{\mathbf{R}}$ in each training time. The vertices of the silhouette are updated in the end, as shown in red on the upper right and lower-right.

are labeled on the mean neutral face with a frontal view, which causes the problem that the correspondences between 3D vertices and 2D landmarks are not correct for non-frontal faces as shown in Fig. 4. To solve this problem, we update the indices of 3D silhouette landmarks during training according to the estimated rotation matrix and vertices' coordinates. In each training iteration, we select some vertices from each horizontal line that has a vertex lying on the silhouette and project them onto the image plane according to the estimated rotation matrix $\hat{\mathbf{R}}$. Then for each 2D silhouette landmark, we set the vertex whose projection is closet to it (see Fig. 4) as its current corresponding 3D silhouette landmark.

The total loss function is given in the following form:

$$\mathbf{E} = \lambda_1 \mathbf{E}_{ver} + \lambda_2 \mathbf{E}_{lan}, \tag{10}$$

where λ_1, λ_2 are hyperparameters and their setting will be discussed in the experiment section.

IV. Experiments

In this section, we give the implementation details, ablation studies, qualitative and quantitative evaluation of our proposed method, as well as comparisons with several related methods.

Implementation Details We train our model via the PyTorch [60] framework. CNN takes the input of a color caricature image with size $224 \times 224 \times 3$. We use Adam solver [61] with the mini-batch size of 32 and train the model with 2K iterations. The base learning rate is set to 0.0001. We set $\lambda_1 = 1, \lambda_2 = 0.00001$ during the first 1K iterations, and set $\lambda_1 = 1, \lambda_2 = 0.001$ during the last 1K iterations. The reason why the magnitudes of parameters are quite different is that the magnitude of vertices' coordinates has a big difference from that of 2D pixels.

All the tests, including our method and comparison methods, were conducted on a desktop PC with a hexa-core Intel CPU i7 at 3.40 GHz, 16GB of RAM, and NVIDIA TITAN Xp GPU. As for the running time for each caricature, our method takes about 10ms to obtain both 3D mesh and 68 2D landmarks. The number of vertices of our reconstructed mesh is 6144.

A. Ablation Study

We first conduct ablation studies to demonstrate the importance of each component. The ablation studies are designed for the augmented data, PCA initialization, and silhouette updating strategy.

As Tab. I shown, we evaluate the detection performance with several commonly used landmark error metrics, which are also used in [21]. Specifically, the second row shows the error metrics of our method without using augmented data, which are generated by CariGANs [4] in Sec. III-A. The third row shows the errors of our method without PCA initialization for the weight of the last FC layer mentioned in Sec. III-C. And the fourth row shows the errors of our method without adopting the strategy to update the indices of silhouette landmarks displayed in Fig. 4. As demonstrated in Tab. I, the mean error of estimated landmarks decreases from 5.85 to 5.64 with the help of augmented data, from 6.91 to 5.64 thanks to the PCA initialization, and from 5.99 to 5.64 owing to the silhouette updating strategy.

Fig. 5 shows the detection and reconstruction results of ablation studies. The reconstructed mesh of our method without using augmented data shows that the learned model does not show good generalization ability, which leads to misalignment in the detection of silhouette points. The recovered mesh of our method without PCA initialization demonstrates that the PCA model makes the results to be smooth and more natural. We can observe that the predicted silhouette landmarks by the model without silhouette updating strategy deviate from the accurate position, which confirms the effectiveness of the silhouette updating strategy.

TABLE I
Results of the ablation studies with metrics of landmark detection errors. Values of mean error with normalization are shown as the percentage of the normalization metric.

	mean error	inter -pupil	inter -ocular	diagonal
w/o Augmented	5.85	9.29	6.34	2.38
w/o PCA	6.91	11.01	7.52	2.82
w/o Sil. Update	5.99	9.49	6.48	2.44
Ours	**5.64**	**8.93**	**6.10**	**2.30**

B. Detection Comparison

As far as we know, there is no existing method for landmark detection for general caricatures. We compare our method with some benchmark methods. The first type is the face alignment methods, which are designed for normal human faces, and we select three typical methods, including DAN [21], ERT [18], and vanilla CNN (VCNN) designed by [20]. As their released

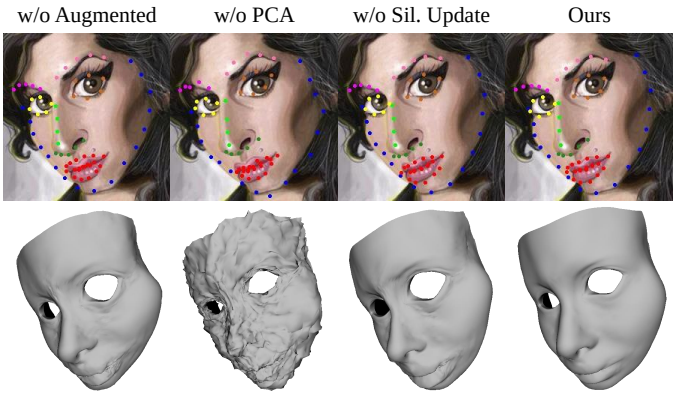

Fig. 5. Landmark detection and reconstruction results of the ablation studies. Left to right: results by method without using augmented data, results by method without PCA initialization, results by method without silhouette updating strategy, and results by our full method.

TABLE II
STATISTICS OF LANDMARK DETECTION ERRORS AND COMPUTATION TIME (MS/IMAGE) ON THE TEST SET. VALUES OF MEAN ERROR WITH NORMALIZATION ARE SHOWN AS THE PERCENTAGE OF THE NORMALIZATION METRIC.

	mean error	inter -pupil	inter -ocular	diagonal	time (ms)
DAN	5.78	9.93	6.80	2.59	25.9
ERT	8.24	14.52	9.95	3.71	2.7
VCNN	14.04	24.33	16.67	6.39	**1.6**
L-PCA	5.87	10.08	6.91	2.64	4.8
V-PCA	6.20	10.68	7.32	2.79	6.4
DR-PCA	5.75	9.89	6.77	2.58	9.3
Ours	**4.98**	**8.51**	**5.82**	**2.23**	9.8

trained models are trained with normal facial images, we retrain their models based on the author's training code. For a fair comparison, their methods are trained and tested with the same training and testing dataset as our method. We randomly split our dataset into 80% for training and 20% for testing.

DAN: Deep Alignment Network (DAN) [21] is a robust face alignment method based on deep neural network architecture. Its algorithm pipeline includes multiple stages, where each stage improves the locations of the facial landmarks estimated by the previous stage.

ERT: In [18], an ensemble of regression trees (ERT) has been used to directly estimate the facial landmark positions from a sparse subset of pixel intensities. This method achieves super-realtime performance with high-quality predictions. It has been integrated into the Dlib library [19].

VCNN: Vanilla CNN is proposed in [20], which introduces hierarchical and discriminative processing to the existing CNN design for facial landmark regression.

Except for the above three methods, we also implement some baseline methods.

L-PCA: Inspired by [4], we extract the PCA basis of 2D caricature landmarks from the labeled landmark dataset. In this way, the landmarks of caricature image can be represented by the coefficient of PCA basis. We use the same ResNet framework in our method to directly regress the coefficient.

V-PCA: We extract the PCA basis of 3D caricature shape set represented by the Euclidean coordinates. The network structure is the same as our algorithm pipeline in Fig. 2, and regresses the PCA coefficient and orientation.

DR-PCA: We extract the first 210 principal components of a PCA basis from the 3D caricature shape set represented by the deformation representation. The pipeline is the same as our method by changing the decoder (the last 3 FC layers) to the matrix multiplication with the extracted PCA basis. Specifically, the deformation gradients $\{(\log \hat{\mathbf{R}}_i, \hat{\mathbf{S}}_i), i = 1, \ldots, n_v\}$ can be directly computed via matrix multiplication between the extracted PCA basis and the latent vector $\boldsymbol{\chi}_s \in \mathbb{R}^{210}$.

We compare our method with these benchmark methods. Fig. 6 shows some visual results of landmark detection. It can be observed that the detected landmarks of ERT [18] and VCNN [20] can not match the face shape. The method of DAN [21] performs quite well for the facial feature parts, including eyes, nose, and mouth. However, its silhouette landmarks may deviate from the accurate positions. V-PCA and L-PCA are also not good for the landmarks on the silhouette. Though DR-PCA representation shows nice performance, it still can not match the facial feature parts precisely. In contrast, the detection results by our method are quite close to the ground truth landmarks, even for the silhouette landmarks. We also quantitatively compare our method with these methods on several frequently used landmark error metrics and average computation time. We show the statistics in Tab. II and the cumulative errors distribution (CED) curves of these methods on the mean error in Fig. 7. We can see that the mean error, mean error normalized separately by inter-pupil distance, inter-ocular distance, and bounding box diagonal of our methods are all smaller than those of other methods.

The reason why our method performs better includes the following aspects. First, rather than directly regressing the 2D landmarks, we regress the 3D shape and orientation. In this way, a challenging problem is decomposed into two easier problems. Second, to better represent the 3D caricature shape, we learn a nonlinear parametric model, which is more suitable to represent the 3D caricature shape than 3D morphable model [37] and FaceWareHouse [43].

C. 3D Reconstruction Comparison

Reconstructing 3D caricature shape from caricature image is also a challenging problem. In Fig. 8, we show eight reconstruction examples from the test set. The reconstructed mesh is overplayed on the image, and we can observe that the shape is recovered quite well. The recovered mesh from two different views and with texture are also shown to demonstrate the effectiveness of our method.

We also compare our method with an existing state-of-the-art method [12], which is the only universal method of 3D caricature reconstruction. Then as shown in their paper, classical parametric models like 3DMM [62], [63] and FaceWareHouse [43] cannot reconstruct exaggerated meshes well due to their limited extrapolation ability. As Fig. 9 shows, compared to the results of [12], the reconstructed 3D meshes by our method are quite natural and vivid. There are two advantages of our method. One is the computation time. It

Fig. 6. We provide visual landmark detection results on the test dataset using DAN [21], ERT [18], VCNN [20] and some baselines, including Landmark PCA (L-PCA), Vertex PCA (V-PCA), and DR-PCA.

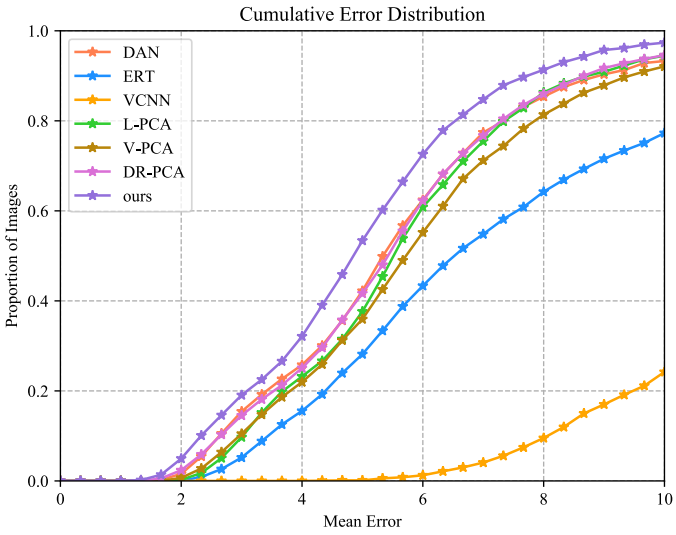

Fig. 7. Comparisons of cumulative errors distribution (CED) curves on the test set.

takes around 10ms (real-time) to produce the result with our method, while 12.5s for their method. Another difference is that our reconstruction method does not need to label the landmarks manually, while [12] needs labeled landmarks as input.

Moreover, we also compare our method with the reconstruction methods by 3DMM [63], FaceWareHouse [43], [38] and Alive [12]. We compare the mean square error between the projected landmarks and ground-truth landmarks, which also has been used in [12]. For these optimization based methods, the 3D caricature mesh is reconstructed by minimizing the residuals between the projected landmarks and the ground-truth landmarks. It needs to be pointed out that the compared methods all need labeled landmarks input, while our method is automatic. As shown in Tab. III, our method even outperforms both 3DMM and FaceWareHouse fitting over test data, although our method does not have the ground truth 2D caricature landmarks as input.

TABLE III
THE MEAN SQUARE ERROR BETWEEN PROJECTED LANDMARKS AND GROUND-TRUTH LANDMARKS OVER TEST DATA. THE FIRST ROW SHOWS THE METHODS, AND THE SECOND ROW SHOWS THEIR CORRESPONDING MEAN SQUARE ERRORS OF LANDMARKS. NOTE THAT THE COMPARED METHODS ALL NEED LABELED LANDMARKS INPUT, WHILE OUR METHOD AUTOMATICALLY DETECTS THE LANDMARKS AND RECONSTRUCTS THE 3D MESH FROM THE INPUT CARICATURE.

3DMM	FaceWareHouse	Alive	Ours
6.32	7.61	0.02	4.98

From the above quantitative and qualitative experiments, we can see that our proposed method performs quite well on landmark detection and reconstruction for caricatures. In Fig. 10,

Fig. 8. Left to right: input caricature, predicted mesh overlaying on the image, predicted mesh in two different views, predicted mesh with texture.

more experimental results and comparisons with the benchmark landmark detection methods [21], [18], [20] and the existing state-of-the-art caricature reconstruction method [12] on 10 test caricatures are given. These results further validate the superior effect of our proposed method on the tasks of landmark detection and 3D reconstruction on caricature.

V. Conclusion

We have presented an effective and efficient algorithm for automatic landmark detection and 3D reconstruction for 2D caricature images. This challenging problem is well solved by separately regressing the 3D face shape and face pose, and then 2D landmarks and 3D shape can both be obtained. To represent the non-regular 3D caricature face, we construct a 3D caricature shape dataset to learn the latent representation. Extensive experimental results show that the detected 2D landmarks and reconstructed 3D face shape fit the caricature

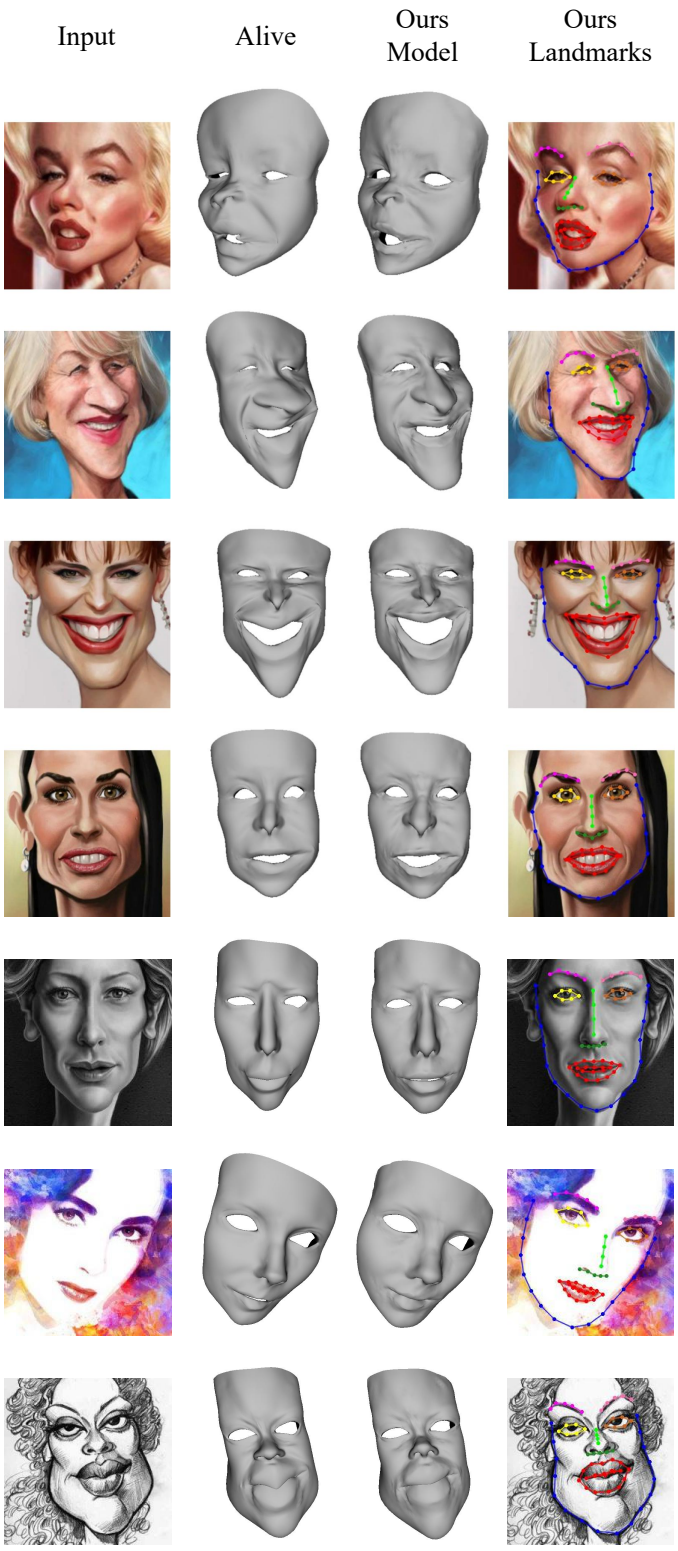

Fig. 9. Reconstruction results by our method and [12] which needs labeled landmarks. From the first column to the last column are input images, reconstruction results by [12], reconstruction results by our method, and the projected 2D landmarks by our method, respectively.

Fig. 10. Landmark detection comparisons with benchmark methods DAN [21], ERT [18], VCNN [20] and reconstruction comparisons with state-of-the-art method [12] which needs labeled landmarks. It can be seen that our method can detect landmarks and reconstruct 3D face shapes quite well.

quite well, which outperforms the existing state-of-the-art methods in both computation speed and accuracy.

Acknowledgement This work was supported by the National Natural Science Foundation of China (No. 61672481) and Youth Innovation Promotion Association CAS (No. 2018495).

REFERENCES

[1] S. E. Brennan, "Caricature generator: The dynamic exaggeration of faces by computer," *Leonardo*, vol. 18, no. 3, pp. 170–178, 1985.

[2] L. Liang, H. Chen, Y. Xu, and H. Shum, "Example-based caricature generation with exaggeration," in *10th Pacific Conference on Computer Graphics and Applications*, 2002, pp. 386–393.

[3] H.-Y. Shum, Y.-Q. Xu, M. F. Cohen, and H. Zhong, "Sample based face caricature generation."

[4] K. Cao, J. Liao, and L. Yuan, "Carigans: unpaired photo-to-caricature translation," *ACM Transactions on graphics (TOG)*, vol. 37, no. 6, pp. 244:1–244:14, 2018.

[5] Y. Shi, D. Deb, and A. K. Jain, "Warpgan: Automatic caricature generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 762–10 771.

[6] B. Klare, S. S. Bucak, A. K. Jain, and T. Akgul, "Towards automated caricature recognition," in *5th IAPR International Conference on Biometrics, ICB*, 2012, pp. 139–146.

[7] S. Ouyang, T. M. Hospedales, Y. Song, and X. Li, "Cross-modal face matching: Beyond viewed sketches," in *12th Asian Conference on Computer Vision*, 2014, pp. 210–225.

[8] B. Abaci and T. Akgul, "Matching caricatures to photographs," *Signal, Image and Video Processing*, vol. 9, no. Supplement-1, pp. 295–303, 2015.

[9] T. Lewiner, T. Vieira, D. Martínez, A. Peixoto, V. Mello, and L. Velho, "Interactive 3d caricature from harmonic exaggeration," *Computers & Graphics*, vol. 35, no. 3, pp. 586–595, 2011.

[10] R. C. C. Vieira, C. A. Vidal, and J. B. C. Neto, "Three-dimensional face caricaturing by anthropometric distortions," in *XXVI Conference on Graphics, Patterns and Images, SIBGRAPI*, 2013, pp. 163–170.

[11] X. Han, C. Gao, and Y. Yu, "Deepsketch2face: a deep learning based sketching system for 3d face and caricature modeling," *ACM Transactions on graphics (TOG)*, vol. 36, no. 4, pp. 126:1–126:12, 2017.

[12] Q. Wu, J. Zhang, Y.-K. Lai, J. Zheng, and J. Cai, "Alive caricature from 2d to 3d," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7336–7345.

[13] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin, "Webcaricature: a benchmark for caricature recognition," in *British Machine Vision Conference (BMVC)*, 2018, p. 223.

[14] W. Chu, W. Hung, Y. Tsai, Y. Chang, Y. Li, D. Cai, and M. Yang, "Learning to caricature via semantic shape transform," *CoRR*, vol. abs/2008.05090, 2020.

[15] K. Chen, J. Zheng, J. Cai, and J. Zhang, "Modeling caricature expressions by 3d blendshape and dynamic texture," in *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, Eds. ACM, 2020, pp. 3228–3236.

[16] S. Sadimon and H. Haron, "Neural network model for prediction of facial caricature landmark configuration using modified procrustes superimposition method." *International Journal of Advances in Soft Computing & Its Applications*, vol. 7, no. 3, pp. 42–66, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1867–1874.

[19] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[20] Y. Wu, T. Hassner, K. Kim, G. G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3067–3074, 2018.

[21] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2017, pp. 2034–2043.

[22] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 585–601.

[23] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N. M. Robertson, and J. Wang, "Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3467–3476.

[24] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 360–368.

[25] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1546–1555.

[26] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3486–3496.

[27] D. Merget, M. Rock, and G. Rigoll, "Robust facial landmark detection via a fully-convolutional local-global context network," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 781–790.

[28] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng, "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1294–1307, 2019.

[29] S. Liu, Y. Zhang, X. Yang, D. Shi, and J. Zhang, "Robust facial landmark detection and tracking across poses and expressions for in-the-wild monocular video," *Comput. Vis. Media*, vol. 3, no. 1, pp. 33–47, 2017.

[30] J. Wang, J. Zhang, C. Luo, and F. Chen, "Joint head pose and facial landmark regression from depth images," *Comput. Vis. Media*, vol. 3, no. 3, pp. 229–241, 2017.

[31] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[32] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5163–5172.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 146–155.

[34] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 534–551.

[35] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 183:1–183:14, 2015.

[36] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2387–2395.

[37] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces." in *Siggraph*, vol. 99, no. 1999, 1999, pp. 187–194.

[38] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Transactions on graphics (TOG)*, vol. 33, no. 4, p. 43, 2014.

[39] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[40] C. M. Grewe and S. Zachow, "Fully automated and highly accurate dense correspondence for facial surfaces," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 552–568.

[41] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu, "3d face reconstruction with geometry details from a single image," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4756–4770, 2018.

[42] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch, "Fitting 3d morphable face models using local features," in *IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 1195–1199.

[43] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[44] Y. Jin, D. Jiang, and M. Cai, "3d reconstruction using deep learning: a survey," *Commun. Inf. Syst.*, vol. 20, no. 4, pp. 389–413, 2020.

[45] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 7346–7355.

[46] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1155–1164.

[47] M. Feng, S. Zulqarnain Gilani, Y. Wang, and A. Mian, "3d face reconstruction from light field images: A model-free approach," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 501–518.

[48] A. J. O'toole, T. Vetter, H. Volz, and E. M. Salter, "Three-dimensional caricatures of human heads: distinctiveness and the perception of facial age," *Perception*, vol. 26, no. 6, pp. 719–732, 1997.

[49] A. J. O'Toole, T. Price, T. Vetter, J. C. Bartlett, and V. Blanz, "3d shape and 2d surface textures of human faces: The role of "averages" in attractiveness and age," *Image and Vision Computing*, vol. 18, no. 1, pp. 9–19, 1999.

[50] M. Sela, Y. Aflalo, and R. Kimmel, "Computational caricaturization of surfaces," *Computer Vision and Image Understanding*, vol. 141, pp. 1–17, 2015.

[51] J. Liu, Y. Chen, C. Miao, J. Xie, C. X. Ling, X. Gao, and W. Gao, "Semi-supervised learning in reconstructed manifold space for 3d caricature generation," in *Computer Graphics Forum*, vol. 28, no. 8. Wiley Online Library, 2009, pp. 2104–2116.

[52] M. Stricker, O. Augereau, K. Kise, and M. Iwata, "Facial landmark detection for manga images," *CoRR*, vol. abs/1811.03214, 2018.

[53] A. Mishra, S. N. Rai, A. Mishra, and C. V. Jawahar, "IIIT-CFW: A benchmark database of cartoon faces in the wild," in *Computer Vision - ECCV Workshops*, 2016, pp. 35–47.

[54] M. Botsch and O. Sorkine, "On linear variational surface deformation methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 213–230, 2007.

[55] J. Diebel, "Representing attitude: Euler angles, unit quaternions, and rotation vectors," *Matrix*, vol. 58, no. 15-16, pp. 1–35, 2006.

[56] L. Gao, Y. Lai, J. Yang, L. Zhang, S. Xia, and L. Kobbelt, "Sparse data driven mesh deformation," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 3, pp. 2085–2100, 2021.

[57] M. Alexa, "Linear combination of transformations," in *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 380–387.

[58] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.

[59] Z. Huang, J. Yao, Z. Zhong, Y. Liu, and X. Guo, "Sparse localized decomposition of deformation gradients," *Comput. Graph. Forum*, vol. 33, no. 7, pp. 239–248, 2014.

[60] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[62] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, 2-4 September 2009, Genova, Italy*, 2009, pp. 296–301.

[63] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 787–796.