ASM: Adaptive Skinning Model for High-Quality 3D Face Modeling

Kai Yang, Hong Shang[*] Tianyang Shi, Xinghan Chen, Jingkai Zhou, Zhongqian Sun, Wei Yang

Tencent AI Lab

{arvinkyang, hongshang, tirionshi, xinghanchen, fszhou, sallensun, willyang}@tencent.com

Abstract

The research fields of parametric face model and 3D face reconstruction have been extensively studied. However, a critical question remains unanswered: how to tailor the face model for specific reconstruction settings. We argue that reconstruction with multi-view uncalibrated images demands a new model with stronger capacity. Our study shifts attention from data-dependent 3D Morphable Models (3DMM) to an understudied human-designed skinning model. We propose Adaptive Skinning Model (ASM), which redefines the skinning model with more compact and fully tunable parameters. With extensive experiments, we demonstrate that ASM achieves significantly improved capacity than 3DMM, with the additional advantage of model size and easy implementation for new topology. We achieve state-of-the-art performance with ASM for multi-view reconstruction on the Florence MICC Coop benchmark. Our quantitative analysis demonstrates the importance of a high-capacity model for fully exploiting abundant information from multi-view input in reconstruction. Furthermore, our model with physical-semantic parameters can be directly utilized for real-world applications, such as in-game avatar creation. As a result, our work opens up new research direction for parametric face model and facilitates future research on multi-view reconstruction.

1. Introduction

A key preliminary decision factor for 3D face modeling is a proper choice of face representation, as there is no one representation that fits all. For reconstruction with abundant constraints from multiple calibrated images (high-end), high capacity in the form of raw 3D points is essential to achieve high-fidelity scans with fine-grained details within the Multi-view Stereo (MVS) framework [5, 6, 12, 13, 20]. For reconstruction with a single in-the-wild image (low-end), an intrinsically ill-posed problem, parametric face models with a strong prior are indispensable to ensure robust reconstruction with consistent topology [10, 11, 36, 18]. Reconstruction with multi-view uncalibrated images (middle-end) is a previously less explored scenario with performance on par with the low-end setting, and far behind the high-fidelity scans in the high-end setting. This suggests that the additional constraints from multi-view uncalibrated images are not fully exploited. Previous studies in this category [1, 15, 33, 3, 2] have used parametric face models interchangeably with the low-end setting. We contend that parametric face models with a higher representation capacity should be employed to accommodate extra constraints from multi-view images. Consequently, this study investigates the design of high-capacity parametric face models for reconstruction with multi-view uncalibrated images. This understudied scenario is increasingly relevant in real-world applications due to the widespread use of high-quality camera-equipped mobile phones and the need for precise reconstruction for applications such as avatar creation and facial animation.

The parametric face model is an extensively researched field. The majority of studies are based on the 3D Morphable Model (3DMM), originally introduced in the pioneering work of Blanz and Vetter [7]. Subsequent studies have continued to refine the 3DMM method by either improving the amount and diversity of data [19, 34] or proposing new methods [26, 30, 8] for dimensional reduction given such data. Simultaneously, a different trend has emerged in the game and film industries, where parametric face models are primarily represented in the form of human-designed skinning models. These models employ a set of controllable bones and skinning weights, which determine the degree to which each vertex on the mesh is influenced by the surrounding bones. This representation has demonstrated sufficient capability for extensive applications such as facial animation and avatar customization [16, 28, 29].

[*] Corresponding author.

Comparing human-designed skinning models with data-dependent 3DMMs for 3D face modeling presents an intriguing yet understudied topic. These two models are fundamentally different in terms of constraint mechanism and capacity scaling. While 3DMMs derive their constraints from data, skinning models acquire proper constraints through the design process, such as converting empirical knowledge of real faces into the placement of bones and the definitions of skinning weights. Regarding capacity scaling, 3DMMs heavily rely on the collection of facial scan data, which is prohibitively expensive to scale. In contrast, the capacity of skinning models can be easily scaled by merely adjusting the number of parameters for bones and skinning weights, making it a more cost-effective and ideal candidate for high-capacity parametric face models.

With a closer look into standard skinning models with the vanilla Linear Blend Skinning (LBS), we find that their capacity can be further improved. Standard skinning models, which typically feature hundreds of bones on tens of thousands of vertices, usually posses tens of parameters for bone position, hundreds of parameters for transformation, and millions of parameters for skinning weights. These extensive skinning weights must be determined beforehand and remain fixed during subsequent 3D face modeling. They are usually determined either by professional animators or through data-driven learning [21, 22], with certain initial estimations [4]. Since skinning weights depend on bone position, which also needs to be predefined and fixed, transformation remains the sole variable in face modeling. Within this paradigm, improving model capacity relies on increasing the number of bones or refining predefined skinning weights. We refer to these standard skinning models as Static Skinning Models (SSM). We argue that the current paradigm of SSM fundamentally limits capacity, as the critical skinning weights are fixed.

A neglected fact is that skinning weights, despite being defined in the form of a high-dimensional matrix, invariably result in low-dimensional patterns that are smooth, concentrated, and sparse. Given the strong structural nature of the human face, the movement space of each vertex is highly correlated and constrained. Consequently, skinning weights do not necessitate high-dimensional definition initially. We introduce the Adaptive Skinning Model (ASM), which defines skinning weights in a more compact form using the Gaussian Mixture Model (GMM). This new design significantly reduces the dimension of skinning weights to a level comparable with the transformation matrix. As a result, all parameters of skinning weights, transformation, and bone position can be simultaneously solved during reconstruction. This eliminates not only the labor-intensive manual design required in SSM but also the need for training data as in 3DMMs. Compared to SSM, our model can achieve a significantly increased capacity with even fewer total parameters.

The main contributions of this paper are as follows:

- A novel parametric face model is proposed, named ASM, by redefining skinning model with fully tunable parameters via introducing a more compact skinning weights representation with Gaussian Mixture Model.
- We demonstrate that ASM outperforms existing models in terms of capacity, model size, ease of implementation with arbitrary topology, and manual editing with semantic parameters. Moreover, it eliminates the need for laborious manual design and costly training data collection.
- State-of-the-art performance in 3D face reconstruction with multi-view uncalibrated images is achieved using ASM.

2. Related Work

3D Morphable Models was first proposed by Blanz and Vetter [7] as a parametric face model. They used Principal Component Analysis (PCA) to reduce a set of topology-consistent face mesh into a low-dimensional space as a set of basis representing facial shape and texture. Paysan *et al.* [24] introduced Basel Face Model (BFM), which is a widely used 3DMM in recent years, calculated from registered 3D scans from 100 male and 100 female faces. FLAME [19] became popular recently, which used 3,800 face scans to construct a shape basis and 33,000 scans to construct the expression basis. FaceScape [34] collected high-quality facial data of 938 individuals and each with 20 expressions to build 3DMM with the bilinear PCA method.

To further improve the representation capacity of 3DMM, increasing attention has been drawn into non-linear dimensionality reduction methods, especially using neural networks to train and reduce facial library to latent vector features [26, 30, 8, 35]. Ranjan *et al.* [26] introduced CoMA to extract the latent vector features from the mesh using an encoder-decoder network structure, resulting in better representations of the mesh from the training sets. Zheng *et al.* [35] proposed ImFace, which used Signed Distance Function (SDF) and implicit neural representation to model human faces, achieving impressive results. Nevertheless, either linear or non-linear 3DMM methods are data dependent, making these methods intrinsically difficult to generalize and scale, considering collecting a large number of high-quality 3D facial models is prohibitively expensive.

Skinning Model has a group of bones placed in 3D space, which can be controlled by the bones' translation, rotation, and scaling parameters. Once binding the bones with a mesh by defining the vertex-bone skinning weights matrix, the mesh can be deformed together with the bones via LBS. Skinning models have human-friendly semantic parameters, enabling the easy human design of bone placement and skinning weights. Besides, these models do not need to store basis and are computationally efficient. With these advantages, skinning models are widely used in the game and film industry for character modeling and animation of whole body and face.

Although popular in the game industry, skinning models receive less attention in 3D face modeling research. JNR [31] is the closest study to ours, which modeled face shape entirely by a skinning model with 52 bones and learned skinning weights. To the best of our knowledge, JNR is the only previous study that applied skinning models for face registration and reconstruction. Our study differs substantially from JNR in terms of design concepts and experimental findings. Firstly, JNR reduced the skinning weight matrix using a neural network, while we redesign the skinning model in a compact form in the first place, so that further dimension reduction or data-dependent learning are completely avoided, and all the parameters of skinning weights and bone positions can be freely solved online. Secondly, JNR demonstrated that skinning models achieved slightly worse capacity than state-of-the-art (SOTA) methods, such as FLAME, while our model achieves SOTA performance for both capacity and multi-view reconstruction.

3. Method

In this section, we will begin by providing a brief overview of LBS, followed by an introduction of our proposed Adaptive Skinning Model (ASM).

3.1. Linear Blend Skinning

LBS is a fundamental algorithm used for skeletal shape deformation in computer graphics [17]. It requires three types of input data: vertex data from a polygon mesh, bone transformation data in the skeleton, and skinning weight data that defines the influence of each bone on each vertex. Given a vertex $\mathbf{v} \in \mathbb{R}^3$, the LBS algorithm computes its deformed position $\mathbf{v}'$ as follows:

$$\mathbf{v}' = \sum_{j=1}^{J} w_j \mathbf{T}_j \mathbf{v} \tag{1}$$

where $\mathbf{v}$ and $\mathbf{v}'$ are in homogeneous coordinate format, w_j is the skinning weight of bone j on vertex $\mathbf{v}$ with the constraint $\sum_{j=1}^{J} w_j = 1$, $\mathbf{T}_j \in \mathbb{R}^{4\times4}$ is the bone j's transformation matrix and J is the total number of bones. In Eq. 1, the deformation is performed by $\mathbf{T}_j$ according to the following formula:

$$\mathbf{T}_j = \mathbf{M}_j^{l2w}\mathbf{M}_j^{w2l} = \mathbf{M}_p^{l2w}\mathbf{M}^{trs}(\boldsymbol{\tau}_j)\mathbf{B}_j^{-1} \tag{2}$$

where the vertex $\mathbf{v}$ is firstly projected from world space to local bone space by world-to-local transformation matrix $\mathbf{M}_j^{w2l}$ and then projected back into world space using $\mathbf{M}_j^{l2w}$. $\mathbf{M}_j^{l2w}$ can be decomposed into its parent bone's transformation matrix $\mathbf{M}_p^{l2w}$ multiply its local transformation $\mathbf{M}^{trs}(\boldsymbol{\tau}_j)$, where transformation parameters $\boldsymbol{\tau} \in \mathbb{R}^9$ include the translation, rotation, and scale parameters of the bone and $\mathbf{M}^{trs}(\cdot)$ is the composite matrix of these transformation parameters. $\mathbf{M}_j^{w2l}$ is defined as the inverse of pre-calculated bind-pose matrix $\mathbf{B}_j \in \mathbb{R}^{4\times4}$.

Based on Eq. 1 and Eq. 2, for the vanilla LBS-based skinning model, only transformation parameters $\boldsymbol{\tau}$ can be adjusted for deformation, while the skinning weights and initial bone position are fixed, which significantly limits its capacity.

3.2. Adaptive Skinning Model

To further enlarge the capacity of the vanilla LBS-based skinning model, we redesign its skinning weights and binding strategy by introducing GMM skinning weights and dynamic binding. The proposed ASM can be written as:

$$\begin{aligned} &ASM(\mathbf{v}|\boldsymbol{\zeta},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\tau}) = \\ &\sum_{j=1}^{J} W^g(\mathbf{v}|\boldsymbol{\zeta}_j,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})\mathbf{M}_p^{l2w}\mathbf{M}^{trs}(\boldsymbol{\tau}_j)B_j(F'(\boldsymbol{\zeta}))^{-1}\mathbf{v} \end{aligned} \tag{3}$$

where $W^g(\cdot)$ denotes GMM skinning weight function. $B(\cdot)$ is no longer the pre-calculated bind-pose matrix, but the standard bind-pose calculation method which takes positions and orientation in the world space of all the bones as inputs and outputs the bind-pose for each bone. $F'(\cdot)$, $\boldsymbol{\zeta}$, $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\tau}$ will be described in detail below. Fig. 1 presents an overview of our proposed model.

GMM Skinning Weights. Observing that skinning weights painted by human artists resemble a mixture of multiple Gaussian distributions, we introduce GMM to simulate the hand-painting process, so that we can build a more compact representation while maintaining strong capacity. Specifically, we define skinning weights as 2D-GMM in the unwrapped UV space.

Given the vertex $\mathbf{v}_i$ on the polygon mesh, there is a known unwrapping function $\mathbf{u}_i = F(\mathbf{v}_i)$ that maps the topology of the mesh vertex index to the UV space coordinate $\mathbf{u}_i \in \mathbb{R}^2$. The skinning weight of the point on the UV space influenced by bone j is:

$$W(\mathbf{v}|\boldsymbol{\zeta}_j,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(F(\mathbf{v})|\boldsymbol{\mu}_k + \boldsymbol{\zeta}_j, \boldsymbol{\Sigma}_k) \tag{4}$$

where $\pi_k \in \mathbb{R}$ ($\sum_{k=1}^{K} \pi_k = 1$), $\boldsymbol{\mu}_k \in \mathbb{R}^2$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^3$ are the GMM parameters, and K controls the complexity of GMM. Since $\boldsymbol{\Sigma}$ is a symmetric matrix, it has only 3 degrees of freedom. $\boldsymbol{\zeta}_j \in \mathbb{R}^2$ is the projection of the bone j onto UV space, and we use this projection as an initial guess of GMM's center. To find this projection, we firstly project the bone j with initial placement position $\boldsymbol{\psi}_j^0 \in \mathbb{R}^3$ in 3D space

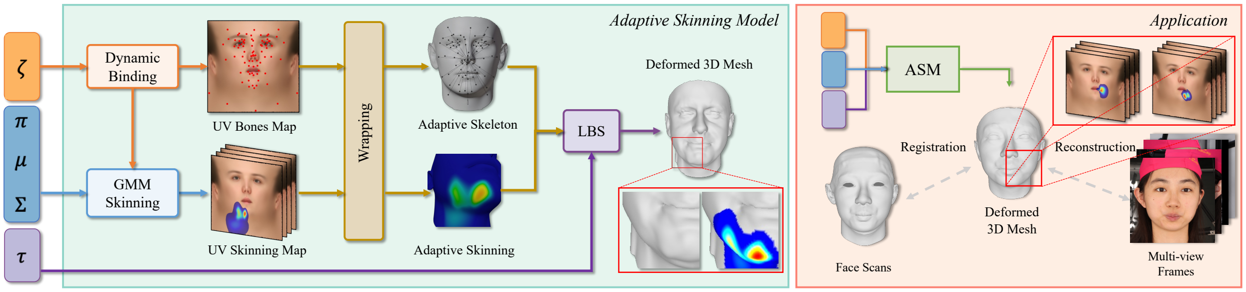

Figure 1. Illustration of Adaptive Skinning Model. The bone positions in the UV space are adjusted by the parameters $\boldsymbol{\zeta}$, which also provide an initial guess for the GMM skinning module. The parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ generate personal-specific skinning weights for each bone in the UV space, which is then wrapped into 3D space to obtain the updated skinning model. The output 3D mesh is deformed using LBS with the parameters $\boldsymbol{\tau}$. ASM can be used for tasks such as multi-view reconstruction and scan registration.

along with the z-axis (i.e. front-view) and then search the nearest vertex with index t as a proxy to obtain:

$$\zeta_j = F(\mathbf{v}_t) \tag{5}$$

For the LBS-based skinning model, all the skinning weights on vertex $\mathbf{v}$ have to add up to 1, thus we normalize 2D GMM-based skinning weights as below:

$$W^g(\mathbf{v}|\boldsymbol{\zeta}_j, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{W(\mathbf{v}|\boldsymbol{\zeta}_j, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sum_{i=1}^{J} W(\mathbf{v}|\boldsymbol{\zeta}_i, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \tag{6}$$

where J is the total number of bones. With this method, we can compress a large number of skinning weights into a few 2D GMM parameters.

Dynamic Bone Binding. In the previous GMM skinning weights calculation, $\boldsymbol{\zeta}_j$ is the UV position of the predefined bone j. Taking these estimations as the initialization and jointly optimizing $\boldsymbol{\zeta}$ with skinning weights is a straightforward way to further increase model capacity. During the joint optimization process, the gradient not only comes from $W^g(\cdot)$, but also from the bind-pose calculation $B_j(F'(\boldsymbol{\zeta}))$, where $F'(\boldsymbol{\zeta}_j)$ should be a differentiable wrapping function that maps the given UV space coordinate $\boldsymbol{\zeta}_j$ to the corresponding 3D position $\boldsymbol{\psi}_j$. Here we define this wrapping function as follows:

$$\begin{aligned} \boldsymbol{\psi}_j &= F'(\boldsymbol{\zeta}_j) = \alpha\mathbf{v}_A + \beta\mathbf{v}_B + \gamma\mathbf{v}_C - \mathbf{v}_t + \boldsymbol{\psi}_j^0 \\ \alpha, \beta, \gamma &= Barycentric(\boldsymbol{\zeta}_j, \mathbf{u}_A, \mathbf{u}_B, \mathbf{u}_C) \end{aligned} \tag{7}$$

where α, β and γ are the barycentric weights of $\boldsymbol{\zeta}_j$ with respect to the triangle f_{ABC} which $\boldsymbol{\zeta}_j$ fall within. The vertices of triangle f_{ABC} are $\mathbf{u}_A = F(\mathbf{v}_A)$, $\mathbf{u}_B = F(\mathbf{v}_B)$, and $\mathbf{u}_C = F(\mathbf{v}_C)$. $\mathbf{v}_t$ is the same vertex referred in Eq. 5 and $\boldsymbol{\psi}_j^0$ is the initial position of bone j.

Once we wrap $\boldsymbol{\zeta}$ to the 3D position $\boldsymbol{\psi}$ by vertex interpolation, we can use $B(\boldsymbol{\psi})$ to calculate the updated bind-pose matrix and evaluate the loss subsequently. As the whole process is differentiable, $\boldsymbol{\zeta}$ can be joint optimized with GMM skinning weights using backpropagation.

Up to this point, we achieve a fully parameterized representation of the LBS-based skinning model. The detailed proof process and formulas can be found in the supplemental materials.

3.3. Implementation Details.

To set up the initial placement of the bones, we use Blender[1] and place $J = 84$ bones with a hierarchical structure, which provides higher degrees of freedom than JNR [31]. We use Blender's automatic skinning weights generation method to obtain the initial skinning weights and fit our GMMs for initial parameters $\boldsymbol{\zeta}$, $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. These parameters serve as the starting point for optimization when using ASM in reconstruction tasks. For different scenarios, we suggest using different K values for the GMM model ($K = 2 \sim 5$). In total, each bone of ASM has $(11 + K * 6)$ tunable parameters. The dimension counting is shown in Tab. 1.

Parameters	$\boldsymbol{\zeta}$	$\boldsymbol{\pi}$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$\boldsymbol{\tau}$
Dimension	2	K	$K * 2$	$K * 3$	9

Table 1. Dimension of parameters for each bone.

4. Experiments

4.1. Model Characteristics

Representation capacity of parametric face models was assessed by fitting the models to 3D face scans and measuring the scan-to-mesh error. We utilized the Adam optimizer in PyTorch [23] with a learning rate of 1e-3 and 300 iterations to solve the transformation parameters of rigid ICP

[1] https://www.blender.org

and the model parameters as an optimization problem. Our error measurement adhered to the NoW-benchmark [27] prototype and was confined to the same facial region for fair comparison among models with different face coverage. We used two publicly available datasets: the LYHM dataset [9], which includes 1,212 scanned meshes of neutral faces with inconsistent topology, and a dataset from FaceScape [34], with the same setting as ImFace [35], containing 10 individuals with 20 different expressions per person, resulting in 200 total meshes with consistent topology. Note that FaceScape is not in a metrical space, hence the units of measurements on FaceScape are not in millimeters.

Methods	LYHM	FaceScape
BFM [24]	$0.372_{\pm 0.163}$	$0.462_{\pm 0.052}$
FLAME [19]	$0.246_{\pm 0.072}$	$0.341_{\pm 0.039}$
CoMA [26]	$0.756_{\pm 0.186}$	$1.088_{\pm 0.162}$
FaceScape [34]	$0.341_{\pm 0.185}$	$0.216_{\pm 0.048}$
ImFace [35]	$0.339_{\pm 0.119}$	$0.257_{\pm 0.061}$
MetaHuman [14]	$0.234_{\pm 0.089}$	$0.269_{\pm 0.063}$
Ours	$\mathbf{0.228}_{\pm \mathbf{0.072}}$	$\mathbf{0.210}_{\pm \mathbf{0.025}}$

Table 2. Scan-to-fitting error with the metric of 3D-Normalized Mean Error (NME) (mm for LYHM). (Lower is better)

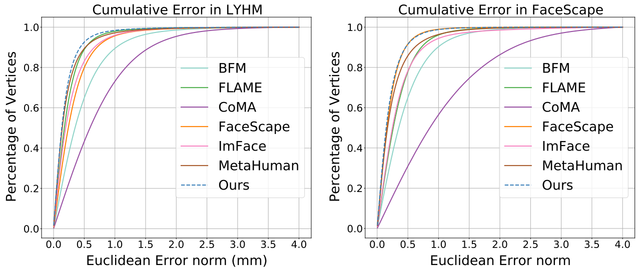

Figure 2. Scan-to-fitting cumulative error curve.

The proposed ASM was compared to widely used and SOTA parametric face models, including BFM [24] with entire 199 parameters of identity and 79 parameters of expression, FLAME [19] with entire 300 parameters of identity and 100 parameters of expression, FaceScape [34], CoMA [26], and ImFace [35]. For CoMA, we used a 64-dimensional latent vector and retrained on its datasets, considering the original 8-dimensional latent vector would limit its performance. For nonlinear 3DMM (CoMA, ImFace) the latent vector served as parameters during fitting while the weights of the decoder network were fixed. Additionally, the state-of-the-art human-designed skinning model from MetaHuman Creator [14] was also compared, which included 887 bones, far more than our model. JNR [31] was not compared as its implementation and data were not open sourced.

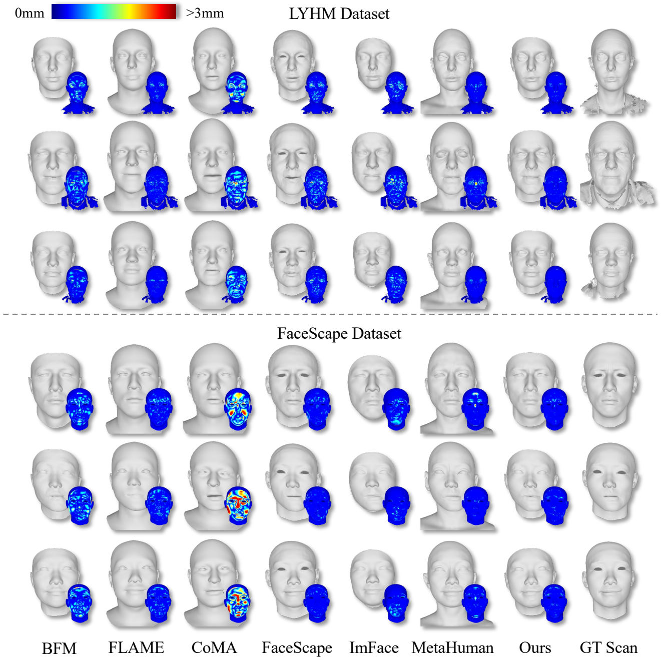

Figure 3. Exemplar fitting result. GT Scans stands for the ground truth scan used for fitting.

Results in the form of mean error with standard deviation and cumulative error curve were shown in Tab. 2 and Fig.2 respectively, with some examples shown in Fig. 3. Within the group of linear 3DMM, FLAME had the highest capacity and stable performance on both datasets. The extraordinary performance of FaceScape on its own dataset was illusive. When tested on a new dataset of LYHM, its performance dropped significantly, which illustrated the difficulty of generalization, a shared problem for all data-dependent methods. For non-linear 3DMM methods, CoMA had difficulty fitting these two datasets. ImFace behaved well on FaceScape datasets, but degraded on the LYHM dataset, similarly to FaceScape. Noted that both ImFace and FaceScape were trained using the FaceScape datasets, and both suffered from the generalization issue. Skinning models, including MetaHuman and our proposed ASM, though less studied previously, outperformed all data-dependent models. The intrinsic design of skinning models made it very cost-effective to increase capacity by simply adding more parameters. Compared to MetaHuman, the proposed ASM further improved capacity on both datasets with fewer tunable parameters, demonstrating the contribution of converting fixed skinning weights into compact and tunable skinning weights. Besides, skinning models avoided training data and the derived generalization issue, thus, leading to consistently excellent performance on both datasets.

Implementation cost is a practical consideration when adapting a face model to a new topology. It is common that different topologies are used by different groups in various applications. Off-the-shelf 3DMMs bring certain topologies, which may not be the desired ones in some applications. Adapting 3DMM to a new topology requires re-topologizing its data library and replicating the dimension reduction process, which is cumbersome for large-scale data as shown in Tab. 4. It is even impossible if the data library is not accessible considering the risk of privacy. On the other hand, MetaHuman is a sophisticated human-designed SSM with 887 bones. Adapting MetaHuman to a new topology requires tremendous domain expertise and time-consuming painting of skinning weights.

In contrast, the implementation of our model is simply determining the number of bones and placing them on a facial mesh, which can be easily replicated on any new topology. For example, 84 bones were used in this work, which took around 20 minutes in total to go through the making process. As a demonstration, our original model with the topology from BFM was duplicated twice with the topology of FLAME and topology of a game character [2]. Note the number and initial location of bones were kept the same among these three models. The representation capacity of these three models was tested on the LYHM datasets, with results shown in Tab. 3 and some examples shown in Fig. 4. Our method was robust for all different topologies.

Topology	BFM	FLAME	GAME
3D-NME↓	$0.228_{\pm 0.072}$	$0.236_{\pm 0.029}$	$0.235_{\pm 0.063}$

Table 3. Representation capacity of ASM with different topology.

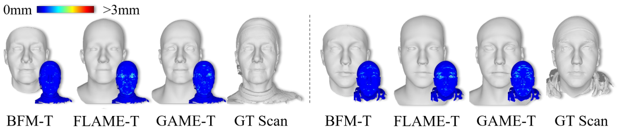

Figure 4. Exemplar fitting results of ASM with different topologies. BFM-T, FLAME-T, and GAME-T stand for the topology of BFM, FLAME, and a game character respectively.

Model size refers to the disk space required to store the model, which is divided into the fixed part and headcount proportional part. The fixed part comes from the 3DMM basis, weights of neural networks, and predefined skinning weights. The headcount proportional part comes from 3DMM parameters, feature vectors of the neural networks, and skinning model tunable parameters. As shown in Fig. 5, our model size is significantly lower than all other models, especially within the range of 100 faces, which is a common range for real-world applications. This makes our model advantageous for mobile device applications.

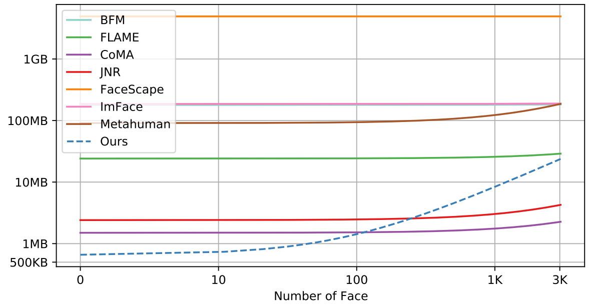

Figure 5. Model size as a function for storing the number of faces.

Methods	CPU	GPU	Dim.	Data
BFM [24]	0.082s	0.007s	278	200
FLAME [19]	0.028s	0.002s	406	3,800
CoMA [26]	1.880s	0.012s	64	12
FaceScape [34]	30.661s	0.034s	351	938
ImFace [35]	94.660s	20.816s	256	355
MetaHuman [14]	0.489s	0.007s	7,983	-
Ours	2.658s	0.066s	1,932	1

Table 4. Statistics of different face models. CPU and GPU refer to inference time measured on CPU or GPU. Dim refers to the dimension of parameters. Data refers to the number of individuals used to construct the face model.

Inference time refers to the time it takes to generate a face mesh given the input parameters. Inference time measurement was conducted with a batch size of 32 and averaged over 1,000 repetitions. It was measured on either CPU of Intel(R) Xeon(R) Gold 6133 CPU @ 2.50GHz or the GPU of NVIDIA Tesla V100 32G. As shown in Tab. 4, ASM was slower compared to linear 3DMM (BFM and FLAME) and SSM (Metahuman), but still within an acceptable range. ImFace with a much longer inference time increases the difficulty of being used.

4.2. Model Application

3D face reconstruction with multi-view uncalibrated images was evaluated. The Florence MICC benchmark is widely used for multi-view 3D face reconstruction with three subsets (Coop, indoor, and Outdoor). The Coop and Indoor subsets have video segments of 53 individuals with stable indoor lighting, differing by camera distance, portrait distance for Coop, and roof camera for Indoor. Coop is closer to our targeted setting with high-quality images, and both were used in our evaluation. For each video segment, we manually selected 15 frames at different angles

[2] We obtain the mesh file from the open game mods community: https://steamcommunity.com/sharedfiles/filedetails/?id=2326367687

with close expressions. Reconstruction was solved as an optimization problem with our proposed face model and photometric consistency constraints [1, 15]. A learning-based method [10] was used to serve as initialization to accelerate the convergence of optimization. For detailed experimental settings and energy function, please refer to the supplementary materials. As shown in the Tab. 5, We achieved SOTA performance on the Florence MICC Coop benchmark. For the Indoor benchmark with video taken in the distance, which is out of our targeted setting, methods with the advantage of robustness behave better, such as [32].

Methods	Coop↓	Indoor↓
Piotraschke and Blanz [25]	1.68	1.67
Deng *et al.* [10]	1.60	1.61
Wood *et al.* [32]	1.43	**1.42**
Ours	**1.34**	1.53

Table 5. Multi-view reconstruction error with metric of 3D-RMSE(mm) on Florence MICC benchmark. (↓Lower is better.)

The MICC benchmark does not accurately represent our intended setting due to the allowance of speech and facial expression changes during video collection. To address this limitation, we conducted further evaluations on the FaceScape dataset, which captured a large number of high-definition images synchronously using a camera rig. Calibration information of this dataset was not used, and we randomly selected 3, 5, 10, and 20 images from 10 subjects to conduct multi-view 3D face reconstruction using various models, including BFM, FLAME, ASM-K2, ASM-K5, and MetaHuman, while maintaining consistent settings as previously stated. ASM-K2 and ASM-K5 referred to our model with different parameter K settings, with ASM-K2 being the default setting used in all other experiments. Additionally, we also compared with MVS implemented by commercial software MetashapePro[3].

Tab. 6 and Fig. 6 demonstrated that skinning models, including ours and MetaHuman, outperform 3DMM (BFM and FLAME) in the multi-view setting. Skinning models can continuously improve results with more views, while 3DMM exhibited a less noticeable improvement. This highlighted the importance of using skinning models with higher capacity to accommodate more constraints from multi-view input. MVS failed with only 3 or 5 images, but achieved high-fidelity results with 20 images, as expected. While MetaHuman results exhibited bizarre shapes, our model achieved natural and high-fidelity results. This can be attributed to the fact that MetaHuman adds extra bones, far beyond the physical number of joints on human face. As a result, the added capacity may not align well with the actual human face, resulting in an unnatural appearance. In contrast, our proposed model increases capacity in a more balanced manner by allowing all skinning model parameters to be tuned simultaneously, leading to a better representation of human face.

Images	BFM	FLAME	ASM-K2	ASM-K5	MetaHuman	MVS
3	1.64	1.56	1.30	**1.29**	1.47	-
5	1.56	1.54	1.06	**1.06**	1.34	-
10	1.52	1.48	0.94	0.92	1.15	**0.88**
20	1.50	1.33	0.86	0.84	1.04	**0.55**

Table 6. Multi-view reconstruction error with metric of 3D-RMSE on selected FaceScape dataset. (↓Lower is better.)

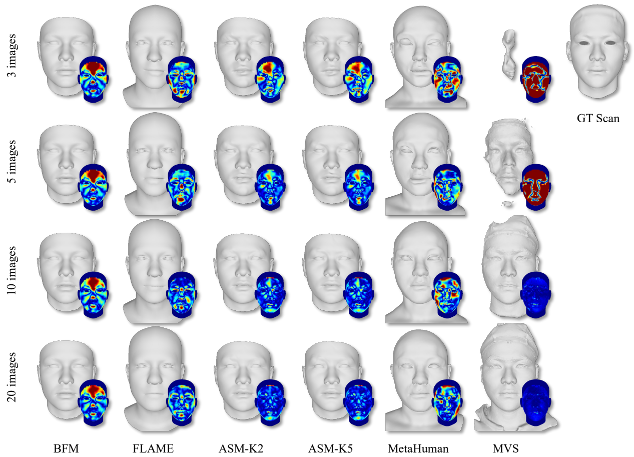

Figure 6. Multi-view reconstruction result on FaceScape.

We also obtained in-house data by capturing 6 images with a high-quality mobile camera and requesting participants to remain stationary. Using the same set-up, we performed multi-view 3D face reconstruction and compared our model to FLAME and MVS. Our model outperformed FLAME with more identifiable results, as shown in Fig. 7, while MVS failed. These findings demonstrate that our model is the proper choice for reconstruction with multi-view uncalibrated images, especially when the number of images is not adequate for successful MVS.

In-game avatar creation is another application benefiting from the proposed model, which is to customize in-game avatars given input images. Character's face is mostly represented in the form of skinning models with certain topology in games [28, 29]. Our model belongs to skinning models and can be easily adapted to new topology, therefore, the reconstruction results of our model can be directly transferred into the game system without a performance drop. The implementation of reconstruction had the same setting as above, except the model was based on the topology from the game, as previously illustrated in Fig. 4. As shown in Fig. 8, in-game avatar from reconstruction result was

[3] https://www.agisoft.com

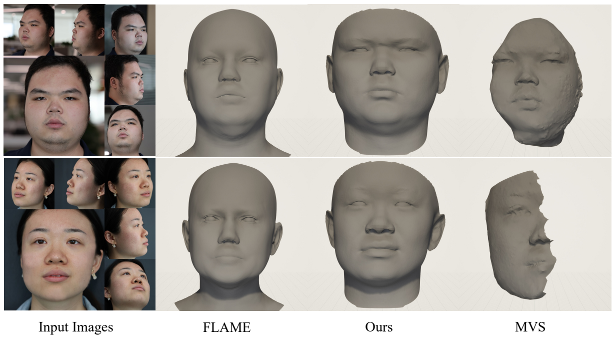

Figure 7. Multi-view reconstruction result on in-house data.

achieved, and post-editing was allowed, due to the advantage of skinning models with physical-semantic parameters.

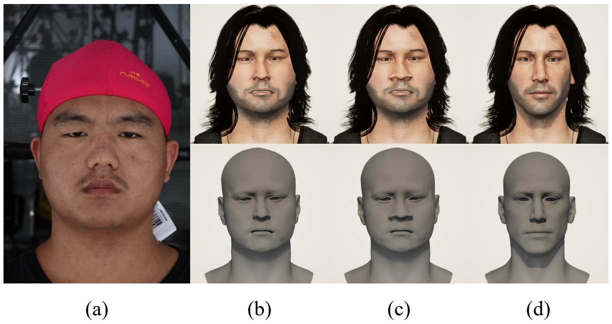

Figure 8. (a) exemplar image out of 5; (b) customized avatar with reconstruction result; (c) avatar with further manual edit, for example, adjusting the bone of the nose wing; (d) the original avatar.

4.3. Ablation Studies

A quantitative ablation study was conducted to investigate the key design components for fitting and reconstruction performance using the LYHM and MICC datasets, respectively. The study compared the following methods. SSM referred to a static skinning model with fixed bone binding and skinning weights provided by Blender. DBB referred to dynamic bone binding with the bone position as tunable variable. GSW referred to tunable GMM based skinning weights. RD referred to replacing the initial skinning weights provided by Blender with random ones. In Tab. 7, the last row represented the default setting as in previous evaluations. Results indicated that SSM had a higher representation capacity than most 3DMM models, except FLAME, leading to improve multi-view reconstruction performance. Converting bone location and skinning weights into tunable parameters further improved capacity. Careful consideration was required for the initialization of GMM skinning weight.

Additionally, a qualitative ablation study was conducted to illustrate the facial prior of ASM. Unlike 3DMMs learning constraints from data, ASM, as well as skinning models in general, encodes proper constraints within the design of initial bone placement and skinning weights. Therefore, random bone placement or skinning weights lead to failed modeling, as shown in Fig. 9(a) and Fig. 9(b) respectively. Besides, regularization terms can be used to provide additional constraints for local facial regions, thanks to the explicit semantics of skinning model parameters. For instance, enlarging the regularization weight for bones near the eyebrows can reduce the impact of noise for that region, as shown in Fig. 9(c).

SSM	DBB	GSW	RD	Registration	Reconstruction
✓				$0.322_{\pm 0.118}$	$1.36_{\pm 0.48}$
✓	✓			$0.282_{\pm 0.094}$	$1.36_{\pm 0.46}$
✓	✓	✓	✓	$0.416_{\pm 0.107}$	$1.47_{\pm 0.45}$
✓	✓	✓		$\mathbf{0.228}_{\pm \mathbf{0.072}}$	$\mathbf{1.34}_{\pm \mathbf{0.51}}$

Table 7. Ablation study on registration (with metric of 3D-NME) and reconstruction (with metric of 3D-RMSE).

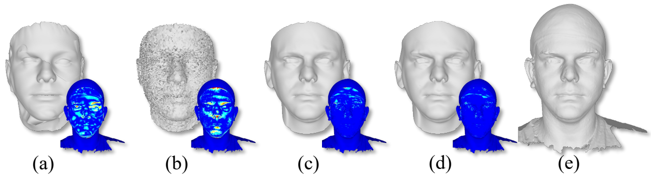

Figure 9. Illustration of fitting results. (a) random bones placement; (b) random skinning weights; (c) enlarged regularization weight on the eyebrow area; (d) default setting of ASM; (e) GT scans.

5. Discussion

This study demonstrated that parametric face models have varying characteristics and should be tailored for specific applications. When dealing with low-quality input, such as the MICC Indoor benchmark, 3DMM with strong prior achieved robust and SOTA performance. For high-quality calibrated input captured within a camera rig, parametric face models were unnecessary, and MVS with raw 3D points achieved high-quality facial scans, considered as the ground truth. For intermediate-level input of high-quality multi-view but uncalibrated images, skinning models based ASM with higher capacity achieved SOTA performance on MICC Coop benchmark, FaceScape (without calibration), and our in-house data. Compared to a sophisticated human-designed static skinning model, ASM with fully tunable parameters can further improve capacity in a more natural and effective way.

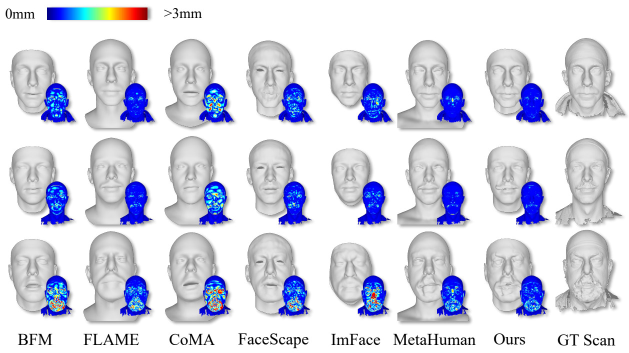

Figure 10. Exemplar failed fitting results. GT Scans stands for the ground truth scan used for fitting.

This study did not cover other aspects of multi-view reconstruction, such as the design of constraints or optimization process. We believe that our proposed model with higher capacity will facilitate future research on multi-view reconstruction, enabling better use of increased capacity to improve reconstruction performance. Another potential future work is to explore decoupling the identity and expression of the skinning parameters to enable expression transfer between different individuals and customization of personal-specific expressions. It would be interesting to combine data dependent decoupling techniques as used in 3DMMs with skinning models. Moreover, note that this study did not fully investigate the unique case of faces with beards. Given that beards are not part of the facial topology, they pose a significant challenge for all parametric face models. This is particularly true for models with higher capacity, such as ASM, which is more prone to artifacts, as illustrated in Fig. 10.

6. Conclusion

We proposed ASM, a high-capacity parametric face model, to be used for reconstruction with multi-view uncalibrated images. ASM offers stronger capacity than data-dependent 3DMMs with compact and fully tunable parameters. Our experiments demonstrated that ASM achieved SOTA performance for multi-view reconstruction on the MICC Coop benchmark, and its high capacity was crucial to exploit abundant information from multi-view input. Furthermore, the semantic parameters of ASM made it suitable for real-world applications like in-game avatar creation. The study opens up new research direction for the parametric face model and facilitates future research on multi-view reconstruction.

7. Acknowledgment

We would like to thank Jiawen Zheng and Bishan Wang for their assistance in the data collection process.

References

[1] Brian Amberg, Andrew Blake, Andrew Fitzgibbon, Sami Romdhani, and Thomas Vetter. Reconstructing high quality face-surfaces using model based stereo. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[2] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3d face reconstruction via in-network optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6216–6225, 2021.

[3] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5850–5860, 2020.

[4] Ilya Baran and Jovan Popović. Automatic rigging and animation of 3d characters. *ACM Transactions on graphics (TOG)*, 26(3):72–es, 2007.

[5] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9. 2010.

[6] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011.

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[8] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7213–7222, 2019.

[9] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128:547–571, 2020.

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[11] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[12] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[13] Graham Fyffe, Koki Nagano, Loc Huynh, Shunsuke Saito, Jay Busch, Andrew Jones, Hao Li, and Paul Debevec. Multi-view stereo on consistent face topology. In *Computer Graphics Forum*, volume 36, pages 295–309. Wiley Online Library, 2017.

[14] Epic Games. Metahuman creator. *Available: https://www.unrealengine.com/en-US/metahuman-creator*, 2021.

[15] Matthias Hernandez, Tal Hassner, Jongmoo Choi, and Gerard Medioni. Accurate 3d face reconstruction via prior constrained structure from motion. *Computers & Graphics*, 66:14–22, 2017.

[16] Doug L James and Christopher D Twigg. Skinning mesh animations. *ACM Transactions on Graphics (TOG)*, 24(3):399–407, 2005.

[17] Ladislav Kavan. Part i: direct skinning methods and deformation primitives. In *ACM SIGGRAPH*, volume 2014, pages 1–11, 2014.

[18] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. Robust model-based face reconstruction through weakly-supervised outlier segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 372–381, 2023.

[19] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[20] Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. Topologically consistent multi-view face inference using volumetric sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3824–3834, 2021.

[21] Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[22] Xiaoyu Pan, Jiancong Huang, Jiaming Mai, He Wang, Honglin Li, Tongkui Su, Wenjun Wang, and Xiaogang Jin. Heterskinnet: A heterogeneous network for skin weights prediction. In *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, volume 4. Association for Computing Machinery, 2021.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[24] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[25] Marcel Piotraschke and Volker Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3418–3427, 2016.

[26] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018.

[27] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019.

[28] Tianyang Shi, Yi Yuan, Changjie Fan, Zhengxia Zou, Zhenwei Shi, and Yong Liu. Face-to-parameter translation for game character auto-creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 161–170, 2019.

[29] Tianyang Shi, Zhengxia Zuo, Yi Yuan, and Changjie Fan. Fast and robust face-to-parameter translation for game character auto-creation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1733–1740, 2020.

[30] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018.

[31] Noranart Vesdapunt, Mitch Rundle, HsiangTao Wu, and Baoyuan Wang. Jnr: Joint-based neural rig representation for compact 3d face modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 389–405. Springer, 2020.

[32] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 160–177. Springer, 2022.

[33] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2019.

[34] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 601–610, 2020.

[35] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20343–20352, 2022.

[36] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 250–269. Springer, 2022.