HiFace: High-Fidelity 3D Face Reconstruction by Learning Static and Dynamic Details

Zenghao Chai[1,2*] Tianke Zhang[2] Tianyu He[3] Xu Tan[3†] Tadas Baltrušaitis[4]
HsiangTao Wu[5] Runnan Li[5] Sheng Zhao[5] Chun Yuan[2] Jiang Bian[3]
[1]National University of Singapore [2]Tsinghua University [3]Microsoft Research Asia
[4]Microsoft Mixed Reality & AI Lab [5]Microsoft Cloud + AI
zenghaochai@gmail.com ztk21@mails.tsinghua.edu.cn yuanc@sz.tsinghua.edu.cn
{tianyuhe,xuta,tabaltru,musclewu,runnan.li,sheng.zhao,jiang.bian}@microsoft.com

Abstract

*3D Morphable Models (3DMMs) demonstrate great potential for reconstructing faithful and animatable 3D facial surfaces from a single image. The facial surface is influenced by the coarse shape, as well as the static detail (*e.g.*, person-specific appearance) and dynamic detail (*e.g.*, expression-driven wrinkles). Previous work struggles to decouple the static and dynamic details through image-level supervision, leading to reconstructions that are not realistic. In this paper, we aim at high-fidelity 3D face reconstruction and propose HiFace to explicitly model the static and dynamic details. Specifically, the static detail is modeled as the linear combination of a displacement basis, while the dynamic detail is modeled as the linear interpolation of two displacement maps with polarized expressions. We exploit several loss functions to jointly learn the coarse shape and fine details with both synthetic and real-world datasets, which enable HiFace to reconstruct high-fidelity 3D shapes with animatable details. Extensive quantitative and qualitative experiments demonstrate that HiFace presents state-of-the-art reconstruction quality and faithfully recovers both the static and dynamic details. Our project page: https://project-hiface.github.io.*

1. Introduction

The reconstruction of a 3D face from a single image has drawn much attention recently [67, 21, 41, 81]. It has tremendous potential applications like face recognition [11, 63, 4, 59], face animation [16, 75], virtual reality [7, 58, 31], *etc*. For example, the reconstructed 3D face representation can be driven by an audio [16], or a video from another person [38].

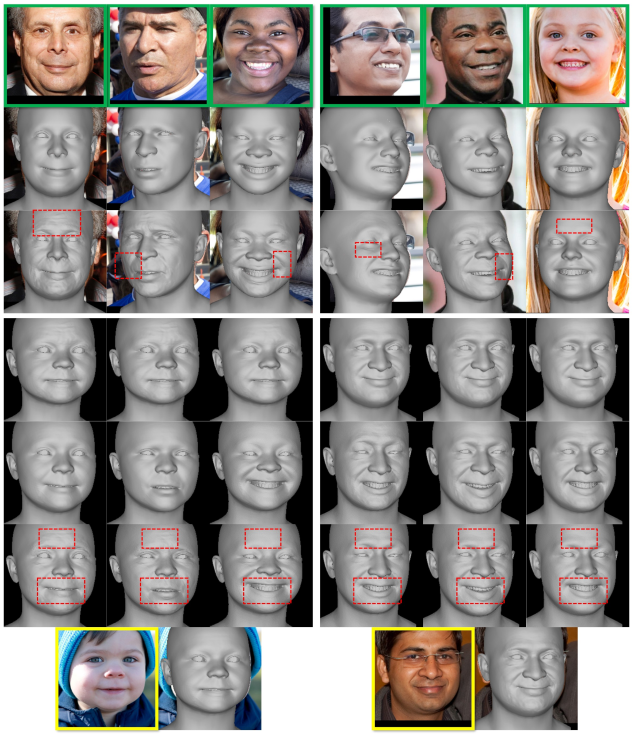

Figure 1. We propose HiFace to reconstruct high-fidelity 3D face with realistic and animatable details. **Reconstruction**: given a single image (1st-row), HiFace faithfully reconstructs a coarse shape (2nd-row) with vivid details (3rd-row). **Animation**: given a source face (yellow box), HiFace can animate the static (4th-row), dynamic (5th-row), or both (6th-row) details of the driving images (green box). Images are taken from FFHQ [37] and CelebA [40].

To build a flexible and animatable facial representation, a popular way is to leverage the success of 3D Morphable Models (3DMMs) [5, 6, 10, 45, 65], which decouple the influence of shape, expression, albedo, and others by modeling them in separate coefficients. Typically

[*]Work done when the author was an intern at MSRA.
[†]Corresponding author: Xu Tan (xuta@microsoft.com).

in literature, one can achieve coarse shape reconstruction in coefficients-fitting optimization [27, 73, 3, 77, 2], or an analysis-by-synthesis pipeline [67, 21, 81, 50]. As 3DMMs typically capture only the coarse facial geometry and are not capable of representing fine details (*e.g*., wrinkles), recent advances model such details with a displacement map [15, 77, 13, 9, 76]. However, previous work fails to model the distinction between static and dynamic factors of fine detail, leading to errors in reconstructions. For example, given that one may drive the expression of a young man from an old man, trivially transferring all wrinkles from the old man to the young man could make the young man look unnatural. In this sense, Feng *et al*. [24] implicitly leverages the person-specific identity and expression as conditions to generate the details. Although effective, they optimize the model in an analysis-by-synthesis pipeline with only the image-level supervision, leading to insufficient decoupling of static and dynamic details and inconsistent animation results (see Fig. 7).

Therefore, we propose HiFace to explicitly model the static and dynamic details for high-fidelity 3D face reconstruction, by designing SD-DeTail module to decouple the static and dynamic factors. More specifically, for person-specific static detail, instead of directly predicting the displacement map that may increase the difficulty of detail prediction [24, 19], we follow the spirit of 3DMMs to build a displacement basis from the captured facial scans with age diversity [57, 72]. In this way, the model is trained to predict the coefficients of the displacement basis, and make the detail prediction easier. For dynamic detail, since it is highly expression-dependent, directly modeling it with one displacement basis is quite difficult. Therefore, based on the fact that the expression can be interpolated by a compressed and a stretched expressions [57], we build two displacement bases for the compressed and stretched expressions from the captured scans respectively, and learn to regress the displacement coefficients with the ground-truth labels. Therefore, we can obtain the dynamic detail by linearly interpolating the compressed and stretched displacement maps, which are derived from the displacement bases and the predicted coefficients. Finally, the predicted static and dynamic details are merged with the coarse shape to formulate the final output.

Since we would like the final output to contain both the coarse shape and high-frequency detail, we propose several novel loss functions to learn coarse shape and details simultaneously from both the synthetic and real-world datasets. For details, we leverage the ground-truth static and dynamic displacement maps of the synthetic dataset [72, 57] as supervision. While for the coarse shape, we leverage the ground-truth vertex of the synthetic dataset as supervision. We also follow the previous methods [24, 21, 73] to leverage self-supervised losses for all training images.

Overall, with the above insights and techniques, HiFace enables the reconstruction of high-fidelity 3D faces from a single image, and decouples static and dynamic details that are naturally animatable (see Fig. 1). We demonstrate that the proposed HiFace reconstructs realistic and faithful 3D faces, reaching state-of-the-art performance both quantitatively and qualitatively. In addition, HiFace is compatible with optimization-based methods [73], and is flexible to transfer vivid expressions and details from one person to another. In summary, our contributions are:

- We propose HiFace to model the static and dynamic details explicitly, and demonstrate the benefits of synthetic data in decoupling the static and dynamic factors for detailed 3D face reconstruction.
- We propose novel loss functions in HiFace to learn 3D representations of coarse shape and fine details simultaneously from both the synthetic and real-world images.
- We achieve state-of-the-art reconstruction quality both quantitatively and qualitatively, with over 15% performance gains in the region-aware benchmark [12].
- We show that our SD-DeTail is easy to plug-and-play into optimization-based methods and can transfer expressions and details from one to another for face animation.

2. Related Work

3D face reconstruction from monocular images has received much attention in the past decades. Among them, 3D Morphable Models (3DMMs) are widely used to build 3D representations. Below we review the works that are related to them, and a full in-depth review can be found in recent surveys [82, 53, 23].

3D Morphable Model (3DMMs) [23] are statistical models widely used to constrain the distribution of 3D faces. The seminal work [5] presents 200 scans to generate shape and texture bases with Principal Component Analysis (PCA) [1], and formulate 3DMMs as linear models by the generated bases. After that, expression models [69, 45, 10] are proposed to support face manipulation. Recent advances [12, 55, 65, 18, 44] are proposed to expand the expressiveness of 3DMMs and play a crucial role in 3D face reconstruction. 3DMMs make it possible to simplify the 2D-to-3D problem into a regression task, which typically presents an analysis-by-synthesis fashion to estimate the coefficients of 3DMMs. In this paper, we follow the spirit of the 3DMMs family to present the decoupled static and dynamic details for 3D face reconstruction.

Coarse Shape Reconstruction. Traditional optimization-based methods [27, 73, 3, 77, 2] directly optimize the 3DMM coefficients of given 2D images. While such methods work well in controlled settings (*e.g*., frontal view, no occlusion), they heavily rely on high-quality annotations. Learning-based methods leverage the advances of

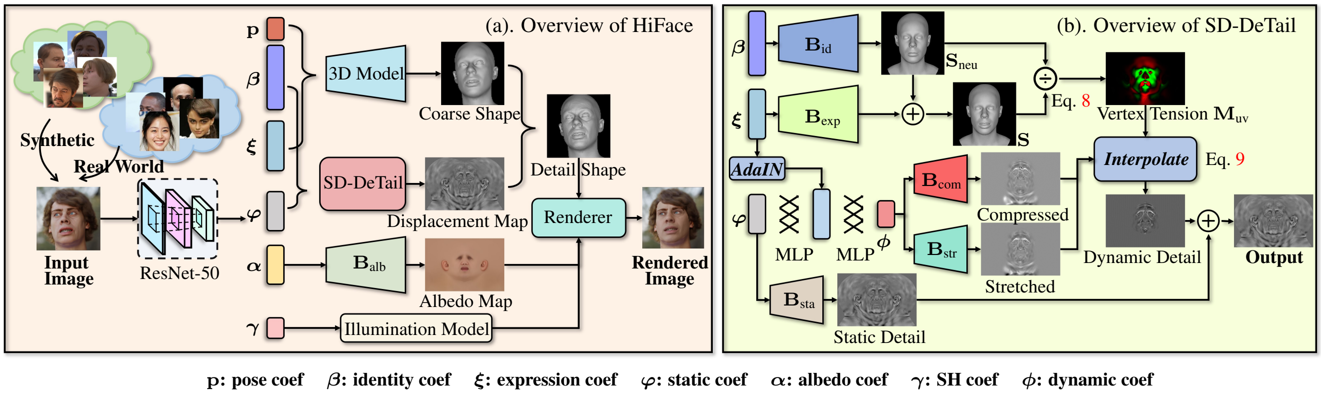

Figure 2. **Illustration of HiFace**. (a). Learning framework of HiFace. Given a monocular image, we regress its shape and detail coefficients to synthesize a realistic 3D face, and leverage a differentiable renderer [28] to train the whole model end-to-end from synthetic [72, 57] and real-world [40, 52] images. (b). The pipeline of **S**tatic and **D**ynamic **De**coupling for De**Tail** Reconstruction (SD-DeTail). We explicitly decouple the static and dynamic factors to synthesize realistic and animatable details. Given the shape and static coefficients, we regress the static and dynamic details through displacement bases and interpolate them into the final details through vertex tension [57].

CNNs [67, 21, 19, 62] and GCNs [47, 41, 25] to learn high-level representations from large-scale images in the wild. These methods show plausible generalization over diverse environments. To train the network end-to-end, recent methods leverage the differentiable renderers [28, 22, 80, 49], along with the photo loss, perceptual loss, and landmark loss [21, 19, 28, 67, 71] to optimize the network in a self-supervised manner. Different from these coarse shape reconstruction methods, we aim at high-fidelity 3D face reconstruction with both coarse shape and fine details.

Detail Reconstruction. While 3DMMs can reconstruct coarse 3D face shapes from 2D images, they struggle with reconstructing fine-level details, such as forehead wrinkles and crows-feet. To fill this gap, shape by shading (SfS) [34, 46, 26, 66] methods reconstruct the facial details from monocular images or videos. However, these methods are sensitive to occlusions and large poses. Recent advances [24, 48, 15, 13, 42] leverage displacement maps to present details. These methods explicitly re-topologize the coarse shape and present residual bias to generate geometric details. The main challenge of detail reconstruction is the difficulty in learning the nuances and disentangling the static and dynamic details from only self-supervised learning. Ground-truth labels of the details are helpful to guide the learning process. However, it is difficult to obtain such fine-grained labels on real data.

Synthetic Dataset. Several methods [81, 77, 33, 68, 51] utilize rendered faces or fitted coefficients to synthesize 3D-2D pairs. These ground-truth pairs lack diversity over background, illumination, and assets, making them hard to generalize well to real-world images. Recent advances in synthetic data generation [72, 57] demonstrate its ability to generalize to real-world settings, and diverse to compensate for the domain gap to real-world images. In this paper, we leverage high-quality data with ground-truth labels to explore the detailed 3D face reconstruction.

3. Methodology

3.1. Preliminary

We adopt a common practice [21, 24] to represent a textured coarse shape with a 3D face model, an illumination model, and a camera model.

3D Face Model. The 3D shape $\mathbf{S}$ and albedo $\mathbf{A}$ are represented by:

$$\begin{aligned}\mathbf{S} &= \bar{\mathbf{S}} + \boldsymbol{\beta}\mathbf{B}_{\text{id}} + \boldsymbol{\xi}\mathbf{B}_{\text{exp}} \\ \mathbf{A} &= \bar{\mathbf{A}} + \boldsymbol{\alpha}\mathbf{B}_{\text{alb}}\end{aligned}, \qquad (1)$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{A}}$ are the mean shape and albedo. $\mathbf{B}_{\text{id}}$, $\mathbf{B}_{\text{exp}}$, and $\mathbf{B}_{\text{alb}}$ are bases [73] of 256-dim identity, 233-dim expression, and 300-dim albedo, respectively. The coarse shape $\mathbf{S}$ in the bind pose is deformed from a neutral shape $\mathbf{S}_{\text{neu}} = \bar{\mathbf{S}} + \boldsymbol{\alpha}\mathbf{B}_{\text{id}}$ with expression component $\boldsymbol{\xi}\mathbf{B}_{\text{exp}}$. $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\alpha}$ are the corresponding identity, expression, and albedo coefficients for generating a coarse shape. In this paper, the coarse shape $\mathbf{S}$ contains $n_v = 7,667$ vertices and $n_f = 14,832$ triangles with $512 \times 512 \times 3$ albedo.

Pose & Camera Model. To estimate the face pose, we follow [72, 73] to predict skeletal pose $\mathbf{p} = [\boldsymbol{\theta}|\mathbf{t}]$, where $\boldsymbol{\theta} \in \mathbb{R}^{3j}$ and $\mathbf{t} \in \mathbb{R}^3$ are the local joint rotations and root joint translation, respectively. $j = 4$ indicates 4 skeletal joints w.r.t. the head, neck, and two eyes. We perform a standard linear blend skinning (LBS) function [43] (with per-vertex weights $\mathbf{W} \in \mathbb{R}^{j \times n_v}$) to rotate $\mathbf{S}$ about joint locations $\mathbf{J} \in \mathbb{R}^{3j}$ by $\mathbf{p}$ to obtain $\mathbf{S}_{\mathbf{p}}$:

$$\mathbf{S}_{\mathbf{p}} = \text{LBS}(\mathbf{S}, \mathbf{p}, \mathbf{J}; \mathbf{W}), \qquad (2)$$

where $\mathbf{J}$ is the joint locations in the bind pose determined by $\mathbf{J} = \mathcal{J}(\boldsymbol{\beta}) : \mathbb{R}^{|\boldsymbol{\beta}|} \to \mathbb{R}^{3j}$. Then we use an orthographic camera model to project 3D vertices in $\mathbf{S_p}$ to the 2D plane.

Illumination Model. We follow previous work [21] to use Spherical Harmonics (SH) [56] to estimate the illumination of a given image. The shaded texture $\mathbf{T}$ is computed as:

$$\mathbf{T} = \mathbf{A} \odot \sum\nolimits_{k=1}^{9} \gamma_k \boldsymbol{\Psi}_k(\mathbf{N}), \tag{3}$$

where $\odot$ denotes the Hadamard product, $\mathbf{N}$ is the surface normal of $\mathbf{S}$ in UV coordinates, $\boldsymbol{\Psi} : \mathbb{R}^3 \to \mathbb{R}$ are SH basis function and $\boldsymbol{\gamma} \in \mathbb{R}^9$ is the corresponding SH coefficient.

3.2. Overview of HiFace

Key Idea. The key idea of HiFace is to explicitly model the static (*e.g.*, person-specific properties) and dynamic (*e.g.*, expression-driven wrinkles) details, allowing the model to reconstruct a high-fidelity 3D face from a single image with realistic and animatable details.

Overview. The goal of HiFace is to reconstruct 3D shapes with realistic details from a single image. The overview of HiFace is illustrated in Fig. 2(a). We leverage a feature extractor (*i.e.*, ResNet-50 [30]) to regress corresponding coefficients from an input image. Our model jointly predicts both the coarse-level shapes and the fine-level details. For coarse-level shapes, we regress shape parameters (*i.e.*, identity, expression, albedo, illumination, and pose) of a parametric face model. For the fine-level details, we propose a novel way to model it through the separation of static and dynamic factors and formulate the generation of details into the problems of 3DMM coefficients regression and displacement maps interpolation.

Note that the facial details are based on the coarse shape, we thereby exploit novel loss functions to learn 3D representations of coarse shape and details simultaneously from the synthetic dataset with ground-truth labels. To generalize our model to real-world images, we also present several self-supervised losses to train the model with both synthetic data and real-world images coherently. As a result, HiFace can faithfully reconstruct the facial details of a given image, or animate a face by combining the decoupled static and dynamic coefficients that come from different individuals.

3.3. Decoupling Static and Dynamic Details

We propose **S**tatic and **D**ynamic **De**coupling for De**Tail** Reconstruction (SD-DeTail). The facial details are basically composed of a static factor and a dynamic factor:

$$\mathbf{D} = \mathbf{D}_{\text{sta}} + \mathbf{D}_{\text{dyn}}, \tag{4}$$

where $\mathbf{D}_{\text{sta}}$ and $\mathbf{D}_{\text{dyn}}$ indicate details from static and dynamic factors, respectively.

Concretely, the static factor is the inherent property of the identity (*i.e.*, the given 2D face), and originates from the appearance and age attributes. As for the dynamic factor, it is typically driven by the expression and influenced by person-specific properties.

Static Detail Generation. To simplify the problem, we are inspired by 3DMMs, which parameterize the statistical models to simplify the 2D-to-3D problem. We build a 300-dim displacement basis $\mathbf{B}_{\text{sta}}$ from the captured 332 scans [57] by PCA [1]. The scans contain diverse age groups in a neutral expression. Then we regress the coefficient $\boldsymbol{\varphi}$ to synthesize the static detail $\mathbf{D}_{\text{sta}}$ from the image:

$$\mathbf{D}_{\text{sta}} = \bar{\mathbf{D}}_{\text{sta}} + \boldsymbol{\varphi}\mathbf{B}_{\text{sta}}, \tag{5}$$

where $\bar{\mathbf{D}}_{\text{sta}}$ and $\mathbf{B}_{\text{sta}}$ are the mean displacement map and displacement basis for static details, respectively.

Dynamic Detail Generation. Due to the high diversity and complexity of expression representation, directly generating dynamic details from expression is quite difficult. Therefore we simplify the expression representation by using an interpolation between two displacement maps: compressed and stretched [57]. For example, the compressed expression may indicate a state of frowning to the extreme, while the stretched expression may indicate a state of complete relaxation between the eyebrows. Other states of this area can be interpolated by these two polarized states.

Consequently, we generate the dynamic details through compressed and stretched displacement maps. Again, we build 26-dim compressed $\mathbf{B}_{\text{com}}$ and stretched $\mathbf{B}_{\text{str}}$ displacement bases by PCA [1] to simplify the generation of displacement maps. To generate the dynamic coefficients $\boldsymbol{\phi} = \{\boldsymbol{\phi}_{\text{com}}, \boldsymbol{\phi}_{\text{str}}\}$, we apply the expression coefficient $\boldsymbol{\xi}$ into the static coefficient $\boldsymbol{\varphi}$ through AdaIN [32], followed by the MLP transformation $\boldsymbol{\Phi}$ to obtain $\boldsymbol{\phi}$:

$$\boldsymbol{\phi} = \boldsymbol{\Phi}\left(\sigma(\tilde{\boldsymbol{\xi}})\left(\frac{\boldsymbol{\varphi} - \mu(\boldsymbol{\varphi})}{\sigma(\boldsymbol{\varphi})} + \mu(\tilde{\boldsymbol{\xi}})\right)\right), \tag{6}$$

where $\tilde{\boldsymbol{\xi}}$ is the affined vector from $\boldsymbol{\xi}$ via MLP transformation. μ and σ indicate the mean and standard deviation. $\boldsymbol{\phi}_{\text{com}}$ and $\boldsymbol{\phi}_{\text{str}}$ are coefficients for compressed and stretched displacement maps respectively.

Similar to Eq. 5, the compressed and stretched displacement maps are formulated as:

$$\begin{aligned} \mathbf{D}_{\text{com}} &= \bar{\mathbf{D}}_{\text{com}} + \boldsymbol{\phi}_{\text{com}}\mathbf{B}_{\text{com}} \\ \mathbf{D}_{\text{str}} &= \bar{\mathbf{D}}_{\text{str}} + \boldsymbol{\phi}_{\text{str}}\mathbf{B}_{\text{str}} \end{aligned}, \tag{7}$$

where $\bar{\mathbf{D}}_{\text{com}}$ and $\mathbf{B}_{\text{com}}$ are the mean displacement map and 26-dim displacement basis for compressed detail, and $\bar{\mathbf{D}}_{\text{str}}$ and $\mathbf{B}_{\text{str}}$ are the mean displacement map and 26-dim displacement basis for stretched detail, respectively.

Considering the coarse shape $\mathbf{S}$ can be obtained by deforming the neutral shape $\mathbf{S}_{\text{neu}}$ with the expression component $\boldsymbol{\xi}\mathbf{B}_{\text{exp}}$, such expression-driven deformation over face

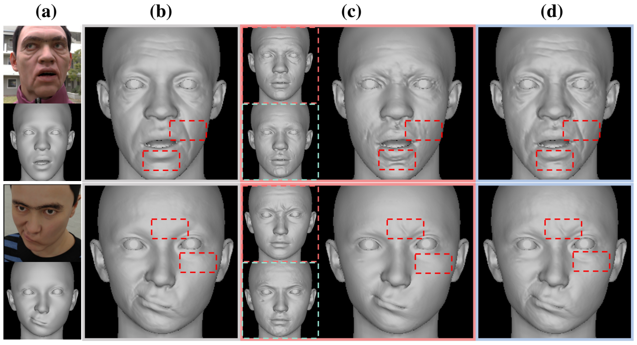

Figure 3. **Illustration of displacement map composition in SD-DeTail.** Given (a). an image (top) to reconstruct its coarse shape (bottom), we formulate the detail as (b). a static factor and (c). a dynamic factor interpolated by polarized states w.r.t. compressed (top) and stretched (bottom). (d). the output displacement map is linearly combined by (b) and (c) to present vivid details.

shape yields the "tension" over each vertex [57], which influences facial details from expression. Since $\mathbf{S}_{\text{neu}}$ and $\mathbf{S}$ posses the same topology, for each vertex $\mathbf{v}_i \in \mathbf{S}$ with K-edges $E_i = \{e_1, \cdots, e_K\}$ connected with $\mathbf{v}_i$, $E'_i = \{e'_1, \cdots, e'_K\}$ are the corresponding edges in $\mathbf{S}_{\text{neu}}$ that are connected to $\mathbf{v}'_i$. Then the tension at $\mathbf{v}_i$ is:

$$t_{\mathbf{v}_i} = 1 - \frac{1}{K}\sum\nolimits_{k=1}^{K}\frac{\|e_k\|}{\|e'_k\|}, \tag{8}$$

where $\|\cdot\|$ represents the edge length. Positive values of $t_{\mathbf{v}_i}$ indicate compression, negative values indicate stretch, and 0-value indicates no change, respectively.

The vertex tension $t_{\mathbf{v}_i}$ in $\mathbf{S}$ composes the tension map $\mathbf{M}_{\text{uv}}$ in UV coordinates. Then, the displacement map of the dynamic detail is the linear interpolation of $\mathbf{D}_{\text{com}}$ and $\mathbf{D}_{\text{str}}$:

$$\mathbf{D}_{\text{dyn}} = \mathbf{M}^{+}_{\text{uv}} \odot \mathbf{D}_{\text{com}} + \mathbf{M}^{-}_{\text{uv}} \odot \mathbf{D}_{\text{str}}, \tag{9}$$

where $\mathbf{M}^{+}_{\text{uv}}$ and $\mathbf{M}^{-}_{\text{uv}}$ indicate the positive and negative value of $\mathbf{M}_{\text{uv}}$, respectively. Fig. 3 shows the effectiveness of SD-DeTail. The dynamic factor interpolated by two polarized states introduces expression-related details and further decorates the static detail, yielding the final vivid output.

3.4. Overall Loss Functions

We propose several loss functions to train HiFace end-to-end. Specifically, we use static and dynamic detail losses to supervise the synthesized displacement maps from $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$. In addition, we leverage the coarse shape loss to supervise the reconstructed shape from $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. Finally, we follow previous methods [21, 24, 73] to leverage the differentiable renderer [28] to map the generated 3D shape into 2D images, by combining $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, $\mathbf{p}$, $\boldsymbol{\varphi}$, $\boldsymbol{\phi}$. Then, we perform self-supervised losses to train in both synthetic and real-world images. See more details in the supplementary.

Static and Dynamic Detail Losses. To explicitly train the details of each component, we leverage the ground-truth annotations from the synthetic dataset [57, 72] as supervision to assist the training process of our model. Specifically, we calculate the detail losses by estimating the l_2 distance between the reconstructed displacement maps and ground-truth w.r.t. static, compressed, and stretched components, and summarize them as $\mathcal{L}_{\text{detail}}$:

$$\begin{aligned}
\mathcal{L}_{\text{sta}} &= \left\|\mathbf{M}_{\text{detail}} \odot (\mathbf{D}_{\text{sta}} - \hat{\mathbf{D}}_{\text{sta}})\right\|_2 \\
\mathcal{L}_{\text{com}} &= \left\|\mathbf{M}_{\text{detail}} \odot (\mathbf{D}_{\text{com}} - \hat{\mathbf{D}}_{\text{com}})\right\|_2 \\
\mathcal{L}_{\text{str}} &= \left\|\mathbf{M}_{\text{detail}} \odot (\mathbf{D}_{\text{str}} - \hat{\mathbf{D}}_{\text{str}})\right\|_2 \\
\mathcal{L}_{\text{detail}} &= \mathcal{L}_{\text{sta}} + \mathcal{L}_{\text{com}} + \mathcal{L}_{\text{str}}
\end{aligned}, \tag{10}$$

where $\mathbf{M}_{\text{detail}}$ is the facial mask in the UV coordinates, and $\hat{\mathbf{D}}_{\text{sta}}/\hat{\mathbf{D}}_{\text{com}}/\hat{\mathbf{D}}_{\text{str}}$ and $\mathbf{D}_{\text{sta}}/\mathbf{D}_{\text{com}}/\mathbf{D}_{\text{str}}$ are the reconstructed and ground-truth displacement maps, respectively.

Coarse Shape Losses. Since the details should be based on realistic coarse shapes, we train the coarse shape to help the learning of details by leveraging the ground-truth vertex as supervision:

$$\mathcal{L}_{\text{ver}} = \left\|\mathbf{M}_{\text{ver}} \odot (\mathbf{S} - \hat{\mathbf{S}})\right\|_2, \tag{11}$$

where $\mathbf{M}_{\text{ver}}$ is frontal face area of the coarse shape. $\hat{\mathbf{S}}$ and $\mathbf{S}$ are the reconstructed and ground-truth face by Eq. 1.

In addition, we make constraints on shape coefficients to prevent overfitting. We enforce the predicted coefficients have a similar distribution to the ground-truth coefficients:

$$\mathcal{L}_{\text{kl}} = \rho(\boldsymbol{\beta})\big(\log\rho(\boldsymbol{\beta}) - \log\rho(\hat{\boldsymbol{\beta}})\big), \tag{12}$$

where ρ denotes *softmax* function to map the predicted coefficients $\hat{\boldsymbol{\beta}}$ and ground-truth $\boldsymbol{\beta}$ into probability distribution.

Finally, the shape loss is $\mathcal{L}_{\text{shp}} = \mathcal{L}_{\text{ver}} + \mathcal{L}_{\text{kl}}$.

Self-supervised Losses. To encourage the generalization of our models in real-world images [40, 52], we follow previous methods [24, 21, 73] to leverage self-supervised loss $\mathcal{L}_{\text{self}}$ for all training images, including photo loss $\mathcal{L}_{\text{pho}}$, perceptual loss $\mathcal{L}_{\text{id}}$, and dense landmark loss $\mathcal{L}_{\text{lmk}}$:

$$\mathcal{L}_{\text{self}} = \mathcal{L}_{\text{pho}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{lmk}}\mathcal{L}_{\text{lmk}}, \tag{13}$$

where λ_{id} and λ_{lmk} are weights to balance the self-supervised losses term.

In addition, considering the static detail heavily correlates to person-specific age attribute, inspired by [19], we leverage the pre-trained age prediction network [36] to learn high-level representations of static details through knowledge distillation, such that the learned coefficients exhibit expressive results. To achieve this, we use several MLP layers on the static coefficient $\boldsymbol{\varphi}$, and map it into age classification probabilities $\hat{\mathbf{p}}_{\text{age}}$. Then we use the pre-trained

Table 1. **Quantitative comparison of 3D face reconstruction methods on REALY benchmark**. "-c" and "-d" indicate coarse and detail shape, respectively. $@\mathcal{R}_N/@\mathcal{R}_M/@\mathcal{R}_F/@\mathcal{R}_C$/all indicate errors in nose/mouth/forehead/cheek/all regions. We highlight the best method for the two groups respectively. HiFace achieves the best reconstruction performance in the overall error by a large margin. Each component in HiFace contributes to a better reconstruction quality. The reconstructed details of HiFace further boost the quality while previous methods [24, 19] modeling details with only image-level supervision even deteriorate the reconstruction accuracy.

Group	Methods /	***frontal-view***					***side-view***				
	e (mm)	$@\mathcal{R}_N$	$@\mathcal{R}_M$	$@\mathcal{R}_F$	$@\mathcal{R}_C$	all	$@\mathcal{R}_N$	$@\mathcal{R}_M$	$@\mathcal{R}_F$	$@\mathcal{R}_C$	all
Coarse	Deep3D [21]	1.719±0.354	**1.368±0.439**	2.015±0.449	1.528±0.501	1.657	1.749±0.343	1.411±0.395	2.074±0.486	1.528±0.517	1.691
	MGCNet [64]	1.771±0.380	1.417±0.409	2.268±0.503	1.639±0.650	1.774	1.827±0.383	**1.409±0.418**	2.248±0.508	1.665±0.644	1.787
	3DDFA-v2 [29]	1.903±0.517	1.597±0.478	2.447±0.647	1.757±0.642	1.926	1.883±0.499	1.642±0.501	2.465±0.622	1.781±0.636	1.943
	DECA-c [24]	1.694±0.355	2.516±0.839	2.394±0.576	1.479±0.535	2.010	1.903±1.050	2.472±1.079	2.423±0.720	1.630±1.135	2.107
	SADRNet [61]	1.791±0.542	1.591±0.488	2.413±0.537	1.856±0.701	1.913	1.771±0.521	1.560±0.462	2.490±0.566	2.010±0.715	1.958
	EMOCA-c [19]	1.868±0.387	2.679±1.112	2.426±0.641	1.438±0.501	2.103	1.867±0.554	2.636±1.284	2.448±0.708	1.548±0.590	2.125
	MICA [81]	1.585±0.325	3.478±1.204	2.374±0.683	**1.099±0.324**	2.134	1.525±0.322	3.567±1.212	2.379±0.675	**1.109±0.325**	2.145
	Ours-c (w/o Syn. Data)[†]	1.227±0.407	1.787±0.439	1.454±0.382	1.762±0.436	1.558	1.187±0.379	1.826±0.490	1.470±0.426	1.653±0.450	1.534
	Ours-c	**1.054±0.317**	1.461±0.430	**1.331±0.347**	1.342±0.384	**1.297**	**0.992±0.246**	1.505±0.454	**1.427±0.400**	1.439±0.429	**1.341**
Detail	DECA-d [24]	2.138±0.461	2.802±0.868	2.457±0.559	1.443±0.498	2.210	2.286±1.103	2.684±1.041	2.519±0.718	1.555±0.822	2.261
	EMOCA-d [19]	2.532±0.539	2.929±1.106	2.595±0.631	1.495±0.469	2.388	2.455±0.636	2.948±1.292	2.606±0.686	1.599±0.563	2.402
	HRN [42]	1.722±0.330	**1.357±0.523**	1.995±0.476	**1.072±0.333**	1.537	1.642±0.310	**1.285±0.528**	1.906±0.479	**1.038±0.322**	1.468
	Ours-d (w/o Syn. Data)[†]	1.465±0.557	1.790±0.425	1.528±0.373	1.618±0.362	1.600	1.422±0.537	1.849±0.473	1.530±0.414	1.572±0.399	1.594
	Ours-d (w/o static)*	1.055±0.290	1.469±0.415	1.336±0.337	1.319±0.374	1.295	1.004±0.233	1.491±0.437	1.418±0.392	1.418±0.415	1.332
	Ours-d (w/o dynamic)*	1.069±0.318	1.469±0.414	1.358±0.336	1.270±0.344	1.292	0.991±0.239	1.496±0.437	1.411±0.393	1.375±0.402	1.318
	Ours-d	**1.036±0.280**	1.450±0.413	**1.324±0.334**	1.291±0.362	**1.275**	**0.985±0.237**	1.489±0.436	**1.399±0.388**	1.360±0.395	**1.308**

[†] To align the dataset scale, w/o Syn. Data indicates we train the model without using the ground-truth labels from the synthetic dataset.

* To eliminate the bias of coarse shape in estimating the reconstruction error, we fix the coarse shape and train the details with/without static and dynamic factors for comparisons.

age recognition model $\mathbf{\Gamma}_{\text{age}}$ to obtain the probabilities of the given input image $\mathbf{I}$. The distillation loss $\mathcal{L}_{\text{kd}}$ enforces the probabilities between $\hat{\mathbf{p}}_{\text{age}}$ and $\mathbf{\Gamma}_{\text{age}}(\mathbf{I})$ be similar:

$$\mathcal{L}_{\text{kd}} = \mathbf{\Gamma}_{\text{age}}(\mathbf{I})\big(\log \mathbf{\Gamma}_{\text{age}}(\mathbf{I}) - \log \hat{\mathbf{p}}_{\text{age}}\big). \tag{14}$$

Regularization. $\mathcal{L}_{\text{reg}}$ regularizes coefficients of each submodule, by minimizing the l_2 loss of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, $\boldsymbol{\varphi}$, $\boldsymbol{\phi}$.

Overall Loss Function. We train the coarse shape and fine details simultaneously, such that each component can collaborate to reconstruct high-fidelity 3D faces with realistic details. Formally, we minimize the total loss function:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{detail}}\mathcal{L}_{\text{detail}} + \lambda_{\text{shp}}\mathcal{L}_{\text{shp}} \\ & + \lambda_{\text{self}}\mathcal{L}_{\text{self}} + \lambda_{\text{kd}}\mathcal{L}_{\text{kd}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}},\end{aligned} \tag{15}$$

where λ is the weight for each component.

4. Experiments

4.1. Implementation Details

Dataset. We use a hybrid dataset made up from both synthetic [72, 57] and real data [40, 52]. We use the synthetic data pipeline [72, 57] to generate a diverse dataset of $200k$ faces with ground-truth vertex, landmark, albedo, and displacement map annotations. The real-world datasets contain $400k$ images in total from diverse age, gender, and ethnicity groups. For the real-world dataset, we use the pre-trained dense landmark detector [72] to detect 669 landmarks for training. We use face parsing [79] to generate and select region-of-interest as facial masks, providing robustness to common occlusions by hair or other accessories. We follow [19, 73, 21] to split the dataset into training and validation sets. The test images are from CelebA [40], FFHQ [37], LS3D-W [8], and AFLW2000 [78].

Implementation Details. We implement HiFace in PyTorch [54] and leverage the PyTorch3D differentiable rasterizer [35] for rendering. We train our model for 35 epochs on 8× NVIDIA Tesla V100 GPUs with a mini-batch of 320. We use the pre-trained ResNet-50 on ImageNet [20] as initialization, and use Adam [39] as optimizer with an initial learning rate of $1e-4$. The input image is cropped and aligned by [14], and resized into 224×224. We empirically set $\lambda_{\text{detail}} = 10$, $\lambda_{\text{shp}} = 1$, $\lambda_{\text{self}} = 1$, $\lambda_{\text{id}} = 0.1$, $\lambda_{\text{lmk}} = 0.5$, $\lambda_{\text{kd}} = 1$, $\lambda_{\text{reg}} = 1e-3$ throughout the experiments.

4.2. Comparisons to State-of-the-art

Quantitative Comparison. We perform the quantitative evaluation on the REALY benchmark [12], which contains 100 frontal-view and 400 side-view images from 100 textured-scans [17]. The REALY benchmark presents a region-aware evaluation pipeline to separately evaluate the metric error (in mm) of the nose, mouth, forehead, and cheek regions. Such an evaluation pipeline is demonstrated to better estimate the actual similarity of the 3D faces and align with human perception. We compare HiFace to previous state-of-the-art methods and report the region-wise and average normalized mean square error (NMSE) in Tab. 1.

As Tab. 1 illustrates, HiFace outperforms prior arts in the overall error by a large margin. HiFace balances the reconstruction quality on each region, compared to those optimum region methods that may fail in specific regions (*e.g.*, MICA [81] fails in mouth region while HRN [42] fails in forehead region). Note that HiFace faithfully recovers the facial details, thus making the reconstruction error smaller

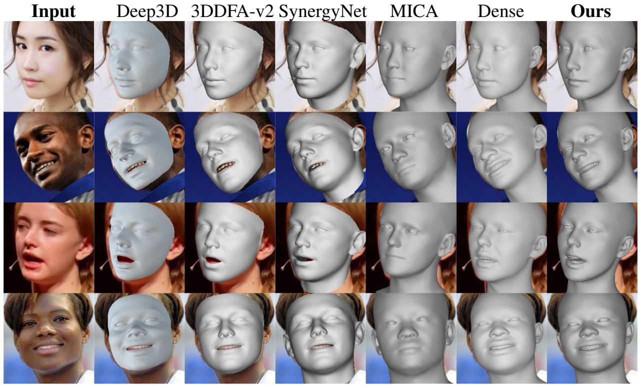

Figure 4. **Comparison on coarse shape reconstruction.** From left to right: Input image, Deep3D [21], 3DDFA-v2 [29], SynergyNet [74], MICA [81], Dense [73], and HiFace (Ours).

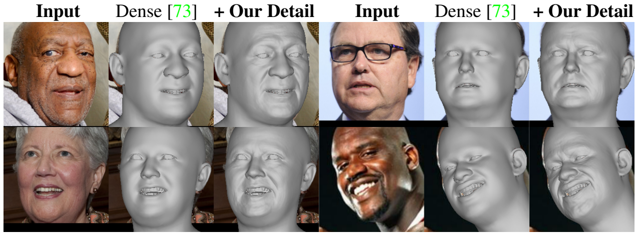

Figure 6. **Illustration on the flexibility of SD-DeTail.** Given the identity and expression coefficients ($\boldsymbol{\beta}$, $\boldsymbol{\xi}$) from the optimization-based method [73], SD-DeTail can generate realistic details based on the coarse shape and further improve the visual quality.

than using the coarse shape alone. As a comparison, although DECA [24] and EMOCA [19] can reconstruct details of given images, they turn out to be noisy, leading to the deterioration of reconstruction quality.

In addition, Tab. 1 also demonstrates the necessity of each component in contributing to a better quality. It can be observed that the synthetic data with ground-truth labels not only improve the coarse shape reconstruction quality but is also crucial for detailed reconstruction. With the synthetic data, the proposed SD-DeTail further boosts the overall reconstruction quality. Both the static and dynamic factors are essential to capture fine-grained details, and the final SD-DeTail achieves the most accurate details in expression-related regions such as the mouth and forehead, which contributes to the overall gains.

Qualitative Comparison. Given a single face image, HiFace reconstructs a high-fidelity 3D shape with details. We present comparisons with previous methods on 1). coarse shape reconstruction [21, 29, 74, 81, 73] in Fig. 4 and 2). detail reconstruction [77, 15, 24, 19, 70] in Fig. 5. See more examples and comparisons in the supplementary.

For coarse shape in Fig. 4, our HiFace faithfully recovers the coarse shape of the given identity and outperforms the previous learning-based methods, and is on par

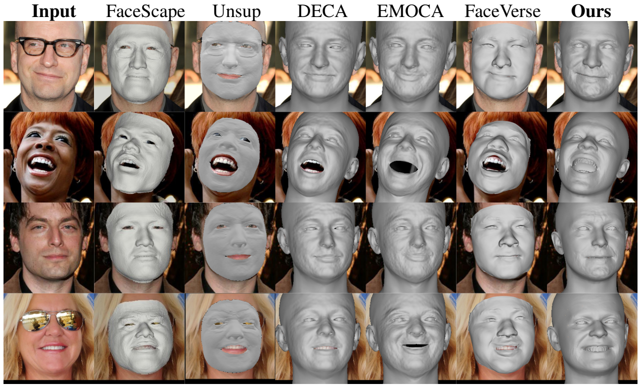

Figure 5. **Comparison on detail shape reconstruction.** From left to right: Input image, FaceScape [77], Unsup [15], DECA [24], EMOCA [19], FaceVerse [70], and HiFace (Ours).

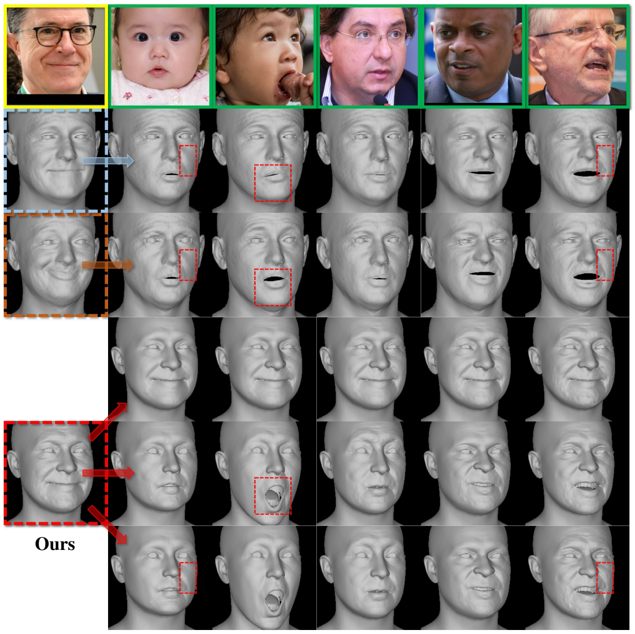

Figure 7. **Comparison on face animation.** Given a source image (yellow box), we use the driving images (green box) to drive its expressions. DECA [24] (2nd-row) and EMOCA [19] (3rd-row) can animate the expression-driven details but lack realistic. As a comparison, HiFace is flexible to animate details from static (4th-row), dynamic (5th-row), or both (6th-row) factors, and presents vivid animation quality with realistic shapes.

with Dense [73], which is the state-of-the-art optimization-based method. For detailed reconstruction in Fig. 5, previous methods [24, 19] fail to reconstruct satisfactory details. Several methods [15, 77, 70] are sensitive to occlusions and large poses. As a comparison, HiFace achieves the most realistic reconstruction quality, and faithfully recovers facial details of a given image, which significantly outperforms previous methods by a large margin.

In addition, given an image and the fitted coefficients $\boldsymbol{\beta}$,

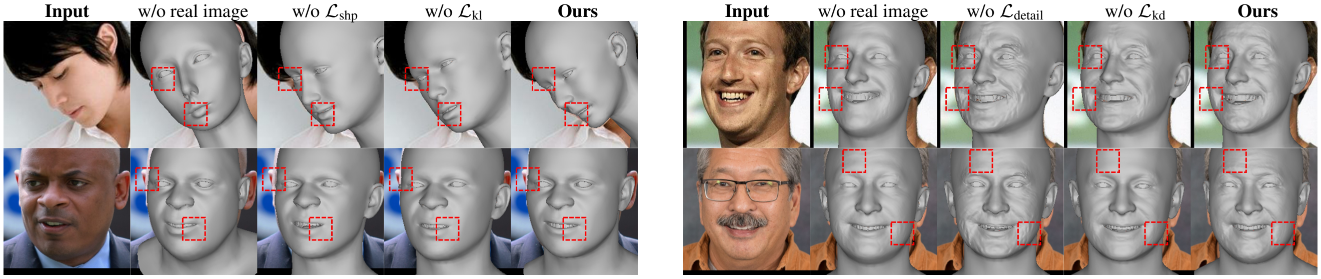
Input
w/o real image
w/o $\mathcal{L}_{\text{shp}}$
w/o $\mathcal{L}_{\text{kl}}$
Ours
Input
w/o real image
w/o $\mathcal{L}_{\text{detail}}$
w/o $\mathcal{L}_{\text{kd}}$
Ours

Figure 8. **Ablation studies on loss functions and training data.** The coarse shape losses $\mathcal{L}_{\text{shp}}/\mathcal{L}_{\text{kl}}$ (left), detail losses $\mathcal{L}_{\text{detail}}/\mathcal{L}_{\text{kd}}$ (right), and hybrid datasets coherently contribute to the reconstruction quality of coarse shapes and details.

ξ from the optimization-based methods such as Dense [73], SD-DeTail synthesizes the details and further strengthens the quality compared to the coarse shape (see Fig. 6). It shows SD-DeTail is flexible and can be easily plugged-and-play into other methods. See more in the supplementary.

Application of HiFace. The HiFace explicitly decouples the static and dynamic details through the proposed SD-DeTail. Therefore, we can animate the facial attributes by simply assigning the expression coefficient ξ and/or static coefficient φ of the driving images to the source images.

In Fig. 7, we demonstrate the animation quality of HiFace outperforms the previous state-of-the-art detail animation methods [24, 19]. It shows that while DECA [24] and EMOCA [19] can animate the expression-driven details but lack realistic, the proposed HiFace is flexible to manipulate the static, dynamic, or both details. When animating the static detail, the person-specific details can be well transferred into the source shape. When animating the dynamic detail, only expression-dependent details are presented. Finally, we can also animate the static and dynamic details simultaneously and achieve satisfactory results.

5. Ablation Studies

Ablation Studies on Loss Functions and Datasets. We present ablation studies on the proposed loss functions and training strategy with hybrid datasets. We train HiFace with synthetic data alone and compare it to the one trained with hybrid datasets. For coarse shape reconstruction, we investigate the contribution of $\mathcal{L}_{\text{shp}}$ and its sub-term $\mathcal{L}_{\text{kl}}$. For detail reconstruction, we compare HiFace without $\mathcal{L}_{\text{detail}}$ and $\mathcal{L}_{\text{kd}}$, respectively. The results are presented in Fig. 8.

Fig. 8 demonstrates that the proposed loss functions and training strategy from hybrid datasets contribute to satisfactory coarse shape and details. First, the model trained with synthetic data alone cannot generalize well to real-world images, which indicates the necessity to train with real-world data. Second, $\mathcal{L}_{\text{shp}}$ improves the coarse shape reconstruction quality. $\mathcal{L}_{\text{shp}}$ is effective in tackling challenging poses and improving alignment. $\mathcal{L}_{\text{kl}}$ can relieve the overfitting risk on the synthetic data and improve the generalization to real-world images. Third, without $\mathcal{L}_{\text{detail}}$ or $\mathcal{L}_{\text{kd}}$, the reconstructed details exhibit random noise and cannot faithfully reflect person-specific details. Such noise misses the correspondence to the person-specific identity.

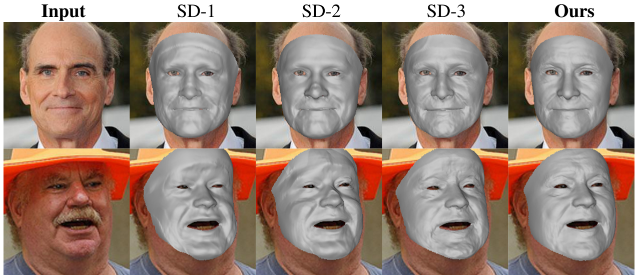

Figure 9. **Ablation studies on SD-DeTail.** Results show that directly synthesizing the static or dynamic details is rather challenging, leading to unreasonable coarse shapes and details (SD-1, SD-2, and SD-3). As a comparison, we leverage the statistical bases with SD-DeTail and regard the detail generation problem as a coefficients regression and interpolation problem, leading to more realistic details.

Ablation Studies on SD-DeTail. To verify the effectiveness of building bases for static and dynamic details, we present detailed ablation studies on SD-DeTail, by replacing the bases (*i.e.*, $\mathbf{B}_{\text{sta}}$ and $\mathbf{B}_{\text{com}}/\mathbf{B}_{\text{str}}$) reconstruction with a U-Net decoder [60] (same as DECA [24]). Therefore, the model learns to directly synthesize displacement maps instead of predicting corresponding coefficients like ours. In Fig. 9, we make comparisons on: 1). directly synthesizing $\mathbf{D}_{\text{dyn}}$ (SD-1), 2). directly synthesizing $\mathbf{D}_{\text{com}}/\mathbf{D}_{\text{str}}$ and interpolating via Eq. 9 (SD-2), 3). directly synthesizing $\mathbf{D}_{\text{sta}}$ (SD-3), 4). SD-DeTail (Ours). It can be seen that, due to the high diversity and complexity of expression representation, it is hard to directly learn realistic details even with ground-truth labels of $\mathbf{D}_{\text{dyn}}$ from synthetic data (see SD-1, SD-2 and SD-3 in Fig. 9). More specifically, for the static details, directly synthesizing displacement maps bring much noise (SD-3). For example, the hollow eyebrow is demonstrated in the second row. For the dynamic details, directly synthesizing displacement maps even leads to unnatural results (SD-1 and SD-2). For example, the reconstructed 3D faces are distorted especially in the second row. We also notice that, directly synthesizing $\mathbf{D}_{\text{dyn}}$ (SD-1) achieves inferior results than directly synthesizing $\mathbf{D}_{\text{com}}/\mathbf{D}_{\text{str}}$ and interpolating via Eq. 9 (SD-2). This demonstrates that it is beneficial to simplify the expression representation by using interpolation between two displacement maps (*i.e.*, compressed and stretched). In conclusion, these observations further verify our insight on relaxing the challenging detail generation problem into a feasible coefficients regression problem.

6. Conclusion

We propose HiFace to reconstruct high-fidelity 3D faces with realistic and animatable details from a single image. Our motivation and insights stand on the spirit of 3DMMs to simplify the challenging detail generation into more accessible regression and interpolation tasks. To achieve this, we elaborately design SD-DeTail to decouple the static and dynamic factors explicitly, and interpolate the dynamic details through vertex tension. We succeed in learning the coarse shape and details jointly by proposing several novel loss functions to train on synthetic and real-world data. Extensive experiments demonstrate that HiFace achieves state-of-the-art face reconstruction both quantitatively and qualitatively in the coarse shape and detail shape, and the details are well decoupled and naturally animatable.

Acknowledgement

This work was partially supported by the National Key R&D Program of China (2022YFB4701400/4701402), SZSTC Grant (JCYJ20190809172201639, WDZC20200820200655001), Shenzhen Key Laboratory (ZDSYS20210623092001004), and Beijing Key Lab of Networked Multimedia.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 2, 4

[2] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2023. 2

[3] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics (TOG)*, 41(1):1–21, 2021. 2

[4] Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *Proceedings of fifth IEEE international conference on automatic face gesture recognition*, pages 202–207. IEEE, 2002. 1

[5] 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. 1, 2

[6] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. 1

[7] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics (ToG)*, 32(4):1–10, 2013. 1

[8] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 6

[9] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Realtime high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):1–9, 2015. 2

[10] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 1, 2

[11] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2018. 1

[12] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 6

[13] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019. 2, 3

[14] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016. 6

[15] Yajing Chen, Fanzi Wu, Zeyu Wang, Yibing Song, Yonggen Ling, and Linchao Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020. 2, 3, 7

[16] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 1

[17] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 2019. 6

[18] Hang Dai, Nick Pears, William AP Smith, and Christian Duncan. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE international conference on computer vision*, pages 3085–3093, 2017. 2

[19] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2, 3, 5, 6, 7, 8

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[21] 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 *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 1, 2, 3, 4, 5, 6, 7

[22] Abdallah Dib, Cédric Thébault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12819–12829, 2021. 3

[23] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 2

[24] 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. 2, 3, 5, 6, 7, 8

[25] Zhongpai Gao, Juyong Zhang, Yudong Guo, Chao Ma, Guangtao Zhai, and Xiaokang Yang. Semi-supervised 3d face representation learning from unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–349, 2020. 3

[26] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):1–15, 2016. 3

[27] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1155–1164, 2019. 2

[28] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 3, 5

[29] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 6, 7

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[31] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017. 1

[32] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[33] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017. 3

[34] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27(10):4756–4770, 2018. 3

[35] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3d deep learning with pytorch3d. In *SIGGRAPH Asia 2020 Courses*, pages 1–1. ACM New York, NY, USA, 2020. 6

[36] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 5

[37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 6

[38] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 1

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 6

[40] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3, 5, 6

[41] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6100–6109, 2020. 1, 3

[42] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023. 3, 6

[43] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 3

[44] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3410–3419, 2020. 2

[45] 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. 1, 2

[46] Yue Li, Liqian Ma, Haoqiang Fan, and Kenny Mitchell. Feature-preserving detailed 3d face reconstruction from a single image. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–9, 2018. 3

[47] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5891–5900, 2020. 3

[48] Jingwang Ling, Zhibo Wang, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Structure-aware editable morphable model for 3d facial detail animation and manipulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[49] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 3

[50] Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11672, 2021. 2

[51] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiři Matas, and Viktoriia Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20942–20952, 2022. 3

[52] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 3, 5, 6

[53] Araceli Morales, Gemma Piella, and Federico M Sukno. Survey on 3d face reconstruction from uncalibrated images. *Computer Science Review*, 40:100400, 2021. 2

[54] 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. 6

[55] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10934–10943, 2019. 2

[56] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 4

[57] Chirag Raman, Charlie Hewitt, Erroll Wood, and Tadas Baltrusaitis. Mesh-tension driven expression-based wrinkles for synthetic faces. In *WACV*, 2023. 2, 3, 4, 5, 6

[58] Roger Blanco I Ribera, Eduard Zell, John P Lewis, Junyong Noh, and Mario Botsch. Facial retargeting with automatic range of motion alignment. *ACM Transactions on graphics (TOG)*, 36(4):1–12, 2017. 1

[59] Sami Romdhani, Volker Blanz, and Thomas Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *European Conference on Computer Vision*, pages 3–19. Springer, 2002. 1

[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 8

[61] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Transactions on Image Processing*, 30:5793–5806, 2021. 6

[62] 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. 3

[63] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[64] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020. 6

[65] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5011–5020, 2020. 1, 2

[66] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. Total moving face reconstruction. In *European conference on computer vision*, pages 796–812. Springer, 2014. 3

[67] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 1, 2, 3

[68] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 3

[69] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*, pages 24–es. ACM New York, NY, USA, 2006. 2

[70] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20333–20342, 2022. 7

[71] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3d face reconstruction via conditional estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13289–13298, 2021. 3

[72] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 2, 3, 5, 6

[73] 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 *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 2, 3, 5, 6, 7, 8

[74] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 International Conference on 3D Vision (3DV)*, 2021. 7

[75] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Chen Qifeng, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. In *Advances in Neural Information Processing Systems*, 2022. 1

[76] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

[77] 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. 2, 3, 7

[78] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3990–3999, 2017. 6

[79] Qi Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Computer Vision and Pattern Recognition*, 2022. 6

[80] Wenbin Zhu, HsiangTao Wu, Zeyu Chen, Noranart Vesdapunt, and Baoyuan Wang. Reda: reinforced differentiable attribute for 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4958–4967, 2020. 3

[81] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Oct. 2022. 1, 2, 3, 6, 7

[82] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer graphics forum*. Wiley Online Library, 2018. 2