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

Kai Yang, Hong Shang, Tianyang Shi, Xinghan Chen, Jingkai Zhou,
Zhongqian Sun and Wei Yang

Tencent
Shenzhen, China

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

Abstract

The research fields of parametric face models 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 directions for the parametric face models and facilitates future research on multi-view reconstruction.

1. Introduction

In the field of 3D face modeling, there have been extensive studies achieving satisfying performance for high-end applications with the setting of camera rig [4, 5, 17, 11] and low-end applications with a single in-the-wild image [9, 10, 33]. However, middle-end settings, such as inputting multi-view high-quality images of people standing still, are less explored. Its inputs are more curated than the low-end setting, although they are still uncalibrated, unlike the high-end one. The reconstruction performance of the middle-end setting is on par with the low-end setting and far behind the high-fidelity scan in the high-end setting [30, 2], meaning that the more curated input is not fully exploited. Such an understudied scenario has increasing real-world demand with the widespread use of mobile phones with high-quality cameras and the need for precise reconstruction, such as avatar creation and facial animation.

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 low-end applications with noisy and insufficient images, an intrinsically ill-posed problem, parametric face models with strong prior are crucial to guarantee robust and stable reconstruction with consistent topology [9, 10, 33]. For high-end applications with abundant constraints from multiple calibrated images, high capacity in the form of raw vertices is essential to achieve high-fidelity scans with fine-grained details within the Multi-view Stereo (MVS) framework. Compared to the low end, an ideal representation for middle-end reconstruction would raise the need for capacity and reduce the need for prior. However, previous studies of multi-view reconstruction use parametric face models interchangeably as in low-end applications [13, 1]. We argue that face representation needs to be tailored for middle-end applications with more attention on representation capacity.

The parametric face model is an extensively researched field. The majority of studies are based on the 3D Morphable Face Model (3DMM), which was originally introduced in the pioneering work of Blanz and Vetter [6]. The following studies continue improving the 3DMM method by either improving the amount and diversity of data or proposing new methods for dimensional reduction given such data. Meanwhile, a different trend exists in the game

and film industries where the parametric face models are mainly 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 shown sufficient capability for extensive applications such as facial animation and avatar customization [14, 25, 26]. Comparing human-designed skinning models with data-based 3DMM for 3D face modeling is an interesting yet understudied topic.

In this study, we investigate the design of parametric face models, with a particular focus on the previously less explored middle-end 3D face modeling. A parametric face model with high capacity is desired to accommodate extra constraints from multiple images. The capacity of 3DMM heavily relies on the collection of facial scan data, which is prohibitively expensive. On the contrary, the capacity of human-designed skinning models can be easily adjusted by simply tuning the number of parameters for bones and skinning weights. Thus it is much more cost-effective to increase capacity with skinning models, making it an ideal candidate for middle-end application.

With a closer look into standard skinning models with the vanilla Linear Blend Skinning (LBS), we find its capacity can be even further improved. Standard skinning models, for example with hundreds of bones on tens of thousands of vertices, usually have tens of parameters for bone position, hundreds of parameters for transformation, and millions of parameters for skinning weights. Such a large number of skinning weights have to be determined beforehand and remain fixed in following 3D face modeling. They are usually determined either by professional animators or via learning from data [18, 19] with certain initial estimation [3]. Since skinning weights relies on the bone position, which also has to be predefined and fixed, thus leaving transformation as the only variable in face modeling. Within this paradigm, improving model capacity relies on increasing the number of bones or improving 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, though defined with the form of a high dimensional matrix, always result in low dimensional patterns, being smooth, concentrated, and sparse. As the human face is strongly structured, the movement space of each vertex is highly correlated and restricted. So skinning weights does not need to be defined in high dimension in the first place. We propose the Adaptive Skinning Model (ASM), which defines skinning weights in a more compact form by Gaussian Mixture Model (GMM). This new design significantly reduces the dimension of skinning weights to be on par with the transformation matrix. Thus, all parameters of skinning weights, transformation, and bone position can be solved simultaneously. Compared to SSM, our model achieve significantly increased capacity with even fewer total parameters. Additionally, ASM can be easily replicated with arbitrary topology by eliminating both the need for scan data as in 3DMM and data-driven skinning model and the need for laborious manual design as in SSM.

The main contributions of this paper are as follows:

- A novel parametric face model is proposed, named ASM, which outperforms existing models in terms of capacity, model size, easy implementation with arbitrary topology, and manual editing with semantic parameters.
- We redefine the skinning model with fully tunable parameters by introducing a more compact skinning weights representation with Gaussian Mixture Model.
- We demonstrated that the new model can be applied for multi-view 3D face reconstruction achieving state-of-the-art performance, and also in-game avatar creation.

2. Related Work

3D Morphable Models was first proposed by Blanz and Vetter [6] 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*. [21] 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 [16] become popular in recent years, which used 3,800 face scans to construct a shape basis and 33,000 scans to construct the expression basis. FaceScape [31] 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 [23, 27, 7, 32]. Ranjan *et al*. [23] 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*. [32] proposed ImFace, which uses 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 for the whole body and face.

Although popular in the game industry, skinning models receive less attention in 3D face modeling research. JNR [28] is the closest study to ours, which models 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 applies 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 redesigned 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 learned online. Secondly, JNR demonstrated that skinning models achieved slightly worse capacity than state-of-the-art (SOTA) methods, such as FLAME, while our model achieved SOTA performance for both capacity and multi-view reconstruction.

3. Method

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

3.1. Linear Blend Skinning

LBS is a fundamental algorithm used for skeletal shape deformation in computer graphics [15]. 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 $\boldsymbol{\tau} \in \mathbb{R}^9$ includes 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 to perform skeletal animation, while the skinning weights and initial bone position are frozen during animation, which significantly limits its capacity in 3D modeling.

3.2. Adaptive Skinning Model

The goal of this work is to improve the parametric face model for middle-end applications, where capacity is our key consideration. 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 Adaptive Skinning Model (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 framework.

GMM Skinning Weights. Since skinning weights on the vertices are independent of each other, it is difficult to directly parameterize skinning weights, especially without training data. Observing that skinning weights painted by human artists resemble a mixture of multiple Gaussian distributions, we introduce Gaussian Mixture Model (GMM) to simulate the hand-painting process, so that we can build a more compact representation while maintaining strong capacity. Considering that the UV space can fully utilize the structural information among mesh vertices, we calculate 2D-GMM skinning weights based on the unwrapped UV map.

Given the vertex $\mathbf{v}_i$ on the polygon mesh, there exists a known unwrapping function $\mathbf{u}_i = F(\mathbf{v}_i)$ that maps the topology of the mesh vertices indices to the UV map coordinate $\mathbf{u}_i \in \mathbb{R}^2$. The skinning weight of the point on the

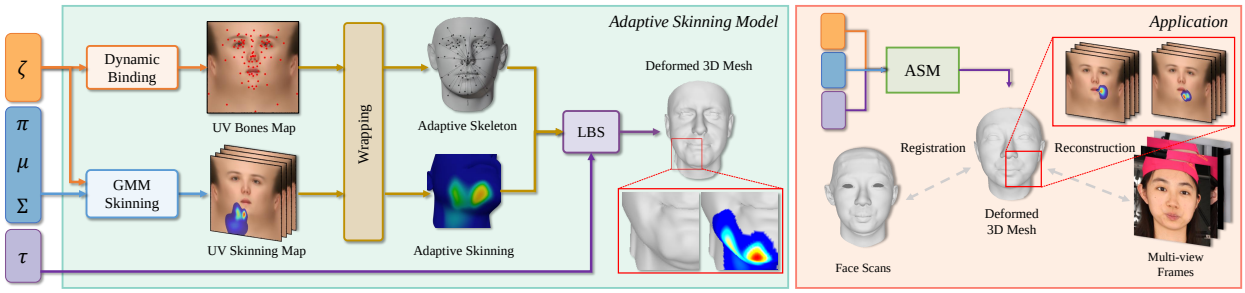

Figure 1. Illustration of Adaptive Skinning Model. The bone positions in the UV space are adjusted by the parameters $\boldsymbol{\zeta}$, which also provide an initial guess for the GMM skinning module. The parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ generate personal-specific 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.

UV map influenced by bone j is:

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

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. $\boldsymbol{\zeta}_j \in \mathbb{R}^2$ 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 along with the z-axis (i.e. front-view) and then search the nearest vertex with index t as a proxy to obtain:

$$\boldsymbol{\zeta}_j = F(\mathbf{v}_t) \quad (5)$$

For the LBS-based skinning model, it is necessary to ensure that all the skinning weights on vertex $\mathbf{v}$ are normalized to 1, thus we combine 2D GMM-based skinning weights among all bones according to the following rule:

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

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 and introduce the UV spatial prior as a constraint.

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 map coordinate $\boldsymbol{\zeta}_j$ to the corresponding 3D position $\boldsymbol{\psi}_j$. Here we define this wrapping function as follows:

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

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 used Blender[1] and placed $J = 84$ bones with a hierarchical structure, which provides higher degrees of freedom than JNR [28]. We used 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, which is sufficient for complete 3D model reconstruction. The dimension counting is shown in Tab. 1.

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

Parameters	ζ	π	μ	Σ	τ
Dimension	2	K	$K*2$	$K*3$	9

Table 1. Dimension of the parameters for each bone to be solved. Since Σ is a symmetric matrix, it has only 3 degrees of freedom.

4. Experiments

4.1. Model Characteristics

Representation capacity of parametric face models is a crucial feature for our middle-end application. The capacity is assessed by fitting the models to 3D face scans and measuring the scan-to-mesh error. We utilized the Adam optimizer in PyTorch [20] with a learning rate of 1e-3 and 300 iterations to solve the transformation parameters of rigid ICP and the model parameters as an optimization problem. Our error measurement adhered to the NoW-benchmark [24] 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 [8], which includes 1,212 scanned meshes of neutral faces with inconsistent topology, and a dataset from FaceScape [31], with the same setting as ImFace [32], containing 10 individuals with 20 different expressions per person, resulting in 200 total meshes with consistent topology.

Methods	LYHM	FaceScape
BFM [21]	$0.372_{\pm 0.163}$	$0.462_{\pm 0.052}$
FLAME [16]	$0.246_{\pm 0.072}$	$0.341_{\pm 0.039}$
CoMA [23]	$0.756_{\pm 0.186}$	$1.088_{\pm 0.162}$
FaceScape [31]	$0.341_{\pm 0.185}$	$0.216_{\pm 0.048}$
ImFace [32]	$0.339_{\pm 0.119}$	$0.257_{\pm 0.061}$
MetaHuman [12]	$0.234_{\pm 0.089}$	$0.269_{\pm 0.063}$
Ours	$\mathbf{0.228_{\pm 0.072}}$	$\mathbf{0.210_{\pm 0.025}}$

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

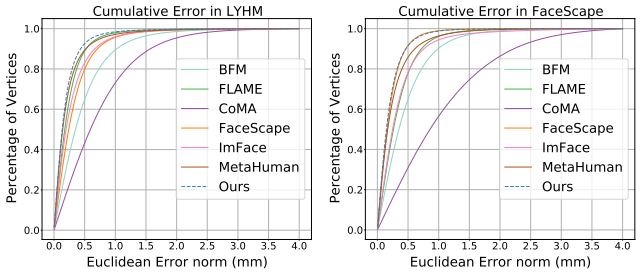

Figure 2. Scan-to-fitting cumulative error curve.

The proposed ASM is compared to widely used and

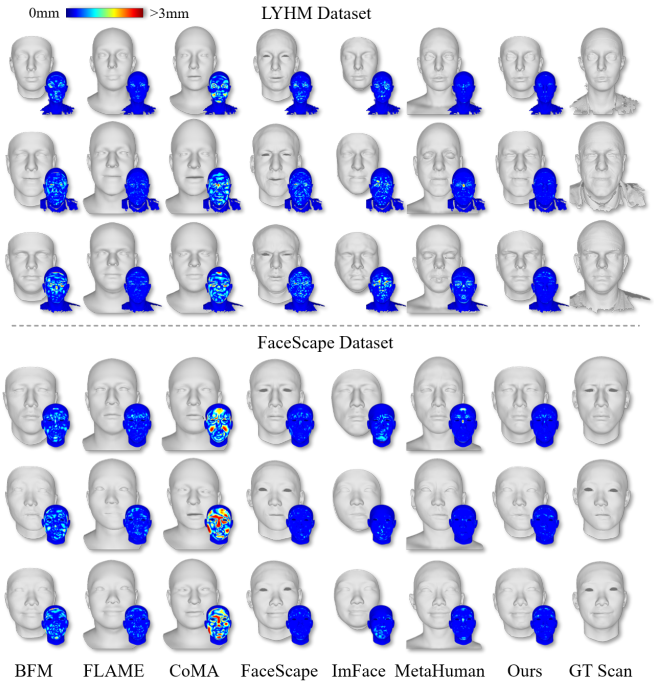

Figure 3. Exemplar fitting result. GT Scans stand for the ground truth scan used for fitting.

SOTA parametric face models, including BFM [21] with entire 199 parameters of identity and 79 parameters of expression, FLAME [16] with entire 300 parameters of identity and 100 parameters of expression, FaceScape [31], CoMA [23], and ImFace [32]. For CoMA, we used a 64-dimensional latent vector and retrain on its public datasets, considering the original 8-dimensional latent vector will limit its performance. For nonlinear 3DMM (CoMA, ImFace) the latent vector serves as parameters during fitting while the weights of the decoder network are fixed. Additionally, the state-of-the-art human-designed skinning model from MetaHuman Creator [12] is also compared, which included 887 bones, far more than our model. JNR [28] was not compared as its implementation and data were not published.

Results in the form of mean error with standard deviation and cumulative error curve are shown in Tab. 2 and Fig.2 respectively, with some examples shown in Fig. 3. Within the group of linear 3DMM, FLAME has the highest capacity and stable performance on both datasets. The extraordinary performance of FaceScape on its own dataset is illusive. When tested on a new dataset of LYHM, its performance dropped significantly, which illustrates 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 behaves well on FaceScape datasets, but degrades on the LYHM dataset,

similarly to FaceScape. Noted that both ImFace and FaceScape were trained using the FaceScape datasets, and both suffer 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 makes it very cost-effective to increase capacity by simply adding more parameters. Compared to the 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 avoid 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 3DMM brought 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 are used in this work, which takes around 20 minutes in total to go through the making process. As a demonstration, our original model with the topology from BFM is duplicated twice with the topology of FLAME and topology of a game character [2]. Note the number and initial location of bones are kept the same among these three models. The representation capacity of these three models is tested on the LYHM datasets, with results shown in Tab. 3 and some examples shown in Fig. 4. Our method is 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.

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

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

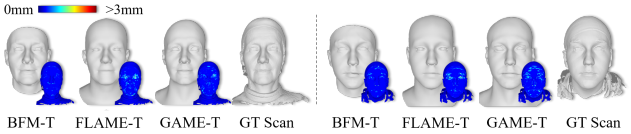

Figure 4. Exemplar fitting results of ASM with different topologies. BFM-T, FLAME-T, and GAME-T stand for the topology of BFM, FLAME, and a game character respectively.

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.

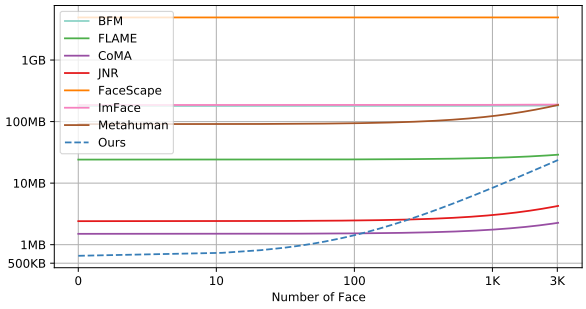

Figure 5. Model size as a function for storing the number of faces.

Methods	CPU	GPU	Dim.	Data
BFM [21]	0.082s	0.007s	278	200
FLAME [16]	0.028s	0.002s	406	3,800
CoMA [23]	1.880s	0.012s	64	12
FaceScape [31]	30.661s	0.034s	351	938
ImFace [32]	94.660s	20.816s	256	355
MetaHuman [12]	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, the proposed ASM is 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 middle-end setting is our targeted application, which refers to multi-view high-quality uncalibrated images of people staying still. There is no public dataset directly in line with this setting. 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, and both were used in our evaluation. For each video segment, we manually selected 15 frames at different angles with close expressions.

Multi-view 3D face reconstruction is solved as an optimization problem with our proposed face model and photometric consistency constraints [13, 1]. A learning-based method [9] is used to serve as initialization to accelerate the convergence of optimization. For detailed experimental settings, 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 should be used, such as [29].

Methods	Coop↓	Indoor↓
Piotraschke and Blanz [22]	1.68	1.67
Deng *et al*. [9]	1.60	1.61
Wood *et al*. [29]	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 captures a large number of high-definition images synchronously using a camera rig. Calibration results were dropped, 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 refer to our model with different parameter K settings, with ASM-K2 being the default setting used in all other experiments. Additionally, we compared the results of our approach to MVS implemented by photogrammetry software.

Tab. 6 and Fig. 6 demonstrate 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 exhibits a less noticeable improvement. This highlights the importance of using skinning models with higher capacity to accommodate more constraints from multi-view input. MVS fails with only 3 or 5 images, but achieves high-fidelity results with 20 images, as expected. While MetaHuman results exhibit bizarre shapes, our model achieves 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 the 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 the 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 (mm) on selected FaceScape dataset. (↓Lower is better.)

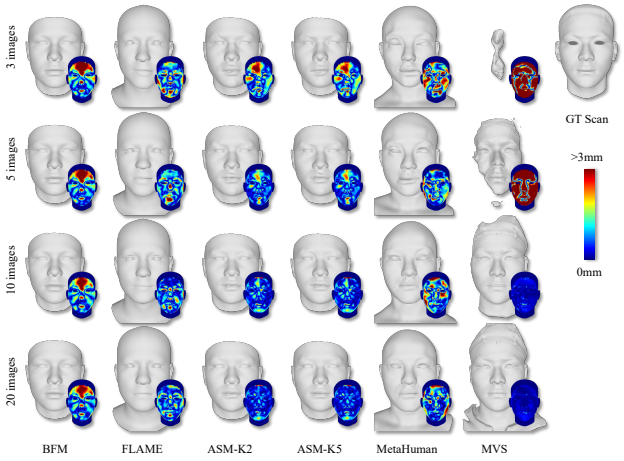

Figure 6. Multi-view reconstruction result on FaceScape.

To more accurately represent our middle-end environment, we 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. We also executed MVS with the photogrammetry software. Our model outperformed FLAME in producing more identifiable results, as depicted in Fig. 7, while MVS failed. These findings demonstrate that our model is

the suitable parametric face model for middle-end applications.

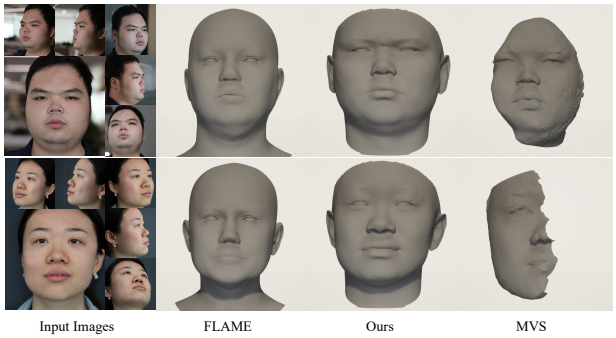

Figure 7. Multi-view reconstruction result on in-house data.

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 [25, 26]. 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 has the same setting as above, except the model is based on the topology from the game, as previously illustrated in Fig. 4. As shown in Fig. 8, in-game avatar from reconstruction result is achieved, and post-editing is allowed, due to the advantage of the skinning model 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 bones of the nose wing; (d) the original avatar.

4.3. Ablation study

An ablation study was conducted to investigate the key design components for fitting and reconstruction performance using the LYHM and MICC datasets, respectively. The study utilized the following methods: SSM, which is a static skinning model with fixed bone binding and skinning weights provided by Blender. DBB refers to dynamic bone binding and makes the bone position a tunable variable. GSW refers to GMM skinning weights, which makes skinning weights tunable parameters. RD refers to replacing the initial skinning weights provided by Blender with random ones. Tab. 7 shows that the default setting in previous evaluations is the setting for the last row. Results indicate that SSM has a higher representation capacity than most 3DMM models, with the exception of FLAME, leading to improved multi-view reconstruction performance. Converting bone location and skinning weights into tunable parameters further improves capacity. Careful consideration is required for the initialization of GMM skinning weight.

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

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

5. Discussion

This study demonstrates that statistical 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 achieves robust and state-of-the-art performance. However, for high-quality input captured within a camera rig, parametric face models are unnecessary, and MVS with raw vertices achieve high-quality facial scans, which are considered the ground truth. For intermediate-level applications using high-quality but uncalibrated images, skinning models with higher capacity, such as the proposed model, achieved state-of-the-art performance on MICC Coop benchmark, uncalibrated data from FaceScape, 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.

This study does not cover other aspects of multi-view reconstruction, such as constraint or optimization design. 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.

6. Conclusion

We propose ASM to address the gap of tailored face models for middle-end reconstruction with multi-view uncalibrated images. ASM offers stronger capacity than data-

dependent 3DMM with compact and fully tunable parameters. Our experiments demonstrate that ASM achieves SOTA performance for multi-view reconstruction on the MICC Coop benchmark, and its high capacity is crucial to exploit abundant information from multi-view input. The semantic parameters of ASM also make it suitable for real-world applications like in-game avatar creation. The study opens up new research directions for the parametric face model and facilitates future research on multi-view reconstruction.

One potential area for 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.

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, 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.