High-Fidelity 3D Digital Human Head Creation from RGB-D Selfies

LINCHAO BAO∗, XIANGKAI LIN∗, YAJING CHEN∗, HAOXIAN ZHANG∗, SHENG WANG, XUEFEI ZHE, DI KANG, and HAOZHI HUANG, Tencent AI Lab, China
XINWEI JIANG, Tencent NExT Studios, China
JUE WANG, DONG YU, and ZHENGYOU ZHANG, Tencent AI Lab, China

We present a fully automatic system that can produce high-fidelity, photo-realistic 3D digital human heads with a consumer RGB-D selfie camera. The system only needs the user to take a short selfie RGB-D video while rotating his/her head, and can produce a high quality head reconstruction in less than 30 seconds. Our main contribution is a new facial geometry modeling and reflectance synthesis procedure that significantly improves the state-of-the-art. Specifically, given the input video a two-stage frame selection procedure is first employed to select a few high-quality frames for reconstruction. Then a differentiable renderer based 3D Morphable Model (3DMM) fitting algorithm is applied to recover facial geometries from multiview RGB-D data, which takes advantages of a powerful 3DMM basis constructed with extensive data generation and perturbation. Our 3DMM has much larger expressive capacities than conventional 3DMM, allowing us to recover more accurate facial geometry using merely linear basis. For reflectance synthesis, we present a hybrid approach that combines parametric fitting and CNNs to synthesize high-resolution albedo/normal maps with realistic hair/pore/wrinkle details. Results show that our system can produce faithful 3D digital human faces with extremely realistic details. The main code and the newly constructed 3DMM basis is publicly available.

CCS Concepts: • Computing methodologies → Reconstruction; Mesh models.

Additional Key Words and Phrases: digital human, 3D face, avatar, 3DMM

∗The first four authors contributed equally to this paper.

Fig. 1. Our system takes a user’s RGB-D selfies as inputs and automatically produce a high-fidelity, riggable head model with high-resolution albedo map and normal map. The model faithfully preserves the user’s facial identity features and can be rendered as a realistic digital human character.

1 INTRODUCTION

Real-time rendering of realistic digital humans is an increasingly important task in various immersive applications like augmented and virtual reality (AR/VR). To render a realistic human face, high-quality geometry and reflectance data are essential. There exist specialized hardware like Light Stage [Alexander et al. 2009] for high-fidelity 3D faces capturing and reconstruction in the movie industry, but they are cumbersome to use for consumers. Research efforts have been dedicated to consumer-friendly solutions, trying to create 3D faces with consumer cameras, e.g., RGB-D data [Thies et al. 2015; Zollhöfer et al. 2011], multiview images [Ichim et al. 2015], or even a single image [Hu et al. 2017; Lattas et al. 2020; Yamaguchi et al. 2018]. While good results have been shown, the reconstructed 3D faces still contain artifacts and are far from satisfactory.

Indeed, faithful 3D facial reconstruction is a challenging problem due to the extreme sensitivity that human perception has towards faces. First, the recovered facial geometry needs to preserve all important facial features like cheek silhouettes and mouth shapes. Single-image based approaches [Hu et al. 2017; Lattas et al. 2020; Yamaguchi et al. 2018] can hardly achieve this due to the lack of reliable geometric constraints. With multiview RGB/RG-D inputs, existing approaches [Ichim et al. 2015; Thies et al. 2015; Zollhöfer et al. 2011] do not fully leverage most recent advances in deep learning and differentiable rendering [Gecer et al. 2019; Genova et al. 2018], leading to inaccurate recovery that does not fully resemble the user’s facial shape. Second, the synthesized facial reflectance maps need to be high-resolution with fine details like eyebrow hair, lip wrinkles, and pore details on facial skin. Several recent work...
We propose a hybrid approach that consists of a regional parametric fitting, and increase system robustness. Moreover, we propose an effective frame selection scheme, as well as an initial model fitting procedure, which can avoid enforcing conflicting constraints and increase system robustness. Moreover, we propose an effective approach that takes advantages of extensive data generation and perturbation to construct the 3DMM, which has much larger expressive capacity compared with previous methods. We show that even with the linear basis of the new 3DMM, our method can consistently recover accurate, personalized facial geometry.

For facial reflectance modeling, we use high-resolution 2K (2048 × 2048) UV-maps consisting of an albedo map and a normal map. We propose a hybrid approach that consists of a regional parametric fitting and CNN-based refinement networks. The regional parametric fitting is based on a set of novel pyramid bases constructed by considering variations in multi-resolution albedo maps, as well as high-resolution normal maps. Faithful but over-smoothed high-resolution albedo/normal maps can be obtained in this step. GAN-based networks are then employed to refine the albedo/normal maps to yield the final high-quality results. Our experiments show that even with the 680 × 480 resolution inputs, our method can produce high-resolution albedo/normal maps, where eyebrow hair, lip wrinkles and facial skin pores are all clearly visible. The high-quality reflectance maps significantly improve the realism of the final renderings in real-time physically based rendering engines.

With the recovered facial geometry and reflectance, we further present a fully automatic pipeline to create a full head rig, by completing a head model, matching a hair model, estimating the position/scale of eyeballs/teeth models, generating the expression blend-shapes, etc. We conduct extensive experiments and demonstrate potential applications of our system.

Our major contributions include:

- A fully automatic system for producing high-fidelity, realistic 3D digital human heads with consumer-level RGB-D selfie cameras. Compared with previous avatar approaches, our system can generate higher quality assets for physically based rendering of photo-realistic 3D characters. The total acquisition and production time for a character is less than 30 seconds. The core code and 3DMM is publicly available.1
- A robust procedure consisting of frame selection, initial model fitting, and differentiable renderer based optimization to recover faithful facial geometries from multiview RGB-D data, which can tolerate data inconsistency introduced during user data acquisition.
- A novel morphable model construction approach that takes advantages of extensive data generation and perturbation. The constructed linear 3DMM by our approach has much larger expressive capacity than conventional 3DMM.
- A novel hybrid approach to synthesize high-resolution facial albedo/normal maps. Our method can produce high-quality results with fine-scale realistic facial details.

2 RELATED WORK

Creating high-fidelity realistic digital human characters commonly relies on specialized hardware [Alexander et al. 2009; Beeler et al. 2010; Debevec et al. 2000] and tedious artist labors like model editing and rigging [von der Pahlen et al. 2014]. Several recent work seek to create realistic 3D avatars with consumer devices like a smartphone using domain specific reconstruction approaches (i.e., with face shape/appearance priors) [Ichim et al. 2015; Lattas et al. 2020; Yamaguchi et al. 2018]. We mainly focus on prior arts along this line and briefly summarize the most related work in this section. Please refer to the recent surveys [Egger et al. 2020; Zollhöfer et al. 2018] for more detailed reviews.

2.1 Face 3D Morphable Model

The 3D morphable model (3DMM) is introduced in [Blanz and Vetter 1999] to represent a 3D face model by a linear combination of shape and texture bases. These bases are extracted with PCA algorithm on topological aligned 3D face meshes. To recover a 3D face model from observations, the 3DMM parameters can be estimated instead. Since the 3DMM bases are linear combinations of source 3D models, the expressive capacity of a 3DMM is rather limited. Researchers tried to increase the capacity either by automatically generating large amounts of topological aligned face meshes [Booth et al. 2016] or turn the linear procedure into a nonlinear one [Lüthi et al. 2017; Tran and Liu 2018]. However, the sampled face models with these 3DMM models are usually flawed (e.g., distorted mesh surfaces around eyes and mouths due to imperfect alignment [Booth et al. 2016], twisted [Lüthi et al. 2017] or noisy [Tran and Liu 2018] face meshes) and not suitable for realistic digital human rendering. Another line to increase the expressive capacity of 3DMM is to segment the face into regions and then employ spatially localized bases to model each region [Blanz and Vetter 1999; Neumann et al. 2013; Tena et al. 2011]. We present a novel data augmentation approach that can effectively increase the capacity of either global or localized 3DMM with the same amount of source 3D face meshes as existing approaches.

2.2 Facial Geometry Capture

Capturing from Single Image. Given a single face image, the 3D face model can be recovered by estimating the 3DMM parameters with analysis-by-synthesis optimization approaches [Blanz and Vetter 2003; Garrido et al. 2013, 2016; Gecer et al. 2019; Hu et al. 2017; Romdhani and Vetter 2005; Thies et al. 2016; Yamaguchi et al. 2018]. A widely adopted approach among them is described in the Face2Face work [Thies et al. 2016], where the optimization objective consists of photo consistency, facial landmark alignment, and statistical regularization. Although there is a recent surge of deep learning based approaches to use CNNs to regress 3DMM parameters [Genova et al. 2018; Fedus et al. 2018; Park et al. 2019; Zollhöfer et al. 2018], no work has been able to recover meaningful 3DMM information from a single face image.

1Code and 3DMM is available at: https://github.com/tencent-ailab/hifi3dface
et al. 2018; Tewari et al. 2017; Tran et al. 2017; Zhu et al. 2016], the results are commonly not in high fidelity due to lack of reliable geometric constraints. Some work go beyond the 3DMM parametric estimation to use additional geometric representations to model facial details [Chen et al. 2020; Guo et al. 2019; Jackson et al. 2017; Kemelmacher-Shlizerman and Basri 2011; Richardson et al. 2017; Sela et al. 2017; Shi et al. 2014; Tewari et al. 2018; Tran et al. 2018], but the results are generally not satisfactory for realistic rendering.

Capturing from Multiview Images. Ichim et al. [2015] present a complete system to produce face rigs by taking hand-held videos with a smartphone. The system relies on a multiview stereo reconstruction of the captured head followed by non-rigid registrations, which is slow and error-prone, especially when motion occurs or no reliable feature points can be detected in face regions. Recent research on multiview face reconstruction with deep learning methods [Dou and Kakadiaris 2018; Wu et al. 2019] do not explicitly model geometric constraints and are not accurate enough for high-fidelity rendering.

Capturing from RGB-D Data. Modeling facial geometries from RGB-D data commonly consists of several separated steps [Bouaziz et al. 2013; Li et al. 2013; Weise et al. 2011; Zollhöfer et al. 2011, 2014]. First, accumulated point clouds are obtained with rigid registration [Newcombe et al. 2011]. Then a non-rigid registration procedure is employed to obtain a deformed mesh from the target mesh model [Bouaziz et al. 2016; Chen et al. 2013]. Finally, in order to obtain a 3DMM parametric representation, a morphable model fitting is applied using the deformed mesh as geometric constraints [Bouaziz et al. 2016; Zollhöfer et al. 2011]. Although the approach is widely adopted as standard practices, it suffers from accumulated errors due to the long pipeline. Thies et al. [2015] propose to use an unified parametric fitting procedure to directly optimize camera poses together with 3DMM parameters, taking into account both RGB and depth constraints. Their method achieves high-quality results in facial expression tracking, but is not specially designed for recovering personalized geometric characteristics.

2.3 Facial Reflectance Capture

Saito et al. [2017] propose to synthesize high-resolution facial albedo maps using CNN style features based optimization like the style transfer algorithm [Gatys et al. 2016]. However, the approach requires iterative optimization and needs several minutes of computation. Yamaguchi et al. [2018] further propose to infer albedo maps, as well as specular maps and displacement maps, using texture completion CNNs and super-resolution CNNs. GANFIT [Gecer et al. 2019] employs the latent vector of a Generative Adversarial Network (GAN) as the parametric representation of texture maps and then use an differentiable renderer based optimization to estimate the texture parameters. The most recent work AvatarMe [Lattas et al. 2020] propose to infer separated diffuse albedo maps, diffuse normal maps, specular albedo maps, and specular normal maps using a series of CNNs. We present a novel hybrid approach that can achieve high-quality results while at the same time is more robust than the above pure CNN-based approaches.

2.4 Full Head Rig Creation

To complete a full head avatar model, accessories beyond face region need to be attached to the recovered face model, e.g., hair, eyebrows, teeth, etc. Ichim et al. [2015] describe a simple solution to transfer these accessories (except hair) from a template model and adapt the scales/positions to the reconstructed face model. Cao et al. [2016] use image-based billboards to deal with eyes and teeth, and coarse geometric proxy to deal with hair model. Nagano et al. [2018] employ a GAN-based network to synthesis mouth interiors. Hu et al. [2017] propose to perform hair digitization by parsing hair attributes from the input image and then retrieving a hair model for further refinement. There are also some approaches working on modeling hairs in strand level [Chai et al. 2015; Hu et al. 2015; Luo et al. 2013; Saito et al. 2018; Wei et al. 2005]. We in this paper adopt a simplified retrieval approach like Hu et al. [2017] to attach hair/eyebrows/teeth models.

For expression blendshape generation, Ichim et al. [2015] present a dynamic modeling process to produce personalized blendshapes, while Hu et al. [2017] adopt a simplified solution to transfer generic FACS-based blendshapes to the target model. The expression blendshapes can also be generated with a bilinear 3DMM model like FaceWarehouse [Cao et al. 2014], where the face identity and expression parameters are in independent dimensions. Wang et al. [2020] recently present a global-local multilinear framework to synthesize high-quality facial expression blendshapes. For simplicity, we adopt a blendshape transfer approach similar to Hu et al. [2017] to obtain expression blendshapes for rigging.

3 OVERVIEW

We first introduce the 3D face dataset used in our system. Then we describe the goal of our system, followed by the user data acquisition process and a summary of the main processing steps.

3D Face Dataset. We use a specialized camera array system [Beeler et al. 2010] to scan 200 East Asians, including 100 males and 100 females, aged from 20 to 50 years old (with their permissions to use their face data). The scanned face models are manually cleaned and aligned to a triangle mesh template with 20,481 vertices and 40,832 faces. Each face model is associated with a 2K-resolution (2048 × 2048) albedo map and a 2K normal map, where pore-level details are preserved. A linear PCA-based 3DMM [Blanz and Vetter 1999] can be constructed from the dataset, which consists of shape basis, albedo map basis, and normal map basis. Note that we propose a novel approach to construct an augmented version of the 3DMM shape basis in Sec. 5.3. Besides, a novel pyramid version of the 3DMM albedo/normal maps is presented in Sec. 6.1.

Goal. The goal of our system is to capture high-fidelity users facial geometry and reflectance with RGB-D selfie data, which is further used to create and render full-head, realistic digital humans. For geometry modeling, we use 3DMM parameters to represent a face, since it is more robust to degraded input data and with more controllable mesh quality than deformation-based representations. For reflectance modeling, we synthesize 2K-resolution albedo and normal maps regardless of the input RGB-D resolution.
**User Data Acquisition.** We use an iPhone X to capture user selfie RGB-D data. Note that it is common nowadays for a smartphone to be equipped with a front-facing depth sensor and any such phone can be used. While a user is taking selfie RGB-D video, our capturing interface will guide the user to consecutively rotate his/her head to left, right, upward, and back to middle. The entire acquisition process takes less than 10 seconds, and a total of 200-300 frames of RGB-D images are collected, with resolution $640 \times 480$. The face region for computation is cropped (and resized) to $300 \times 300$. The camera intrinsic parameters are directly read from the device.

**Processing Pipeline.** We first employ an automatic frame selection algorithm to select a few high-quality frames that cover all sides of the user (Sec. 4). Then an initial 3DM model fitting is computed with the detected facial landmarks in the selected frames (Sec. 5.1). Starting from the initial fitting, a differentiable renderer based optimization with multiview RGB-D constraints (Sec. 5.2) is applied to solve the 3DM parameters as well as lighting parameters and poses. Based on the estimated parameters, high-resolution albedo/normal maps are then synthesized (Sec. 6). Finally, high-quality, realistic full head avatars can be created and rendered (Sec. 7).

4 FRAME SELECTION

There are typically 200–300 frames acquired from a user. For efficiency and robustness, we developed a robust frame selection procedure to select a few high-quality frames for further processing, which considers both view coverage and data quality. As shown in Fig. 2, the procedure consists of two stages as described below.

**Coarse Screening and Preprocessing.** We first apply a real-time facial landmark detector (a MobileNet [Howard et al. 2017] model trained on 300W-LP dataset [Zhu et al. 2016]) on RGB images to detect 2D landmarks for each frame. Then a rough head pose for each frame can be efficiently computed with the correspondences between the 2D landmarks and the 3D keypoints on a template 3D face model using PnP algorithm [Lepetit et al. 2009]. Frames with extreme/invalid poses or closed-eye/opened-mouth expressions can be easily identified and screened out with the 2D landmarks and rough head poses. We categorize the rest frames by poses into groups: front, left, right, and up. Each group only keeps 10–30 frames near the center pose of the group. Note that more groups can be obtained by categorizing the frames with finer-level angle partitioning. We experimented with different number of groups and found four is a good balance between accuracy and efficiency. The remaining depth images are preprocessed to remove depth values beyond the range between 40cm and 1m (the typical selfie distances). Bilateral filtering [Paris and Durand 2009] with a small spatial and range kernel is then applied to the depth images to attenuate noises.

**Frame Selection.** For each group, we further select one frame based on two criteria: image quality and rigidity. To measure the image quality of a frame, we compute the Laplacian of Gaussian (LoG) filter response and use the variance as a motion blur score (images with a larger score are sharper). A front face frame is first selected based on the motion blur score in the front group. We then compute the rigidity between each frame in the other groups and the front face with the help of depth data. Specifically, the detected 2D landmarks for each frame are lifted from 2D to 3D using depth data. Note that occluded landmarks are automatically removed according to the group that a frame belongs to, e.g., for a frame in the left group, the landmarks on the right side of the face are removed. We use RANSAC method to compute the relative pose between each frame in the other groups and the front face using the 3D-3D landmark
correspondences [Arun et al. 1987]. Frames with too many outliers are considered as low rigidness and thus are excluded. Then a best frame in each group can be found based on the motion blur score. The output of this step is four frames with the 3D landmarks.

5 FACIAL GEOMETRY MODELING

5.1 Initial Model Fitting

We use PCA-based linear 3DMM [Blanz and Vetter 1999] for parametric modeling. The shape and albedo texture of a face model is represented as

\[ \text{s} = \bar{s} + \bar{S} \cdot x_{shp}, \]
\[ \text{a} = \bar{a} + \bar{A} \cdot x_{alb}, \]

where \( \bar{s} \) is the vector format of the mean 3D face shape model, \( \bar{S} \) is the corresponding identity parameter vector to be estimated, \( \bar{a} \) is the vector format of the mean albedo map, \( \bar{A} \) is the albedo map basis, \( x_{shp} \) is the corresponding albedo parameter vector to be estimated. The details of the bases are presented in Secs. 5.3 (shape) and 6.1 (albedo).

We fit an initial shape model with the detected 3D landmarks (with depth information) using a ridge regression [Zhu et al. 2015]. A partial texture map can be extracted by projecting the shape model onto each input image. With a predefined mask derived from landmarks for each view (see Fig. 3), the partial texture maps are then blended into a complete texture map using Laplacian pyramid blending [Burt and Adelson 1983]. The initial albedo parameters can be obtained with another ridge regression to fit the blended texture map.

5.2 Optimization

Fig. 4 shows our optimization framework. The parameters to be optimized are

\[ \mathcal{P} = \{ x_{shp}, x_{alb}, x_{light}, x_{pose} \}, \]

and \( x_{shp} \in \mathbb{R}^{500} \) is the shape parameter, \( x_{alb} \in \mathbb{R}^{199} \) is the albedo parameter, \( x_{light} \in \mathbb{R}^{27} \) is the second-order spherical harmonics lighting parameter, \( x_{pose} \in \mathbb{R}^6 \) include the rotation and translation parameters for rigid transformation. Note that we have only one \( x_{shp} \) and one \( x_{alb} \) for an user, while the number of \( x_{light} \) and \( x_{pose} \) equals to the number of views. With a set of estimated parameters and the 3DMM basis, a set of rendered RGB-D frames can be computed via a differentiable renderer [Gecer et al. 2019; Genova et al. 2018]. The distances between the rendered RGB-D frames and the input RGB-D frames can be minimized by backpropagating the errors to update parameters \( \mathcal{P} \). The loss function to be minimized is defined as:

\[
L(\mathcal{P}) = \omega_{rgb} L_{rgb}(\mathcal{P}) + \omega_{dep} L_{dep}(\mathcal{P}) + \omega_{id} L_{id}(\mathcal{P}) + \omega_{lan} L_{lan}(\mathcal{P}) + \omega_{reg} L_{reg}(\mathcal{P}),
\]

where \( L_{rgb}(\mathcal{P}) \) denotes pixel-wise RGB photometric loss, \( L_{dep}(\mathcal{P}) \) indicates pixel-wise depth loss, \( L_{id}(\mathcal{P}) \) is identity perceptual loss, \( L_{lan}(\mathcal{P}) \) represents landmark loss, and \( L_{reg}(\mathcal{P}) \) means regularization terms. Note that the landmark loss, RGB photometric loss, and regularization term are similar to conventional analysis-by-synthesis optimization approaches [Thies et al. 2016]. The identity perceptual loss is also employed in recent differentiable renderer based approaches [Gecer et al. 2019; Genova et al. 2018]. We extend these losses into multiview setting and incorporate depth data for geometric constraints. The details of each term are as follows.

**RGB Photo Loss.** The pixelwise RGB photometric loss is:

\[ L_{rgb}(\mathcal{P}) = \| I_{rgb} - \mathcal{R}(\mathcal{P}) \|^2. \]

where \( I_{rgb} \) is the input RGB image, \( \mathcal{R}(\mathcal{P}) \) is the rendered RGB image from the differentiable renderer. We adopt \( \ell_2 \) -norm because it is more robust against outliers than \( \ell_1 \) -norm.

**Depth Loss.** The depth loss is defined as:

\[ L_{dep}(\mathcal{P}) = \rho(\| I_{dep} - \mathcal{Z}(\mathcal{P}) \|_2^2), \]

where \( \rho(\cdot) \) defines a truncated \( \ell_2 \) -norm that clips the per-pixel mean squared error, \( I_{dep} \) is the input depth image, \( \mathcal{Z} \) is the rendered depth image from the differentiable renderer. The truncated function makes the optimization more robust to depth outliers.

**Identity Perceptual Loss.** To capture high-level identity information, we apply identity perceptual loss defined as

\[ L_{id}(\mathcal{P}) = \| \psi(I_{rgb}) - \psi(I_{render}) \|_2^2, \]

where \( \psi(\cdot) \) is the deep identity features exacted from a pretrained face recognition model. Here we use features from the fc7 layer of VGGFace model [Parkhi et al. 2015].

**Landmark Loss.** We define the landmark loss as the average distances between the detected 2D landmarks and projected landmarks from the predicted 3D model:

\[ L_{lan}(\mathcal{P}) = \frac{1}{|\mathcal{F}|} \sum_{f_j \in \mathcal{F}} \omega_j \| f_j - \Pi(\Phi(x_j)) \|_2^2, \]

where \( f_j \in \mathcal{F} \) are the detected landmarks, \( \Pi(\Phi(x_j)) \) denotes that the vertex \( v_j \) is rigidly transformed by \( \Phi \) and projected by camera \( \Pi \). The weighting \( \omega_j \) is to control the importance of each keypoint, where we set 50 for those located in eye, nose and mouth, while others are 1.

**Regularization.** To ensure the plausibility of the reconstructed faces, we apply regularization to shape and texture parameters:

\[ L_{reg}(\mathcal{P}) = \omega_{shp} \| x_{shp} \|_2^2 + \omega_{alb} \| x_{alb} \|_2^2, \]

where we set \( \omega_{shp} = 0.4 \) and \( \omega_{alb} = 0.001. \)
Table 1. Result comparisons are presented in Sec. 8.2. Note that our different 3DMM fitting approaches. "DR-based" stands for differentiable structures of human faces. This reminds us that we can perturb models. Different reference points may lead to different alignment results. There are no perfect reference points due to the asymmetrical structures of human faces. This reminds us that we can perturb the relative pose between two aligned models to get an “alternative” alignment. In this way, we can actually get additional samples for PCA, since the new alignments introduces new morphing targets. Furthermore, we can use a set of perturbation operations including pose perturbation, mirroring, region replacement, etc., to augment the aligned models. Based on the large amount of generated data, we propose a stochastic iterative algorithm to construct a 3DMM that compresses more capacities into lower dimensions of the basis.

Data Generation and Perturbation. Starting from the 200 aligned face shape models, our data generation and perturbation process consists of the following steps:

- **Region Replacement with Perturbation.** We first replace the nose region of each model with other models, with a rotation perturbation along the pitch angle (uniformly sampled within ±1 degree). Mouth region is also processed in the same way. For eye region, we apply replacement without perturbation. The different perturbations are empirically designed by minimizing the introduced visual defects during processing. The facial regions used in this step are shown in Fig. 5.

- **Rigid Transformation Perturbation.** We then apply rigid transformation perturbations to each face model, where the uniformly sampled range is set to: ±1 degree along yaw/pitch/roll angles for rotation, ±1% along each of the three axes for translation, ±1% for scale.

- **Mirroring.** Finally, we apply a mirroring for all the generated face models along model local coordinate system. In this way, we get over 100,000 face models in total.

#### 5.3 Morphable Model Augmentation

As the constraints incorporated in the optimization are rich, we found the expressive capacity of the linear 3DMM constructed using conventional approaches are very limited. We here present an augmentation approach to effectively boost the 3DMM capacity. Our approach is motivated by the observation that human faces are mostly not symmetrical. This will cause ambiguities when aligning face models. The reason is that during the alignment of two models, the relative rotation and translation between them is determined by minimizing the errors at some reference points on the models. Different reference points may lead to different alignment results. There are no perfect reference points due to the asymmetrical structures of human faces. This reminds us that we can perturb the relative pose between two aligned models to get an "alternative" implementation will be publicly available and can be easily configured to update parameters on 150 iterations to get the results, with a learning rate 0.05 decaying exponentially in every 10 iterations.

| Method     | Input       | Loss Term     | Optimizer   |
|------------|-------------|---------------|-------------|
| Ours       | RGB-D       | $L_{rgb}$, $L_{dep}$, $L_{id}$, $L_{lan}$, $L_{reg}$ | DR-based   |
| GANFIT     | RGB         | $L_{rgb}$, $L_{dep}$, $L_{id}$, $L_{lan}$, $L_{reg}$ | DR-based   |
| Huang et al. 2015 | RGB-D | $L_{rgb}$, $L_{dep}$, $L_{id}$, $L_{lan}$, $L_{reg}$ | Gauss-Newton |

Table 1. Different 3DMM fitting approaches. "DR-based" stands for differentiable renderer based optimizer.

Fig. 4. Our optimization framework. The parameters to be solved include: 3DMM parameters $x_{shp}$ and $x_{alb}$ for a user, lighting parameters $x_{light}$ and poses $x_{pose}$ for each view. The constraints include: landmark loss $L_{lan}$, RGB photo loss $L_{rgb}$, depth loss $L_{dep}$, and identity perceptual loss $L_{id}$.

Fig. 5. Masks for region replacement.
with 99% cumulative explained variance). The results show the expressive power of the principal components with 99% cumulative explained variance.

**Evaluation.** In order to validate the effectiveness of our technique, we design a numerical evaluation experiments with the help of BFM 2009 model [Paysan et al. 2009]. Note that the source face models in our dataset are all East Asians, while those in BFM are mostly not Asians. The domain gap between the two datasets provides us a good benchmark for cross validation (our goal is not to model cross-ethnicity fitting, but use the relative fitting errors between different versions of 3DMM to evaluate their expressive power). For each BFM basis, we compute two 3D face models using the positive and negative standard deviation values. A total of 398 BFM face models are obtained in this way. We register the BFM face models to our mesh topology using Wrap3 software [R3ds 2020]. We use the extracted PCA basis from our dataset to fit the obtained BFM face models and measure the fitting errors. Fig. 6 shows the comparison of fitting errors between the augmented basis and the original basis. If only 100 basis vectors are used, the augmented version has no advantage against the original version. As the number of basis vectors grows, the augmented version clearly outperforms the original version. Note that the maximum number of basis vectors of the original version is 200 since there are only 200 source models for 3DMM construction. For the augmented version, thousands of basis vectors could be obtained since there are over 100,000 models after augmentation. Since our iterative algorithm emphasizes the expressive power of the principal components with 99% cumulative explained variance, most of the expressive capacities are compressed into these components. In our experiments, we found the final number of the components with 99.9% cumulative explained variance in different runs is generally around 500. Thus we use 500 basis vectors through this paper. Fig. 7 shows a comparison of the facial geometries obtained using our optimization algorithm in Sec. 5.2 with different versions of PCA basis.

**Comparison to Direct PCA & Stochastic Stability.** We further compare our stochastic approach to direct PCA. Fig. 8 shows the fitting errors of the BFM face model samples with our method and direct PCA when using different number of principal components (PCs). The results show that when the number of used PCs is less than

---

**Algorithm 1: Iterative 3DMM Construction Algorithm**

```plaintext
Params: n = 1000, m = 25, Threshold
begin
Face model set S ← initial 200 models;
repeat
    Randomly sample (without replacement) a test set D with n face models from the whole dataset (over 100,000 models);
    k ← 0;
    𝜉 ← ∞;
    while 𝜉 > Threshold do
        Apply Principal Component Analysis (PCA) on S;
        Select the principal components with 99% cumulative explained variance to get the 3DMM basis Sk;
        Fit the models in D using basis Sk;
        Select the m models with largest fitting errors as set M;
        Add the the m corresponding mirrored models into M;
        𝜉 ← the median error of the m models;
        S ← S ∪ M;
        k ← k + 1;
    end
until the whole dataset is sampled;
end
Output: PCA basis Sk.
```

![Fig. 6. Mean fitting errors with two versions of basis. Note that the basis without augmentation are constructed from 200 models and thus have a maximum number of dimensions 199. The results show the expressive power of our basis is much larger than original basis.](image)

![Fig. 7. The recovered geometries with two versions of basis. The basis obtained with our method can preserve more personalized facial geometries (note the regions of facial silhouette, mouth shape, and the nose shape).](image)
1500, our approach consistently outperforms direct PCA. When the number grows up to 1500, direct PCA gradually surpasses our stochastic approach. This can be explained by the intentional design of our stochastic algorithm that emphasizes expressive capacities of the lower dimensions of the basis. Consequently, the expressive power of the components beyond the concerned dimensions are compromised. Note that the red shaded regions in the figure are the error variations of 10 different runs of our stochastic algorithm. The algorithm performs quite stably over different runs.

Relation to Localized 3DMM. There are some approaches constructing separate 3DMM for each facial region [Blanz and Vetter 1999; Neumann et al. 2013; Tena et al. 2011]. The localized 3DMMs obtain more capacities compared with global models by separating the deformation correlations between different facial regions. The region replacement augmentation in our approach is in the same spirit as localized 3DMM by explicitly generating samples with possible combinations of facial regions from different subjects. Compared with localized models, our 3DMM avoids online fusion of facial regions and thus is more efficient. Besides, our perturbation scheme and iterative 3DMM construction algorithm can be applied to localized models to improve their capacities as well. In this paper, we employ global model for efficiency consideration.

6 FACIAL REFLECTANCE SYNTHESIS

In this section, we present a hybrid approach to synthesis high-resolution albedo and normal maps. We notice that super-resolution based approaches [Lattas et al. 2020; Yamaguchi et al. 2018] cannot yield high-quality, hair-level details of the eyebrows. On the other hand, directly synthesizing high-resolution texture maps [Saito et al. 2017] may lead to overwhelming details, which also makes the rendering not realistic. Our approach addresses the problems with the help of a pyramid-based parametric representation. Fig. 9 shows the pipeline of our approach, which we explain as follows.

6.1 Regional Pyramid Bases

Fig. 10 illustrates the process to construct our regional pyramid bases. We first compute image pyramids consisting of two resolutions (512 × 512 and 2048 × 2048) for the 200 albedo maps in our dataset. We divide facial regions into 8 sub-regions indicated as the different colors in the left UV-map. The region partitioning is based on the fact that different regions have different types of skin/hair details. Denote the set of all regions as \( \mathcal{K} \). For each region \( k \in \mathcal{K} \), we construct a linear PCA-based blending model. We define each sample in our dataset as a triplet \((a_k^{512}, a_k^{2048}, g_k^{2048})\), where \(a_k^{512}\) stands for 512 × 512 albedo map of region \( k \), and \(a_k^{2048}\) and \(g_k^{2048}\) are the albedo map and normal map in 2048 × 2048 resolution. Then the triplet is vectorized into a vector format by fetching and concatenating all pixel values together from the three maps in the region. Note that during the process, the pixel indices in the three maps are recorded such that the vectorized sample can be "scatter back" into UV-map format. For each region \( k \), we apply a PCA on the the 200 vectorized samples to get the basis. Finally, the vectorized basis can be scattered back into UV-map format to obtain the blending basis \( \{A_k^{512}, A_k^{2048}, G_k^{2048}\} \) for each region, where \(A_k^{512} \in \mathbb{R}^{n_k \times 199}\) is the low-resolution albedo basis, \(A_k^{2048} \in \mathbb{R}^{16n_k \times 199}\) is the high-resolution albedo basis, \(G_k^{2048} \in \mathbb{R}^{16n_k \times 199}\) is the high-resolution normal basis, and \(n_k\) is the number of pixels within region \( k \) in the 512-resolution.

The constructed regional pyramid bases have several advantages compared with conventional bases. First, the expressive capacity is larger than global linear bases, while each region can be processed

---

^We use the "scatter\_nd" function in Tensorflow as the "scatter back" operation.
individually to accelerate the runtime. Second, the bases capture variations in both albedo and normal map. Third, the incorporated textures, followed by an unwrapping and blending to yield an initial 512 × 512 albedo-UV map. Then we use a novel regional fitting approach (Sec. 6.2) to fit the albedo map to get a high-resolution albedo UV-map. Then we use a novel regional fitting approach (Sec. 6.2) to fit the albedo map to get a high-resolution albedo UV-map. Then we use a novel regional fitting approach (Sec. 6.2) to fit the albedo map to get a high-resolution albedo UV-map.

Fig. 9. Our albedo/normal map synthesis pipeline. We first extract textures from the source images using the estimated shape and poses from Sec. 5.2. Then a model-based delighting using the estimated lighting parameters from Sec. 5.2 is applied on the extracted textures, followed by an unwrapping and blending to yield an initial 512 × 512 albedo UV-map. Then we use a novel regional fitting approach (Sec. 6.2) to fit the albedo map to get a high-resolution albedo UV-map. Then we use a novel regional fitting approach (Sec. 6.2) to fit the albedo map to get a high-resolution albedo UV-map. Then we use a novel regional fitting approach (Sec. 6.2) to fit the albedo map to get a high-resolution albedo UV-map. Then we use a novel regional fitting approach (Sec. 6.2) to fit the albedo map to get a high-resolution albedo UV-map.

Fig. 10. Construction of the regional pyramid bases. Take the eyebrow region (denoted as region $k$) as an example, the pyramid basis is constructed as follows. First, for each sample in our dataset, we fetch the pixel values in the masked region from the 512 albedo map, the 2048 albedo map, and the 2048 normal map to get $a_{A512}^{k}$, $a_{A2048}^{k}$, and $g_{A2048}^{k}$, respectively. Then they are vectorized and concatenated together to get a single sample for PCA. Note that each row in the “samples” (middle left) is a vectorized sample containing information from all three maps. Thus the PCA process simultaneously captures variations in three maps across resolutions. Finally, the vectorized PCA basis is scattered back into image format by filling the corresponding pixels back into region $k$.

6.2 Regional Fitting

Since the albedo parameters $x_{alb}$ obtained in Sec. 5.2 are based on conventional global bases, the resulting albedo maps are not satisfactory due to limited expressive power. Here we directly extract textures from the source images using the estimated shape and poses from Sec. 5.2. Then a model-based delighting using the estimated lighting parameters from Sec. 5.2 is applied on the extracted textures, followed by an unwrapping and blending to yield an initial 512 × 512 albedo map $I_{init}$. We use the 512-resolution regional bases $\{A_{512}^{k}\}_{k \in K}$ to fit the initial albedo map:

$$L(x_{alb}) = \|I_{fit}(x_{alb}) - I_{init}\|_2^2 + \omega_{tot}f_{tot}(I_{fit}(x_{alb})) + \omega_{alb}\|x_{alb}\|_2^2,$$

where $I_{fit}(x_{alb}) = \sum_{k \in K} A_{512}^{k} x_{alb}^{k}$ denotes the total variation function, $\omega_{tot} = 0.0001$ and $\omega_{alb} = 0.001$. Note that the total variation term is essential to eliminate the artifacts in the resulting albedo maps near boundaries between regions. After obtaining $x_{alb}$, we can directly compute a high-resolution albedo map $\hat{a}_{2048}$ and a normal map $\hat{g}_{2048}$ as

$$\hat{a}_{2048} = \sum_{k \in K} A_{2048}^{k} x_{alb}^{k}$$

$$\hat{g}_{2048} = \sum_{k \in K} G_{2048}^{k} x_{alb}^{k}.$$
6.3 Detail Synthesis

We adopt two refinement networks to synthesize details for albedo and normal map respectively. The refinement networks employ the architecture of a GAN-based image translation model, pix2pix [Isola et al. 2017]. As shown in Fig. 9, for albedo refinement, the network takes the fitted 2048-resolution albedo map as input and outputs a refined albedo map in the same resolution. For normal refinement, the refined albedo map and the fitted normal map are concatenated along channel dimension. The refinement network takes the concatenation as inputs and outputs a refined normal map.

During training, we first use facial region replacement and skin color transfer [Reinhard et al. 2001] to augment the 200 high-quality albedo/normal maps (from the dataset for constructing the 3DMM) into 4000 maps, which serve as ground-truth supervision for training the two networks. Then we perform regional fitting (Sec. 6.2) on the 4000 maps to get the fitted albedo/normal maps, which serve as inputs of the networks during training. We only use the facial regions out of the whole UV maps for computing training losses. Similar to pix2pix [Isola et al. 2017], we keep $L_1$ loss and GAN loss in both networks. For albedo refinement, we additionally apply total variation loss to reduce artifacts and improve skin smoothness. The weights for $L_1$, GAN and total variation losses are 100, 1, 0.001. For normal refinement, we additionally employ a pixel-wise cosine distances between the predictions and ground-truth maps to increase the accuracy of normal directions. The weights for $L_1$, GAN and cosine distance losses are 100, 1, 0.001. The networks are trained with Adam optimizer for 75000 iterations.

Relation to Existing Approaches. There are three recent CNN-based approaches that can be adopted to synthesize high-resolution facial UV-maps, which are Yamaguchi et al. [2018], GANFIT [Gecer et al. 2019], and AvatarMe [Lattas et al. 2020]. However, these approaches cannot produce satisfactory results in our case. GANFIT [Gecer et al. 2019] needs about 50 times more training data than ours to train a GAN as the nonlinear parametric representation of texture maps. In their work, the 10,000 texture maps are obtained using unwrapped photos, where shadings and specular highlights are not removed. In our system, the 200 albedo maps and normal maps are created with very high-quality artistic efforts, where shadings and specular highlights are completely removed and hair-level details are preserved. It is rather difficult to extend our data amount to theirs while keeping such high data quality. Regarding the other two approaches, we also tried super-resolution based network as in Yamaguchi et al. [2018] and pure CNN-based synthesis in AvatarMe [Lattas et al. 2020], and found the results obtained with their approaches are generally inferior to ours. We present some comparison in Sec. 8.3.

7 FULL HEAD RIG CREATION AND RENDERING

Head Completion. Although the scanned 200 models in our dataset are full head models, there are usually no reliable geometric constraints beyond facial regions for a RGB-D selfie user. We employ an algorithm to automatically complete a full head model given the recovered facial model. The regions involved in our algorithm are:

- A: facial region;
- B: back head region;
- C: intermediate region;
- D: overlapped region between A and C.

Our goal is to compute a full head shape such that region A matches the facial shape and region B matches a reference back head shape. The reason of using a reference shape for back head region B is it can further ease the difficulties to attach accessories like hair models. To this end, we construct a head morphable model using only regions B ∪ C of the 200 source models. Note that this model does not need strong expressive power as face models, thus we do not employ the technique in Sec. 5.3 but directly use PCA to extract basis. Given the recovered facial shape of a user, we apply a ridge regression similar to Sec. 5.1 to get the head 3DMM parameters, using the constraints of region B ∪ D. Then the full head model is obtained by combining the resulting shape (B ∪ C) with facial region A.

Accessories. We perform a hairstyle classification on the user’s front photo (using a MobileNet [Howard et al. 2017] image classification model trained on labeled front photos) and attach the corresponding hair model (created by artists in advance) to the head model according to the predicted hairstyle label. There are in total 30 hairs models in different hairstyles in our system (see supplementary materials). For eyeballs, we use template models and calculate the positions and scales based on reference points on the head model. For teeth, we employ an upper teeth model and a lower teeth model. The upper teeth model is placed according to reference points near nose, and it remains still when facial expression changes. The lower teeth model is placed according to reference points on the chin. When mouth opens or closes, the lower teeth model moves with the chin. Note that the accessory models are not the focus of this work, their modeling and animation can be found in dedicated research work [Bérard et al. 2016, 2019; Velinov et al. 2018; Wu et al. 2016; Zoss et al. 2019, 2018].

Expression Rigging. We adopt a simple approach similar to Hu et al. [2017] to transfer generic FACS-based blendshapes to the target model to obtain expression blendshapes. Note that our approach can be extended to further acquire user’s expression data and construct personalized blendshapes like Ichim et al. [2015].

Rendering. The recovered full head mesh model, as well as the high-quality albedo map and normal map, can be rendered with any physically based renderer. In this work, we show rendered results using Unreal Engine 4 (UE4), with a skin PBR material template simplified from the official skin PBR material template provided by...
8 RESULTS AND EVALUATION

8.1 Acquisition and Processing Time
The selfie data acquisition typically takes less than 10 seconds (200-300 frames). The total processing time after data acquisition is about 15 seconds. Note that some of the processing steps like real-time landmark detection, coarse screening, and bilateral filtering can be computed on the smartphone client while the user is taking a selfie. The preprocessed data are streamed to a server via WiFi during acquisition. The rest steps of the processing are computed on the server. Table 2 shows the runtime on our server with an Nvidia Tesla P40 GPU and an Intel Xeon E5-2699 CPU (22 cores). Note that the frame selection and expression blendshape generation are computed during data acquisition and thus do not need additional processing time. GPU and multi-thread CPU are used.

| Processing Step          | Runtime |
|--------------------------|---------|
| Landmark Detection       | ~       |
| Coarse Screening         | ~       |
| Bilateral Filtering      | ~       |
| Frame Selection          | 0.2s    |
| Initial Model Fitting    | 0.1s    |
| Initial Texture          | 0.5s    |
| Optimization             | 10s     |
| Regional Parametric Fitting | 1.5s |
| Detail Synthesis         | 1s      |
| Head Completion          | 0.5s    |
| Accessories              | 0.1s    |
| Expression Rigging       | 1s      |
| Total                    | ~15s    |

Table 2. The runtime for each step in our system. Note that the first three steps are computed during data acquisition and thus do not need additional processing time. GPU and multi-thread CPU are used.

| Avatar Creation System | Acquisition Time | Processing Time | Manual Interaction |
|------------------------|-------------------|-----------------|--------------------|
| [Ichim et al. 2015]    | 10 minutes        | ~1 hour         | 15 minutes         |
| [Cao et al. 2016]      | 10 minutes        | ~1 hour         | needed             |
| [Hu et al. 2017]       | <1 second         | ~6 minutes      | –                  |
| Ours                   | <10 seconds       | ~15 seconds     | –                  |

Table 3. Time comparison with other avatar creation systems.

N-ICP [Bouaziz et al. 2016; Weise et al. 2011], etc. Note that all the results are obtained with our 3DMM basis for fair comparisons.

Quantitative Evaluation. We use the same workflow as the production of our dataset to manually create the ground-truth models of two users. Since the ground-truth obtained in this way is very expensive, we only perform numerical evaluations on the two models to get quantitative observations. The corresponding geometries recovered from RGB-D selfie data with different variants of our approach are evaluated. The results are in Fig. 11. It can be seen from the results that our method yields the lowest mean errors, closely followed by N-ICP and the single-view variants of our method (single rgbd+id). Compared with N-ICP, our method performs better on detailed facial geometries near eyes, nose, and mouth. It conforms to our motivation that appearance constraints (photo loss and identity perceptual loss) help capture more accurate facial features.

Qualitative Evaluation. Fig. 12 shows two examples of the shape models obtained with different variants of our method. The version with multiview RGB-D data generally outperforms other variants. Besides, as shown in the figure, random initialization of our optimization can lead to flawed models. The initial fitting in Sec. 5.1 improves the system robustness. We further show results comparisons between our method with Face2Face [Hu et al. 2017; Thies et al. 2016; Yamaguchi et al. 2018], GANFIT [Gecer et al. 2019; Lattas et al. 2020], and N-ICP [Bouaziz et al. 2016; Weise et al. 2011] in Fig. 13. In general, both our method and N-ICP can reconstruct more accurate facial shapes than the methods using only RGB data (Face2face and GANFIT). This can be clearly observed from the silhouettes near the cheek region in the results. Watching more closely, our method

N-ICP [Bouaziz et al. 2016; Weise et al. 2011], etc. Note that all the results are obtained with our 3DMM basis for fair comparisons.

Quantitative Evaluation. We use the same workflow as the production of our dataset to manually create the ground-truth models of two users. Since the ground-truth obtained in this way is very expensive, we only perform numerical evaluations on the two models to get quantitative observations. The corresponding geometries recovered from RGB-D selfie data with different variants of our approach are evaluated. The results are in Fig. 11. It can be seen from the results that our method yields the lowest mean errors, closely followed by N-ICP and the single-view variants of our method (single rgbd+id). Compared with N-ICP, our method performs better on detailed facial geometries near eyes, nose, and mouth. It conforms to our motivation that appearance constraints (photo loss and identity perceptual loss) help capture more accurate facial features.

Qualitative Evaluation. Fig. 12 shows two examples of the shape models obtained with different variants of our method. The version with multiview RGB-D data generally outperforms other variants. Besides, as shown in the figure, random initialization of our optimization can lead to flawed models. The initial fitting in Sec. 5.1 improves the system robustness. We further show results comparisons between our method with Face2Face [Hu et al. 2017; Thies et al. 2016; Yamaguchi et al. 2018], GANFIT [Gecer et al. 2019; Lattas et al. 2020], and N-ICP [Bouaziz et al. 2016; Weise et al. 2011] in Fig. 13. In general, both our method and N-ICP can reconstruct more accurate facial shapes than the methods using only RGB data (Face2face and GANFIT). This can be clearly observed from the silhouettes near the cheek region in the results. Watching more closely, our method
Fig. 13. Visual comparison with state-of-the-art approaches. As pointed out by the red arrows, our method is able to generate face geometries with more accurate cheek silhouettes and more faithful mouth shapes to the input photos. In comparison, the mouth shapes obtained by N-ICP lack personalized features and are similar to each other among all the subjects. For fair comparison, all the results are obtained with our 3DMM.

Fig. 14. Examples of our synthesized albedo and normal maps.
Fig. 15. Overlay results with our synthesized albedo maps. Note that the eyebrow and mouth shapes in our albedo maps are faithful to input photos.

Fig. 16. Ablation results of intermediate or alternative steps of our albedo synthesis pipeline. The global fitting results are obtained from our 3DMM optimization in Sec. 5.2. The recovered facial structures (like the eyebrow shapes) are not faithful to inputs due to limited capacities of the global texture basis. In contrast, our regional fitting results are much more faithful. The final refinement CNN further improves the realism by adding more hair/pore details.
can recover more faithful and personalized facial shapes than N-ICP, especially near mouth region. The mouth shapes obtained with N-ICP are similar among all the faces, while our results preserve personalized mouth features and are more faithful to the photos. This can be explained by the additionally incorporated photometric loss and identity loss in our method.

**8.3 Quality of Facial Reflectance**

*Results.* Our method can produce albedo and normal maps with high-quality, realistic details while preserving major facial features of the users. Fig. 14 shows several examples of our obtained albedo and normal maps. Hair-level details in the albedo and normal maps are clearly visible. Fig. 15 shows the overlay results with synthesized albedo maps. The major facial features are consistent between the synthesized albedo maps and the input photos, especially in the eyebrow region. More importantly, the hair-level details near eyebrow region and the pore-level skin details are clearly visible. Note that the input RGB-D images in our method are in $640 \times 480$ resolution, where actual skin micro/meso-structures and hair-level details are hardly visible (see Fig. 15 left column). The synthesized skin/hair details by our approach are actually plausible hallucination, which is critical for realistic rendering of digital humans.

*Ablation Study.* Since our albedo synthesis pipeline consists of several major steps, we show the ablation results of intermediate or alternative steps of our pipeline in Fig. 16. The unwrapped textures are typically blurry and dirty due to low-resolution inputs and imperfect lighting removal. The global fitting results cannot faithfully preserve major facial structures (like the eyebrow shapes) due to limited expressive power of global texture basis. Our regional fitting can recover more accurate facial structures, while lacking hair/pore details. The final refinement CNN supplements hallucinated high-quality, realistic details to the albedo maps. Fig. 17 shows several additional close-ups of the intermediate results. Fig. 18 shows an example of the results obtained with/without delighting.
photos of Asian people. Then the reconstructed results of the 36 subjects are obtained using different approaches and rendered into realistic images using UE4. We compute the feature distances of each rendered image to all the photos in the “distracting” dataset with a face recognition network [Deng et al. 2019]. Ideally, a user’s selfie photo among all the “distracting” photos should be with the shortest distance to the rendered image. Thus, for each rendered image, the distance ranking of a user’s selfie photo among all the “distracting” photos is a good indicator of whether the rendered image is more recognizable to himself/herself or not. The average ranking percentage of our identity verification experiment are shown in Table 4 (lower rankings are better). Our results are generally more recognizable to user photos than AvatarMe and GANFIT.

Table 4. Identity verification using rendered images of the 3D reconstructions. The experiment is conducted with 36 subjects’ selfie data against more than 40,000 distracting photos. Lower ranking means the rendered images are more recognizable to user photos from the point of view of a face recognition network. For example, the average ranking percentage 4.5% means there are in average 4.5% photos in the distracting dataset that are more similar to the rendered results than the users’ own photos.

| Method       | Avg. ranking |
|--------------|--------------|
| AvatarMe [Lattas et al. 2020] | 7.0%         |
| GANFIT [Gecer et al. 2019] | 5.1%         |
| Ours         | 4.5%         |

Fig. 19. Comparison between pix2pix and pix2pix-HD. The results obtained with pix2pix-HD are generally with uneven skin colors and obvious artifacts.

Pix2pix vs pix2pix-HD. We experimented with two variants of our detail synthesis CNNs, i.e., pix2pix [Isola et al. 2017] and pix2pix-HD [Wang et al. 2018]. We found pix2pix-HD is more difficult to train and the results are generally worse than pix2pix (see Fig. 19 for an example), possibly due to the small amounts of training data.

8.4 Comparison to State-of-the-art Approaches

Visual Comparison. We compare our results to two state-of-the-art 3D face reconstruction approaches, GANFIT [Gecer et al. 2019] and AvatarMe [Lattas et al. 2020]. Fig. 20 shows two examples of the results comparison. Fig. 21 shows the rendered comparison. Our results are generally more faithful to input photos, with more accurate facial geometries and higher-quality texture maps. Note that the inputs to GANFIT/AvatarMe are front-view single images, while our results are obtained with RGB-D selfie videos.

Identity Verification. To further justify the recognizability of our results, we design a novel face verification experiment for numerical evaluation. We collect selfie data from 36 subjects and put their selfie photos into a large “distracting” face image dataset with over 40,000 photos. With pix2pix-HD are generally with uneven skin colors and obvious artifacts.

Table 4. Identity verification using rendered images of the 3D reconstructions. The experiment is conducted with 36 subjects’ selfie data against more than 40,000 distracting photos. Lower ranking means the rendered images are more recognizable to user photos from the point of view of a face recognition network. For example, the average ranking percentage 4.5% means there are in average 4.5% photos in the distracting dataset that are more similar to the rendered results than the users’ own photos.

8.5 Comparison to Model-free Reconstruction

We notice some commercial systems (e.g., Bellus3D [2020]) utilize RGB-D selfies to reconstruct static 3D face models and directly extract texture maps from input photos. Their systems commonly employ a model-free reconstruction approach like KinectFusion [Newcombe et al. 2011] and the results usually seem very faithful to input photos. However, there are several drawbacks in their results. First, as shown in Fig. 23, their reconstructed meshes are not topologically consistent and are prone to flaws. It would be difficult to attach accessories and animate them. Moreover, the extracted texture maps contain shadows and highlights, which are undesired since they cause severe unnatural issues when the rendered lighting is different from the captured lighting. In comparison, our results are high-quality and ready for realistic rendering and animation (see Fig. 23).

8.6 Robustness to Different Inputs

We conduct experiments for a user taking selfies in different lighting conditions (Fig. 22). The recovered shape and reflectance remain consistent regardless of different lighting conditions and poses. Note that there is an inherent decoupling ambiguity between skin color and illumination. The resulting skin color in the right of the figure is actually a little bit more yellowish due to the yellower input photo. However, the facial structures (like the eyebrow shape) in the resulting reflectance map are consistent.

8.7 Rendered Results

Fig. 27 shows some of our rendered results with UE4. Thanks to the high-fidelity geometry and reflectance maps, the rendered results are realistic and faithful to input faces. Note the wrinkle details on the lips, the pore-level details on the cheek, the hair-level details of the eyebrows. Fig. 24 shows two examples of our results with attached hair models, which are retrieved from our hair model database by performing hairstyle classification on the selfie photos. More results are in the supplementary video.

8.8 Limitations

Our approach does not take into account cross-ethnicity generalization. Since our 3DMM is constructed from East-Asian subjects, our evaluations are designed on the same ethnical population. We tried
Fig. 20. Visual comparison of our results to state-of-the-art approaches. Our recovered facial shapes are more faithful and recognizable. Our synthesized facial textures are with much higher quality, while there are undesired shadows/highlights or severe artifacts in the results of GANFIT and AvatarMe.

Fig. 21. Rendered comparison. Our results are more faithful to input photos, with more realistic details. More results are in the supplementary document.
The avatars created with our method are animation-ready. Animate [Yu et al. 2019] is adopted to simultaneously synthesize speech supplementary video, where a real-time multimodal synthesis system [Yu et al. 2019] is adopted to simultaneously synthesize speech and expression blendshape weights given input texts. Fig. 25 shows several snapshots of the animation. The application enables users to conveniently create high-fidelity, realistic digital humans that can be interacted with in real time. We also include another lip-sync animation result driven by speech inputs [Huang et al. 2020] in the supplementary video.

10 CONCLUSION

We have introduced a fully automatic system that can produce high-fidelity 3D facial avatars with a commercial RGB-D selfie camera. The system is robust, efficient, and consumer-friendly. The total acquisition and processing for a user can be finished in less than 30 seconds. The generated geometry models and reflectance maps are in very high fidelity and quality. With a physically based renderer, the assets can be used to render highly realistic digital humans. Our system provides an excellent consumer-level solution for users to create high-fidelity digital humans.

Future Work The animation with generic expression blendshapes are not satisfactory. We intend to extend our system to capture personalized expression blendshapes like Ichim et al. [2015]. Besides, the current system employs very simple approaches to handle accessories like hair, eyeballs, and teeth. We intend to incorporate more advanced methods to model accessories.

ACKNOWLEDGMENTS

We would like to thank Cheng Ge and other colleagues at Tencent NExT Studios for valuable discussions; Shaobing Zhang, Han Liu, Caisheng Ouyang, Yanfeng Zhang, and other colleagues at Tencent AI Lab for helping us with the videos; and all the subjects for allowing us to use their selfie data for testing.

REFERENCES

Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. 2009. The Digital Emily project: photoreal facial modeling and animation. In ACM SIGGRAPH 2009 Courses. ACM.

K Somani Arun, Thomas S Huang, and Steven D Blostein. 1987. Least-squares fitting of two 3-D point sets. IEEE Transactions on pattern analysis and machine intelligence 5 (1987), 698–700.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality single-shot capture of facial geometry. ACM Trans. Graph. (Proc. SIGGRAPH) 29, 4 (2010), 1–9.

Bellus3D. 2020. Bellus3D. Retrieved Sept 18, 2020 from https://www.bellus3d.com/.

Pascal Bérard, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. Lightweight eye capture using a parametric model. ACM Trans. Graph. (Proc. SIGGRAPH) 35, 4 (2016), 1–12.

Pascal Bérard, Derek Bradley, Markus Gross, and Thabo Beeler. 2019. Practical Person-Specific Eye Rigging. In Computer Graphics Forum, Vol. 38. Wiley Online Library, 441–454.

Volkert Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In Proc. SIGGRAPH. ACM, 187–194.

Volkert Blanz and Thomas Vetter. 2003. Face recognition based on fitting a 3d morphable model. IEEE Transactions on pattern analysis and machine intelligence 25, 9 (2003), 1063–1074.

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan PNannias, and David Dunaway. 2016. A 3d morphable model learnt from 10,000 faces. In Proc. CVPR. IEEE, 5543–5552.

Sofien Bouaziz and Mark Pauly. 2014. Semi-supervised facial animation retargeting. Technical Report.

Sofien Bouaziz, Andrea Tagliasacchi, Hao Li, and Mark Pauly. 2016. Modern techniques and applications for real-time non-rigid registration. In ACM SIGGRAPH Asia 2016 Courses. 1–25.

Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2015. Online modeling for realtime facial animation. ACM Trans. Graph. (Proc. SIGGRAPH) 32, 4 (2013), 1–10.
Fig. 23. Comparison to model-free reconstruction (e.g., Bellus3D [2020]). The model-free reconstruction are not topologically consistent and are prone to flaws, which cause difficulties when being animated. The extracted textures contain undesired shadows and highlights that would result unnatural renderings. Our results are topological consistent and ready for animation. The high-quality albedo/normal maps make our rendering very realistic.

Fig. 24. Rendering results with hair models.

Fig. 25. Snapshots of our lip-sync animation. See supplementary video.

Fig. 26. Results on other ethnicities different from our 3DMM data source. The results roughly resemble the subjects but lack ethnicity-specific features.
Fig. 27. Rendered results with UE4 rendering engine. Our method can recover faithful face models with high-quality, realistic hair/pore/wrinkle details. Note that the selfie photos are with perspective camera projection, while the rendered results are with orthogonal camera projection.
