3DCaricShop: A Dataset and A Baseline Method for Single-view 3D Caricature Face Reconstruction

Yuda Qiu\textsuperscript{1} Xiaojie Xu\textsuperscript{1} Lingteng Qiu\textsuperscript{1} Yan Pan\textsuperscript{1} Yushuang Wu\textsuperscript{1} Weikai Chen\textsuperscript{2} Xiaoguang Han\textsuperscript{1, *}

\textsuperscript{1}SRIBD, The Chinese University of Hong Kong, Shenzhen \textsuperscript{2}Tencent Game AI Research Center

Abstract

Caricature is an artistic representation that deliberately exaggerates the distinctive features of a human face to convey humor or sarcasm. However, reconstructing a 3D caricature from a 2D caricature image remains a challenging task, mostly due to the lack of data. We propose to fill this gap by introducing 3DCaricShop, the first large-scale 3D caricature dataset that contains 2000 high-quality diversified 3D caricatures manually crafted by professional artists. 3DCaricShop also provides rich annotations including a paired 2D caricature image, camera parameters and 3D facial landmarks. To demonstrate the advantage of 3DCaricShop, we present a novel baseline approach for single-view 3D caricature reconstruction. To ensure a faithful reconstruction with plausible face deformations, we propose to connect the good ends of the detail-rich implicit functions and the parametric mesh representations. In particular, we first register a template mesh to the output of the implicit generator and iteratively project the registration result onto a pre-trained PCA space to resolve artifacts and self-intersections. To deal with the large deformation during non-rigid registration, we propose a novel view-collaborative graph convolution network (VC-GCN) to extract key points from the implicit mesh for accurate alignment. Our method is able to generate high-fidelity 3D caricature in a pre-defined mesh topology that is animation-ready. Extensive experiments have been conducted on 3DCaricShop to verify the significance of the database and the effectiveness of the proposed method. We will release 3DCaricShop upon publication.

1. Introduction

A caricature is a vivid art form of depicting persons by abstracting or exaggerating the peculiarities of the facial features. As a way to convey humor or sarcasm, caricatures are widely used in entertainment, social events, electronic games and a variety of artistic creations. While 2D caricatures have gained popularity in comic graphics, there exist many scenarios, including cartoon character creation, game avatar customization, custom-made 3D printing, etc., that the 3D face caricatures remain the mainstream representations. However, creating a high-quality 3D caricature is a labor-intensive and time-consuming task even for a skilled artist. Thereby, generating expressive 3D face caricatures from a minimal input, such as a single image, is a highly-demanding but also challenging task.

Most of the prior works mainly focus on 2D caricature generation [6, 33, 16], while research on reconstructing 3D caricatures from 2D caricature images remains very rare. Wu et al. [39] propose the first work that creates 3D caricature from 2D caricature images using an optimization based approach. They formulate caricature generation as a problem of deforming the standard 3D face. In particular, they
build a intrinsic deformation space based on the exaggerated morphable models of standard faces [25]. The deformation coefficients are then optimized to reduce the landmark fitting errors. Recently, in their follow-up work [42], they employ CNN to automate the task of 2D facial landmark prediction and deformation regression. However, previous works [32, 18] have shown that the traditional 3D morphable models (3DMM) of normal faces have very limited expressiveness in modeling the intricate facial deformations in reality. Thereby, the deformation space based on a synthetically exaggerated 3DMM, as proposed in [39, 42], is far from sufficient to capture realistic 3D caricatures, which are even more diversified and complex than normal faces.

The key to tackling the above problem is a high-quality 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.

In this work, we introduce 3DCaricShop, a large-scale 3D caricature dataset created by artists that can provide realistic shape priors for both learning-based and optimization-based approaches. However, there exist two challenges in constructing such a dataset. First, the 3D models crafted by artists are not topologically consistent, making it infeasible to many downstream applications, including blend-shape creation, face animation, 3D landmark localization, etc. Secondly, the manually created meshes are typically not aligned with the corresponding images. While many face reconstruction techniques require an accurate registration, such misalignment makes the dataset inapplicable to projection-based applications such as landmark fitting, texture restoration and manipulation, etc.
parametric based [5, 34, 10] and shape-from-shading based [29, 35] methods, and remarkable results have been achieved. However, both methods could not apply on our task directly. Parametric methods suffer from the large diversity of geometry shapes in caricature cases. For SfS algorithms, the underlying physical model could not capture various painting styles of artists.

**3D Caricature Generation** Following the parametric based methods of normal face reconstruction, researchers further introduce deformation to enlarge the capability of representation [19, 39, 42]. In [19], a semi-supervised manifold regularization method is proposed to learn a regressive model for mapping from 2D real faces to the enlarged training set with 3D caricatures. Wu et al. [39] formulate the 3D caricature generation as a problem of deformation from the standard 3D face. By introducing local deformation gradients, they build an intrinsic deformation representation with the capability of extrapolation. With the deformation representation, they construct an optimization framework to create caricature model guided by the landmark constraint. Following [39], Zhang et al. [42] employ CNN to learn the deformation parameters of the intrinsic deformation representations. However, due to the lack of 3D caricature data, their works are still far from satisfaction.

**3D Face datasets** 3D face datasets are of great value in face reconstruction tasks. In general, they could be categorized into synthetic and real captured datasets. For normal face, existing 3D datasets, including FaceWareHouse [5] and Facescape [40], are built from scanned 3D data, hence widely used in normal face tasks. They focus on the high accuracy and photo reality of the meshes. However, they could not be applied directly on caricature reconstruction. Researchers [39, 42] tried to perform deformation on real 3D face models to construct synthetic exaggerated data. Although some reasonable results are achieved, they still suffer from the lack of diversity. To tackle this problem, we propose 3DCaricShop, which is the first 3D caricature dataset built by artists, composed pairs of caricature images and meshes. Based on the dataset, 3D caricature shape could be learned in a model-free manner. We further propose a baseline method to reconstruct 3D mesh with uniform topology from single caricature image.

**3. Dataset**

We construct a dataset which contains 2,000 image-model pairs in total. All of the 3D models are annotated with 3D facial landmarks and poses w.r.t images. More details are introduced in the following aspects.

**3D Model Collection** WebCaricature [15] is the largest-to-date dataset of 2D caricatures. It contains around 6,000 caricature images with diverse identities, geometry, and textural styles. We first selected 2,000 images from them, further making them as diverse as possible. Then we recruited 4 paid expert Zbrush artists to create models according to images. The modeling is required to be matched with the image as much as possible, in projection manner. The contour lines for matching include edges of silhouette, lips, eyes, nose and ears. It takes around 40 minutes for each model on average, and around 40 days are cost in total. Several image-model pairs sampled from our dataset are shown in Fig. 2. Each model consists of 300k ∼ 700k vertices.

**Meshing Unification** To support building parametric space for our 3D caricature dataset, we unify the mesh topology for all models in two steps: 1) We first manually annotate 44 3D landmarks (see details in Fig. 3) for each model; 2) The method of Non-rigid ICP [1, 9] is applied to register a pre-defined template mesh to each model, guided by the 3D landmarks. Due to the inherent difficulty to specify vertices on a 3D mesh, the landmark annotation is performed on 3 rendered views of the 3D shape. As described in [4], these 2D landmarks can be easily transformed into their corresponding 3D positions. The template mesh we use is from FaceWareHouse [5] that consists of 11,551 vertices. This procedure is illustrated by an example in Fig. 3.

**Pose Annotation** 3D pose estimation from a single image is the premise of our reconstruction method (see Sec. 4.1). It usually requires pose information supervision of the 3D face w.r.t the image. 3DCaricShop also provides accurate pose labels for each mesh that annotated by artists manually.

**Analysis of the dataset** We quantitatively analyze our dataset by comparing the shape variations with two normal face datasets (FaceWareHouse (FWH) and FaceScape), as well as one synthetic caricature dataset, FaceWareHouse with deformation (Aug. FWH). We measure the shape variation using global and part variance. In particular, the variance is computed between the models and their correspond-
A novel view-collaborative graph convolution network (VC-GCN) to predict key points $k$ from the implicit mesh, where $k \in \mathbb{R}^{34}$.

4. Methodology

Overview In this section, we introduce the proposed baseline method. Given an input caricature image $I$, the task is to generate the corresponding 3D mesh $M$. With the topologically uniform 3D meshes in 3DCaricShop, a straight forward way to tackle the task is to construct a PCA basis using the 3D Morphable Model algorithm [2] to build the caricature face space. However, such a space could not handle the large variation in our data. To capture the diversity of geometry in caricature, we employ Pixel-aligned Implicit Function (PIFu) [31] to generate the 3D shape $M_I$ from $I$. Although the implicit function models the variation in targets, it could not ensure a uniform topology for the predictions. To achieve that, we register a template mesh $M_t$ to $M_I$ using non-rigid registration (NICP) [1]. Then the output of NICP is projected onto the pre-constructed PCA space, to alleviate deformation artifacts, such as self-intersections. We denote the output of NICP as $M_N$ and that of PCA as $M_P$. Considering the large difference between the template and target meshes, a sparse 3D mesh is needed in the stage of NICP. We propose a novel view-collaborative graph convolution network (VC-GCN) to predict key points $k$ from the implicit mesh, where $k \in \mathbb{R}^{34}$.

4.1. The Baseline Approach

Parametric Modeling Our parametric model space is built with standard 3D Morphable Model (3DMM) [2] algorithm. Given $p$ caricature models with uniform topology and $N$ vertices on each mesh, principal component analysis (PCA) is performed on the shape matrix $S_M \in \mathbb{R}^{3N \times p}$, which is formed by stacking the $3D$ coordinates of the $N \times p$ vertices. The generated $d$ eigen-vectors are employed as the shape basis $S_i, i = 1, 2, ..., d$, where $d$ is a hyper-parameter. The mean vector $\overline{S}$ represents the mean shape in the mesh set. With this 3DMM, a novel caricature model $S_N$ could be represented as follows: $S_N = \overline{S} + \sum_{i=1}^{d} a_i S_i$, where $a = [a_1 \cdots a_d]^T$ is the vector of shape coefficients.

Implicit Reconstruction To capture the diversity of geometric variation, we adopt Pixel-aligned Implicit Function (PIFu) [31] to reconstruct the underlying 3D shape from images. PIFu performs 3D reconstruction by estimating the occupancy of a dense 3D shape, which determines whether a point in 3D space is inside the model or not. Given a RGB image $I$, its normal maps from the front view $F$ and back view $B$ are generated to strengthen the local details, by using a pixel2pixel-hd network [37]. Then the implicit binary function $f(X, I, F, B)$ could be written as:

$$f(X, I, F, B) = \begin{cases} 1, & \text{if } X \text{ is inside the mesh surface,} \\ 0, & \text{otherwise.} \end{cases}$$

(1)

where $X$ is a given 3D location in the continuous camera space. This function is modeled by a neural network. The loss function for training is formulated as:

$$L = \frac{1}{n} \sum_{i=1}^{n} |f(X_i, I_i, F_i, B_i) - f^*(X_i)|^2,$$  

(2)

where $f^*(X_i)$ is the ground-truth occupancy.

3D Landmark Detection for Registration The output meshes $M_I$ of the implicit function are not topologically uniform. In order to unify the topology, we adopt non-rigid registration[1] to deform a template $M_t$ into the shape of $M_I$. As shown in [1], without landmarks the cost function of registration could run into a local minimum, where the template is collapsed onto a point on the target surface. Thus it is important to introduce the 3D landmarks of both meshes to guide the deformation. We design a novel framework to detect 3D landmarks for $M_I$. In short, we propose
to perform detection on the rendered views of $M_1$ to leverage the effectiveness of image-based CNN techniques. The process would be detailed in 4.2.

**Landmark-guided Registration** Since the huge difference between $M_1$ and $M_2$, the deformation is likely to generate artifacts, such as self-intersection. To resolve this problem, we iteratively perform NICP and PCA projection to obtain $M_2$ and $M_2$. After projection, we obtain a deformed template which is closer to $M_1$ in shape. Fig. 4 illustrates the process of the progressive deformation.

### 4.2. View-collaborative 3D Landmark Detection

In this section, we discuss more details about how to detect 3D landmarks from $M$, which is the key to supporting the procedure of landmark-guided registration. A straightforward way for this detection is directly applying point-based CNN (e.g., SparseConv [17]) to estimate landmark-whare heatmap on mesh vertices. However, due to the inherent difficulty to conduct CNN on a mesh, this approach tends to produce inaccurate results. We thus propose to perform detection on the rendered views of $M_1$ to leverage the effectiveness of image-based CNN techniques. A coarse locations of the 3D landmarks can be obtained from detected 2D landmarks on those views. More importantly, a stack of View-Collaborative GCN block (VC-GCN) is novelly designed to aggregate and enhance information from multiple views for accurate 3D landmarks locations. As illustrated in Fig. 5, single view graph features (local features) are first extracted for initialization. Then, these local features are enhanced in a progressive manner by continually fusing global information into each view. The final local features are aggregated into the global graph feature for 3D landmark prediction.

**Initialization Stage** In this part, more details about local feature initialization are introduced. Given 2D images rendered from 3 views (front, left, right) ({$f, l, r$} for simplicity), we first utilize a 2D landmark detector [12] to estimate the 2D landmarks $\hat{D}^v \in \mathbb{R}^{k^v \times 2}$, where $k^v$ denotes the key point number under the view $v \in \{f, l, r\}$. Next, 3D landmarks on the mesh are located using the projection matrix of each view. We use the above landmarks which exist in all local views to build local graphs. After that, to enrich the information of each graph node, we extract features from the feature maps generated by the landmark detector for each node, according to their 2D coordinates. Eventually, the initial local view features $F_{init}^v \in \mathbb{R}^{k^v \times (C+3)}$ for VC-GCN can be produced by concatenating the 3D landmark locations $L^v$ with related node features under each view $v$, where $C$ is the image feature dimension.

**View-Collaborative GCN** In order to provide global information for each view, we aggregate 3 local features into a global graph feature. Then the global feature is fused into each view to enhance the local view feature. This procedure is performed by a View-Collaborative GCN block. In each VC-GCN block, local features are first sent into several GCN layers for better representations. The layer-wise operation in GCN is defined as:

$$F_{\ell+1}^v = \sigma \left( D^v \frac{1}{2} A^v D^v \frac{1}{2} F_{\ell}^v W_{\ell}^v \right)$$

(3)

where $A^v$ is the adjacency matrix with self-loops, $D^v$ is its diagonal node degree matrix to normalize $A^v$, $F_{\ell}^v$ represents the local feature in layer $\ell \in \{0, 1, \ldots, \ell'\}$ under the view $v$, $W_{\ell}$ is a trainable parameter matrix for linear projection, and $\sigma(\cdot)$ represents the non-linear activation operation. Then the obtained local features are combined into a global graph feature. For each node in the 3D landmark, its feature can be drawn from the local feature of the corresponding node under one view. Note that for the node that shared in different views, its feature is set as the average of multiple local features. The combined global graph feature is then strengthened through several GCN layers, with same operations as in Equation 3. Hence, the process of feature aggregation now can be formulated as following:

$$F_{\ell}^v = GCN^v(F_{\ell-1}^v), \quad v = f, l, r;$$

$$F_{\ell}^v = GCN^v(f_{comb}(F_{\ell-1}^f, F_{\ell-1}^l, F_{\ell-1}^r)).$$

(4)

where $f_{comb}(\cdot)$ denotes the combination operation, $F_{\ell}^v$ is the strengthened global features. Note that the input $F_{\ell}^v$ is set
as the initial local view feature $F_{init}^{v}$ in the first VC-GCN block, and is set as the prior output features in subsequent blocks.

In the second step, strengthened global features are fused into local features of each view in a non-local manner \[38\], so that global information can guide the model to learn more representative local features. The enhanced local features $F_{enh}^{v}$ of the view $v$ can be obtained as following, where the non-local fusion operation is denoted as $f_{n-loc}$:

$$F_{enh}^{v} = f_{n-loc}(F_{init}^{v}, F_{v})_{v = f, l, r}. \quad (5)$$

More details about the non-local fusion operation are described in the supplementary materials.

It usually takes numerous glimpses to adjust key points to construct a 3D face, even for an expert artist. Thus, several VC-GCN blocks are stacked to progressively enhance local features. In the connection of two blocks, the enhanced local features $F_{enh}$, from the former block are taken as the input $F_{0}^{v}$ of the later block.

**Loss Function** Given the enhanced local features from the last VC-GCN block, we combine and strengthen them using GCN layers to obtain the final global graph features. Next, it is multiplied by a GCN head layer to get the 3D landmark estimation $\hat{L}^{v} \in \mathbb{R}^{N \times 3}$. The predicted 3D landmarks are supervised by 3D and 2D landmark ground truth simultaneously, which leads to more accurate prediction. We now formulate the loss function for the view-collaborative 3D landmark detection training as following:

$$\mathcal{L} = \sum_{i \in \Omega, v} (\mathcal{L}_{detect}(\hat{P}_{v}^{\ell}, P_{v}^{\ell})) + \mathcal{L}_{3D}(\hat{L}_{v}^{l}, L_{v}^{l}) + \mathcal{L}_{2D}(M^{v}(\hat{L}_{v}^{l}, P_{v}^{\ell})), \quad (6)$$

where $\Omega$ is the training set, $i$ is the subscript indicating each training sample, $\mathcal{L}_{detect}$ denotes the 2D landmark detection error in the initialization stage, $\mathcal{L}_{3D}$ and $\mathcal{L}_{2D}$ represents the 3D landmark prediction error in 2D and 3D space, respectively, and $M^{v}$ is the projection matrix to obtain 2D landmarks from 3D landmarks under the view $v$. Note that all loss terms are in smooth-$l_2$ form.

**5. Experiments**

**Implementation details** The proposed framework is trained on our 3DCaricShop. The dataset is separated into 1,600 for training and 400 for testing. The weights of $\mathcal{L}_{detect}$, $\mathcal{L}_{2D}$ and $\mathcal{L}_{3D}$ are set to 0.1, 0.8 and 1.0 respectively. To train the network for learning the implicit reconstruction, a RMSProp optimizer is adopted with learning rate 0.001, and the network is pre-trained with the mini-batch size 2 for 80 epochs. During the training of 2D landmark detection network, an Adam optimizer is used. The learning rate is set to 0.0001 with a cosine decay, and the mini-batch size is set to 24 for 30 epochs. After that, the whole framework is trained in an end-to-end manner with the same strategy as above.

**Results Gallery** We present some typical results of the proposed framework in Fig. 6. As illustrated, our method is robust to caricature images with diverse textures. It can also recover diversified geometric features, such as the exaggerated nose in the second sample of the first row, and the sharp long chin in the third sample of the second row Fig. 6.

**5.1. Comparisons with the State-of-the-arts**

We qualitatively and quantitatively compare the results of our method with a variety of state-of-the-art 3D caricature reconstruction approaches on 3DCaricShop testing set, including linear parametric model (3DMM) \[35, 2\], depth map (DF2Net) \[41\], deformation representation (AlvieCaric-DL) \[42\], and implicit function (PIFu) \[31\] based methods.
Figure 6: Results gallery for our framework on 3DCaricShop. The framework has the capability to reconstruct 3D shapes from caricature images with diverse texture and geometry shapes.

**Qualitative results** In Fig. 7, we visualize some results of caricature reconstruction on images from 3DCaricShop. Among the parametric methods, the nonlinear deformation representation [42] based model outperforms the linear ones on fitting the exaggerated input images, but it is still not precise enough due to its limited expressiveness. Besides, other deep learning based approaches such as DF2Net and PiFu unavoidably yield artifacts like hollows and spikes. However, our method introduces the constraint of PCA parametric space into the deep model, thus can produce highly exaggerated local details upon the foundation of a plausible global shape.

**Quantitative Results** Considering other methods for comparison only reconstruct the face area, we adopt average point-to-surface Euclidean distance (P2S) as the evaluation metric, which measures the unidirectional distance from the source set to the target set. The average point-to-surface Euclidean distance can be computed as:

$$d_{P2S}(P, S) = \frac{1}{\|P\|} \sum_{p \in P} \min_{p' \in S} \|p - p'\|_2$$  \hspace{1cm} (7)

where $P$ is the vertex set of the reconstructed mesh and $S$ is the corresponding ground truth surface. Besides, due to the mismatch in orientation and scale between the generated meshes and ground truth, before calculation, Procrustes alignment is performed and scaling is estimated based on least square error. As shown in Table 2, our method achieves the smallest P2S on the 3DCaricShop.

### Table 2: Quantitative evaluation on 3DCaricShop

<table>
<thead>
<tr>
<th>Methods</th>
<th>P2S</th>
<th>Methods</th>
<th>P2S</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DMMhuman</td>
<td>0.295</td>
<td>PiFuhead</td>
<td>0.153</td>
</tr>
<tr>
<td>3DMM_carli</td>
<td>0.104</td>
<td>PiFuface</td>
<td>0.126</td>
</tr>
<tr>
<td>DF2Net</td>
<td>0.273</td>
<td>Ourshead</td>
<td>0.065</td>
</tr>
<tr>
<td>AliveCaric-DL</td>
<td>0.067</td>
<td>Oursface</td>
<td>0.037</td>
</tr>
</tbody>
</table>

‘V2V’; 3) employing a global graph to refine the 3D landmarks from the first setting, without using VC-GCN block, which denoted as ‘Global only’; 4) Only using local index to gather local features from global view, rather than using non-local operations for local feature enhancement, denoted as ‘w/o G2L’; 5) The basic setting, denoted as Basic.

The metric we evaluate the results is mean per joint position error (MPJPE) which is defined as a Euclidean distance between predicted and ground truth 3D landmarks after root joint alignment. The root joint we define is the top of nose. This metric measures how accurately the root-relative 3D landmark estimation is performed. The quantitative results are listed in Table 3. It confirm the effectiveness of each components in the design of our 3D landmark detector.

**On landmark guided registration** We evaluate three kinds of registration process: 1) directly perform NICP without landmarks information; 2) perform landmark-guided NICP without PCA projection; 3) the process used in our method. The visualized results are shown in Fig. 9. As [1] suggested, without landmark information, NICP could not capture the large discrepancy between $M_T$ and $M_I$. Besides, the deformation without PCA projection is likely to generate meshes with self-intersection. In contrast, our method could obtain meshes with higher quality, and capture enough shape information in $M_I$. For example, the artifacts on ears in Fig. 9(d) are eliminated, while the shape
Figure 7: Qualitative results of our method compared with state-of-the-art methods, including (a) 3DMM-Human [35], (b) 3DMM-cari [2], (c) DF2Net [41], (d) AliveCaric-DL [42] and (e) PiFu [31], on 3DCaricShop. By incorporating deep models with parametric space constraint, our method (g) can reconstruct highly exaggerated geometry without distinct artifacts.

Figure 8: Ablation experiments on 3D landmark detection: (a) input; (b) ground truth; (c) pure projection; (d) global only; (e) w/o G2L; (f) ours. Our method could capture the face geometry more accurately.

Figure 9: Ablation experiments on registration: (a) input; (b) GT; (c) NICP w/o landmark; (d) NICP w/o PCA projection; (e) ours. A better result is obtained with reasonable topology (e.g., the nose) by using PCA projection.

Table 3: Ablation study for 3D landmark detection.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MPJPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o GCN Refinement</td>
<td>0.451</td>
</tr>
<tr>
<td>V2V [22]</td>
<td>0.407</td>
</tr>
<tr>
<td>Global only w/o G2L</td>
<td>0.375</td>
</tr>
<tr>
<td>Basic</td>
<td>0.291</td>
</tr>
</tbody>
</table>

Table 4: Ablation study for landmark-guided registration. The similar data implies the improvement of our method concentrate on the detail structures.

Table 4: Ablation study for landmark-guided registration.

<table>
<thead>
<tr>
<th>Methods</th>
<th>P2S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours w/o Landmark</td>
<td>0.074</td>
</tr>
<tr>
<td>Ours w/o PCA projection</td>
<td>0.076</td>
</tr>
<tr>
<td>Ours</td>
<td>0.065</td>
</tr>
</tbody>
</table>

6. Conclusions

We construct a new dataset and benchmark, called 3DCaricShop, for single-view 3D reconstruction from caricature images. 3DCaricShop is the largest collection by far of 3D caricature models crafted by professional artists. It consists of 2,000 high-quality and diversified 3D caricatures that are richly labeled with paired 2D caricature image, camera parameters, and 3D facial landmarks. A novel baseline approach is also presented to validate the usefulness of the proposed dataset. It combines the merits of flexible implicit functions and the robust parametric mesh representation. Specifically, we transfer the details from implicit reconstruction to a template mesh with the help of VC-GCN that accurately predicts 3D landmarks for the implicit mesh. Extensive benchmarking on our dataset has been performed including a variety of popular approaches. We found that reconstructing 3D caricature from a single 2D caricature image is a highly challenging task with ample opportunity for improvement. We hope 3DCaricShop and our baseline approach could shed light on future research in this field.

Acknowledgements

The work was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund, and by Guangdong Research Project No. 2017ZT07X152, the National Natural Science Foundation of China 61902334, Shenzhen Fundamental Research (General Project) JCYJ20190814112007258.
References


[27] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification


