REALY: Rethinking the Evaluation of 3D Face Reconstruction

Zenghao Chai[1*], Haoxian Zhang[2*], Jing Ren[2], Di Kang[2], Zhengzhuo Xu[1], Xuefei Zhe[2], Chun Yuan[1,3†], and Linchao Bao[2†]

[1] Shenzhen International Graduate School, Tsinghua University, China
[2] Tencent AI Lab, China, [3] Peng Cheng National Laboratory, China

Abstract. The evaluation of 3D face reconstruction results typically relies on a rigid shape alignment between the estimated 3D model and the ground-truth scan. We observe that aligning two shapes with different reference points can largely affect the evaluation results. This poses difficulties for precisely diagnosing and improving a 3D face reconstruction method. In this paper, we propose a novel evaluation approach with a new benchmark REALY, consists of 100 globally aligned face scans with accurate facial keypoints, high-quality region masks, and topology-consistent meshes. Our approach performs region-wise shape alignment and leads to more accurate, bidirectional correspondences during computing the shape errors. The fine-grained, region-wise evaluation results provide us detailed understandings about the performance of state-of-the-art 3D face reconstruction methods. For example, our experiments on single-image based reconstruction methods reveal that DECA performs the best on nose regions, while GANFit performs better on cheek regions. Besides, a new and high-quality 3DMM basis, HIFI3D^{++}, is further derived using the same procedure as we construct REALY to align and retopologize several 3D face datasets. We will release REALY, HIFI3D^{++}, and our new evaluation pipeline at https://realy3dface.com.

Keywords: 3D Face Reconstruction, Evaluation, Benchmark, 3DMM

1 Introduction

3D face reconstruction is a hotspot with broad applications in real world including face alignment [73,29], face recognition [8,13,64], and face animation [10,12] among many others. How to estimate high fidelity 3D facial mesh [57,19,34] from monocular RGB(-D) images or image collections is a challenging problem in the fields of computer vision, computer graphics and machine learning.

Various methods have been proposed to tackle this problem, among which DNNs, especially CNNs [59,51,57] and GCNs [25,36], have made great progress due to their great expressiveness. However, developing new reconstruction methods and evaluating different methods or 3DMM basis are severely constrained

* Equal Contributions.
† Corresponding authors: yuanc@sz.tsinghua.edu.cn; linchaobao@gmail.com.

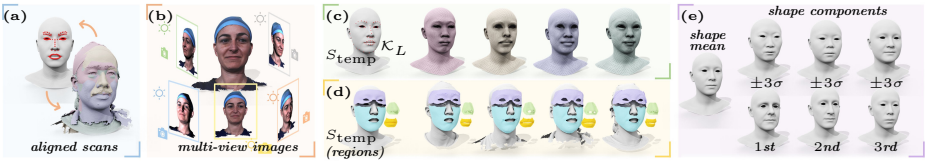


Fig. 1. REALY: a **Re**gion-**a**ware benchmark based on the **LY**HM [18] dataset. Our benchmark contains 100 high-quality face shapes and *each* individual has **(a)** a rescaled and globally aligned scan, **(b)** 5 synthesized *multi-view* images with various GT camera parameters and illuminations, **(c)** a retopologized full-head mesh in HIFI3D [4] topology with consistent and semantically meaningful 68 keypoints, **(d)** 4 consistent region masks defined on both the retopologized mesh and the original scan, and **(e)** HIFI3D^{++} 3DMM: the first three PCs with the mean shape show the ethnic diversity.

by available datasets. Existing open-source 3D face datasets [2,68,46,70] have some unneglectable flaws. For example, the face scans are in different scales and random poses, and the provided keypoints are not accurate or discriminative enough, which makes it extremely hard to align the input shapes to the predicted face for evaluation. Moreover, due to the lack of ground-truth annotations in the original face scans, standard evaluation pipeline relies on nearest-neighboring correspondences to measure the similarity between the scan and the estimated face shape, which completely ignores substantive characteristics and discards shape geometry of human faces.

To fill this gap, we propose a new benchmark named REALY for evaluating 3D face reconstruction methods. REALY contains 3D face scans of 100 individuals from the LYHM [18] dataset, where the face scans are consistently rescaled, globally aligned, and wrapped into topology-consistent meshes. More importantly, since we have predefined facial keypoints and masks of the retopologized mesh template, the keypoints and masks can be transferred to original face scans. In this case, we get the high-quality facial keypoints and masks of the original raw face scans, which enable us to perform more accurate alignments and fine-grained, region-wise evaluations for estimated 3D face shapes. See Fig. 1 for an illustration. Our benchmark contains individuals from different ethnic, age, and gender groups (see Fig. 2 for some examples). Utilizing the retopologizing procedure built for REALY, we further present a high-quality and powerful 3DMM basis named HIFI3D^{++} by aligning and retopologizing several 3D face datasets. We conduct extensive experiments to evaluate state-of-the-art 3D face reconstruction methods and 3DMMs, which reveal several interesting observations and potential future research directions.

Contributions. To summarize, our main contributions are:

- A new 3D face benchmark REALY that contains prealigned scans with accurate facial keypoints and region masks, retopologized meshes, and rendered high-fidelity multi-view images with camera parameters.
- A thorough investigation of the flaws in the standard evaluation pipeline for measuring face reconstruction quality.
- A novel, informative evaluation approach for 3D face reconstruction, with an elaborated region-wise, bidirectional alignment pipeline.

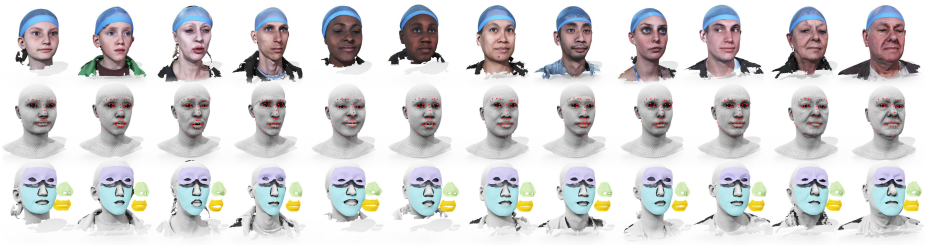


Fig. 2. Examples from REALY. *Top*: aligned high-resolution scans with textures. *Middle*: retopologized meshes in HIFI3D topology with semantically consistent keypoints (red points). *Bottom*: high quality face region masks of each scan.

- Extensive experiments for benchmarking state-of-the-art 3D face reconstruction methods and 3DMMs.
- A new full-head 3DMM basis HIFI3D^{++} built from several 3D face datasets with high-quality, consistent mesh topology.

2 Related work

Face reconstruction has drawn great attention in the past decades in both computer vision and computer graphics communities [73,63,54,37,72,33,65]. Below we review the topics that are most closely related to our work, and a full in-depth review can be found in [21,76,9].

3D Face Database. High quality 3D scan datasets greatly promote the development in the field of 3D face reconstruction. Massive face databases [73,40,32] have made it possible to train models for face reconstruction in a self-supervised manner. However, the 3D face scans in the existing datasets [46,2,18,70] are in different scales and random poses, and only a small set of inaccurate keypoints are provided for alignment. Another type of databases [68,11,70] contains retopologized meshes that are registered from high fidelity scans where all the meshes share the same topology. This type of databases is essential to construct 3D Morphable Models (3DMMs) [6,8,68,45], statistical models of facial shape and texture, which can be used for face regression and editing.

Single-View 3D Face Reconstruction. 3D face reconstruction from a single-view image has received glaring attention over the past decades, though estimating 3D information from a single 2D image is challenging and severely ill-posed. With the help of 3D morphable face models [6,35,21,11,48,55], the reconstruction problem is simplified into a tractable parametric regression. A straightforward solution to estimate 3DMM coefficients is based on analysis-by-synthesis [7,58,67,30], where the optimization objective usually consists of facial landmark alignment, photo consistency and statistical regularizers. These optimization-based approaches are computationally expensive and sensitive to initialization. Recently, many deep learning based models [59,57,28,51] are proposed to predict the 3DMM coefficients in a supervised or self-supervised way. This type of methods is robust for face reconstruction but has limited ex-

pressive power. To address this issue, GANFit [26,27] proposes to parameterize the texture maps using the latent code of a Texture GAN for face regression. Some nonlinear 3DMMs [61,62,60] are proposed for stronger expressiveness of face geometry. Some other work [22,17,75,63,31] utilize additional geometry and appearance representation (such as displacement and normal maps) to recover high-frequency details. Moreover, a recent surge of end-to-end approaches try to reconstruct 3D face shape directly from a depth map or UV position map [23,71,53,41,43,66]. However, these non-3DMM methods are prone to produce unrealistic and malformed faces compared to 3DMM-based methods.

Evaluation of Face Reconstruction. Existing evaluation protocols usually utilize the off-the-shelf datasets [2,11,46,70] to estimate the similarity between the reconstructed shape and the raw scan. Specifically, the reconstructed shape and the input scan are aligned using some predefined keypoints [38,17,24,75] or ICP [5,19,26,27,28,52]. Then Root Mean Square Error (RMSE) [59,19,42,16] or Normalized Mean Square Error (NMSE) [75,39,60,34] is calculated between the corresponding points on the input scan and the reconstructed shape. The correspondences are usually established in two ways, namely finding the nearest neighbor with smallest point-to-point distance [75,17] or point-to-plane distance [19,39,41]. Some benchmarks [24,52,42] propose to use predefined keypoints or disk-shaped region masks to measure shape similarity. Such measurement contains semantic prior but does not faithfully represent the overall face shape.

3 Background

3.1 Notation & Preliminaries

We use *triangle* mesh $S = \{ V, F \}$ to represent face models with vertex positions V and triangle face list F. A *region-of-interest* of the mesh S is denoted as $\mathcal{R}_S$, which can be represented as an indicator function or a list of face IDs. We denote the *keypoints* on the mesh S as $\mathcal{K}_S$, which is a list of manually selected or automatically detected vertices that are semantically meaningful on S.

For a specific face shape, we consider three associated meshes: (1) S_H: the *ground-truth* mesh with *high* resolution, which is constructed from multi-view images; (2) S_L: the *ground-truth* retopologized mesh with *low* resolution, which is obtained by wrapping the HIFI3D [4] mesh topology to the shape S_H; (3) S_P: the *predicted* face mesh constructed from existing techniques. Note that different reconstruction methods may choose different mesh topologies (i.e., different size of V_P and F_P). For simplicity of notation, we denote the *regions* and *keypoints* defined on shape $S_H/S_L/S_P$ as $\mathcal{R}_H/\mathcal{R}_L/\mathcal{R}_P$ and $\mathcal{K}_H/\mathcal{K}_L/\mathcal{K}_P$ respectively.

A *map* between the shape S_i and S_j is denoted as $T_{i\to j}$, where the subscript represents the map *direction*. For example, $T_{p\to h}$ represents a map from the predicted mesh S_P to the high-resolution GT mesh S_H. We consider two different types of map, the vertex-to-vertex map $T_{i\to j}^{\text{vtx}}$ and the vertex-to-point (also called vertex-to-plane) map $T_{i\to j}^{\text{pts}}$. Specifically, $T_{i\to j}^{\text{vtx}}$ maps each vertex in shape S_i to a

Table 1. Overview of 3DMMs.

	BFM [45]	FWH [11]	FLAME [35]	LSFM [8]	LYHM [18]	FS [68]	HIFI3D [4]	HIFI3D^A [4]	***Ours***
# scans	200	140	3800	8402	1212	938	200	200	1957
n_v	53215	11510	5023	53215	11510	26317	20481	20481	20481
n_f	105840	22800	9976	105840	22800	52261	40832	40832	40832
# basis	199	50	300	158	100	300	200	500	526

HIFI3D^A stands for the "augmented" version of HIFI3D [4], which employs data augmentation techniques to construct 3DMM from 200 scans.

vertex on S_j, and $T^{\text{pts}}_{i\to j}$ maps each vertex on shape S_i to a *point* in a face of shape S_j. We will use the superscript to disambiguate the two maps when necessary.

Normalized Mean Square Error (NMSE) computes the distance between two surfaces S_i and S_j based on a given map $T_{i\to j}$ and is denoted as $e(T_{i\to j})$:

$$e(T_{i\to j}) = \frac{1}{n_v}\sum_{v\in S_i} \| v - T_{i\to j}(v) \|_F^2 \tag{1}$$

where n_v is the number of vertices in shape S_i, and $T_{i\to j}(v)$ gives the *coordinates* of the mapped position (of a vertex/point on shape S_j) for vertex $v \in S_i$.

Iterative Closest Point (ICP) can be applied to align two shapes via solving a rigid transformation and a nearest neighbor map iteratively to minimize NMSE:

$$\min_{\mathbf{R},\mathbf{t},T_{i\to j}} \quad \sum_{v\in S_i} \| \mathbf{R}v + \mathbf{t} - T_{i\to j}(v) \|_F^2, \tag{2}$$

where the 3D rotation matrix $\mathbf{R}$ and the 3D translation vector $\mathbf{t}$ are computed to align S_i to S_j. For simplicity, we denote $[\mathbf{R}, \mathbf{t}, T_{i\to j}, S_i^*] = \mathbf{ICP}(S_i \to S_j)$, where S_i^* is obtained by transforming S_i via $(\mathbf{R}, \mathbf{t})$. For the purpose of efficiency, some previous works [38,17,24] only consider a small set of predefined corresponding keypoints to solve for $\mathbf{R}$ and $\mathbf{t}$ instead of using every vertex $v \in S_i$.

3.2 3DMM and Face Reconstruction

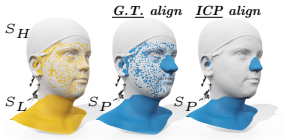
Formulation The goal is to reconstruct a 3D face shape S_P from a single RGB(-D) image or an image collection. To reduce the search space of plausible faces, 3DMM is commonly used for face representation: $S = \bar{S} + \Phi\alpha$, where $\bar{S}$ is the *mean* shape, and Φ is the *principle components* (PCs) trained on some 3D face scans with neutral expression. Then the face reconstruction problem is reduced to a regression problem of solving for the facial parameter α. Existing methods either use deep networks to predict the α [57,28,17,19,29,22] or optimize various energy terms (e.g., most commonly used keypoint loss and photometric loss [7,26,27,3,58,67]) with different regularization terms (e.g., not deviate too far from mean face [58,26,27,4,44]).

3DMM The facial basis Φ determines the expressiveness power of the corresponding 3DMM and affect the reconstruction quality. Publicly available 3DMMs include BFM [6,45], LSFM [8], FLAME [35], LYHM [18], FaceScape (FS) [68], FaceWareHouse (FWH) [11], and HIFI3D/HIFI3D^A [4] (see Tab. 1).

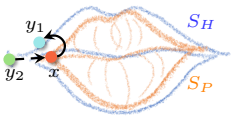
Standard Evaluation Pipeline Some datasets [2,46,52,24,75] also provide ground-truth scans, i.e., S_H with keypoints $\mathcal{K}_H$, that are associated with the input images, which allow us to evaluate the quality of the reconstructed shape S_P. Standard evaluation process consists of three steps: (1) first rescale and align S_P with S_H based on some sparse keypoints or applying ICP, (2) find the map $T_{p\to h}$ [20,41,47,50,16] or $T_{h\to p}$ [52,25,49,69] between the aligned shapes by nearest neighbor searching, (3) compute the NMSE of the nn-map $e(T_{p\to h})$ or $e(T_{h\to p})$.

4 Motivation

Multiple methods have been proposed to tackle the face reconstruction problem. However, to the best of our knowledge, there does not exist an effective evaluation protocol to fairly and reliably compare reconstructed faces from different methods. We observe the following issues in existing protocols:



(1) Global alignment is extremely sensitive to the provided keypoints and local changes. The NMSE metric based on global alignment often fails to reflect the true shape difference. Take the inset figure as an example: we have the ground-truth high-res mesh S_H colored in white, and the ground-truth low-res mesh S_L colored in yellow. We then modify the nose region of S_L while keeping the rest part unchanged, which leads to a toy predicted mesh S_P colored in blue. Perceptually, we expect S_P to be aligned with S_H in the same way as S_L (as shown in the middle) to compute the shape differences. However, ICP computes the alignment in a global way and as shown on the right, S_P is rotated backwards w.r.t. S_H, having the complete facial region of S_P behind S_H, which leads to exaggerated errors. In this case, considering region-based alignment can help to avoid global mismatch (see Fig. 4).



(2) Another limitation is that it is hard to establish accurate and meaningful correspondences based on the global *rigid* alignment from a single direction (i.e., from S_P to S_H only). For example, the inset figure shows two aligned lips and we are supposed to measure the NMSE on the nn-map from S_P to S_H. For the red vertex $x \in S_P$, its nearest neighbor in S_H is the blue vertex y_1. In this case, considering the map in the other direction, i.e., from S_H to S_P, can help to establish the correct correspondence from y_2 to x.

These observations inspire us to consider *region-based* and *bidirectional* alignment for establishing semantically more meaningful correspondences between the predicted mesh S_P to the ground-truth mesh S_H. However, to make this happen, meshes S_H with ground-truth region masks $\mathcal{R}_H$ are requested, which are not available in any of the open datasets [46,2,68,18,56,75]. To fill the gap, we propose a new benchmark REALY, which provides ground-truth high-res mesh S_H and retopologized mesh S_L in HIFI3D topology, with accurate ground-truth keypoints and region masks. To justify the usefulness of our benchmark, we show two applications: evaluating and comparing (1) the reconstruction quality

of faces obtained from 9 different methods in different choices of topology, (2) the expressiveness power of 8 different 3DMM basis for face regression.

Paper Structure Sec. 5 discusses the details of REALY benchmark and our 3DMM. Sec. 6 illustrates our novel region-aware bidirectional evaluation pipeline. In Sec. 7 we extensively justify the usefulness of REALY and the advantages of our evaluation pipeline over the standard one on face reconstruction task, and demonstrate that HIFI3D^{++} is more expressive than existing 3DMMs.

5 REALY: A New 3D Face Benchmark

Overview Our benchmark REALY contains 100 individuals, and each individual is modelled by a ground-truth high-resolution mesh S_H (the aligned 3D scan from LYHM [18]) and a retopologized low-resolution mesh S_L using HIFI3D [4] topology in neutral expression (see Fig. 2 for some examples). The meshes (S_H, S_L) of all individuals are consistently *scaled* and *aligned*. We also provide 68 keypoints and 4 region masks, which are semantically meaningful, for *both* S_H and S_L of *each* individual. For each individual, five high-quality and realistic multi-view images (including one frontal image) are rendered with well designed lighting condition and ground-truth camera parameters.

We explain the detailed construction procedure of REALY as follows.

HIFI3D Topology We choose this topology for our benchmark for the following reasons: (1) LYHM has overdense samplings at the boundary of the eyes and mouth. (2) LSFM does not have edge loops to define the contours of the eyes and mouth. (3) FLAME has unnatural triangulation which cannot model some realistic muscle movements such as raising the eyebrows. As a comparison, HIFI3D has better triangulation and balanced samplings to make realistic and nuanced expressions. Besides, HIFI3D also has eyeballs, interior structure of the mouth, and the shoulder region, which all benefit downstream applications such as talking head. See supplementary for visualized comparisons.

Construction Pipeline We start by collecting 1235 scans from LYHM and preparing a template shape S_{temp} in HIFI3D topology (see Fig. 1(c)) with predefined 68 keypoints $\mathcal{K}_{\text{temp}}$ and 4 region masks $\mathcal{R}_{\text{temp}}$ (including nose, mouth, forehead and cheek region). Firstly, we re-scale and rigidly align the input scans to the template shape S_{temp}, leading to our ground-truth high-resolution meshes S_H (i.e., aligned scans). We then follow [24,52] to define an evaluation region which is a disk centered on the nose tip. Secondly, we "wrap" (i.e., perform non-rigid registration) S_{temp} to each S_H to get the retopologized S_L such that S_L have the same topology as S_{temp} but reflect the shape of S_H. Note that we have keypoints $\mathcal{K}_L = \mathcal{K}_{\text{temp}}$ and regions $\mathcal{R}_L = \mathcal{R}_{\text{temp}}$ since S_L and S_{temp} share the same HIFI3D topology. We then transfer $\mathcal{K}_L$ and $\mathcal{R}_L$ from S_L to S_H for each individual. We also set up a rendering pipeline for synthesizing *multi-view* images for the textured high-resolution mesh S_H. Such a controlled environment enables REALY to focus on reflecting the reconstruction ability of different methods.

Finally, we filter out samples with wrapping error larger than 0.2mm and ask an expert artist with 3 year modeling experience to select 100 individuals among all the processed scans with the highest model quality, across different genders, ethnicity, and ages, to obtain our REALY benchmark.

Challenges & Solutions We observe two major challenges during the above construction procedure: (1) the raw scans are in different scales and poses with inaccurate sparse keypoints [74], which makes it difficult to align them consistently. To tackle this problem, we iterate through the following steps until convergence: first, render a frontal face image of S_H with texture using the initial/estimated transformation to align S_H to S_{temp} (note that the frontal pose needs to be determined from the alignment transformation as the frontal facing pose is unknown for a given scan); second, detect a set of 2D facial keypoints on the rendered image of S_H using state-of-the-art landmark detector; third, project the 2D keypoints into 3D using the rendering camera pose; fourth, update the alignment transformation from S_H to S_{temp} using the correspondences between the projected 3D keypoints on S_H and the known 3D keypoints on S_{temp}.

(2) Another challenge is, after we get the retopologized S_L, how to accurately transfer the region mask from the low-resolution mesh S_L (inherited from S_{temp}) to the high-resolution mesh S_H. One naive solution would be using nearest neighbor mapping from S_L to S_H to transfer the region mask. However, since the resolution of S_H can be $50\times$ larger than that of S_L, this naive solution will introduce disconnected and noisy region mask. To avoid such flaws, we use the vertex-to-point mapping from both directions to find candidate regions on S_H. As much more correspondences can be established during the mapping from S_H to S_L, higher-quality, smoother region masks can be obtained. Finally, we filter out noisy regions (e.g., nostril, eyeballs) and return the largest connected region.

HIFI3D^{++} With the above procedure, we can further construct a 3DMM basis by retopologizing more 3D face models. Specifically, based on the 200 individuals from HIFI3D [4], we additionally process and retopologize 3D face models of 846 individuals from FaceScape dataset [68] into HIFI3D topology. Together with the aforementioned processed models of 1235 individuals from LYHM [18], we collect and then select 1957 most representative meshes consisting of individuals from various ethnic groups. We then apply PCA [6] to obtain our new basis with 526 PCs (with 99.9% cumulative explained variance), which we name as HIFI3D^{++}. Tab. 1 shows the comparison of HIFI3D^{++} to other 3DMMs. Note that previous 3DMMs are more or less ethnics-biased. For example, BFM [45] is constructed mostly from Europeans, FLAME [35] is constructed from scans of the US and European, while HIFI3D [4] and FS [68] is constructed from scans of Asians. The LSFM and FLAME contains $50:1$ and $12:1$ of Caucasian and Asian respectively. In contrast, HIFI3D^{++} is constructed from high-quality models across more balanced ethnic groups that ensures $1:1$ between Caucasian and Asian (plus a few subjects from other ethnicities). We examine the expressive powers and reconstruction qualities of different 3DMMs in Sec. 7.3.

Algorithm 1 $S_P^* = \boldsymbol{r}\mathbf{ICP}\big(S_P \to S_H@\mathcal{R}_H\big)$

Goal Rigidly align S_P to the *region* R_H of S_H.

Input: High-res mesh S_H with region R_H and keypoints $\mathcal{K}_H$; a predicted mesh S_P with keypoints $\mathcal{K}_P$; weights $w_\mathcal{K}$ for keypoints alignment; maximum iteration $\mathbb{K}$.

$S_P^{(0)} = \boldsymbol{g}\mathbf{ICP}\ (S_P \to S_H)$
for $0 \le k \le \mathbb{K}$ **do**
  Find nn-map T from region $\mathcal{R}_H$ to $S_P^{(k)}$.
  Solve $[\mathbf{R}, \mathbf{t}] = \arg\min \sum_{v \in \mathcal{R}_H} \big\|\mathbf{R}T(v) + \mathbf{t} - v\big\|_F^2 + w_\mathcal{K}\big\|\mathbf{R}\mathcal{K}_P + \mathbf{t} - \mathcal{K}_H\big\|_F^2$.
  Obtain $S_p^{(k+1)}$ by transforming $S_p^{(k)}$ via $(\mathbf{R}, \mathbf{t})$.
end for
Set $S_P^* \leftarrow S_P^{(\mathbb{K})}$.

Algorithm 2 $\big[S_P^*, \mathcal{R}_H^*\big] = \boldsymbol{b}\mathbf{ICP}\big(S_P \leftrightarrow S_H@\mathcal{R}_H\big)$

Goal Rigidly align S_P and non-rigidly deform $\mathcal{R}_H$ for better alignment in region $\mathcal{R}_H$

$S_P^* = \boldsymbol{r}\mathbf{ICP}\big(S_P \to S_H@\mathcal{R}_H\big)$ *% call Algo. 1*
Find nn-map T from region $\mathcal{R}_H$ to S_P^*
$\mathcal{R}_H^* = \boldsymbol{n}\mathbf{ICP}\Big(\mathcal{R}_H \,\big|\, V_{\mathcal{R}_H} \to T\big(V_{\mathcal{R}_H}\big)\Big) + w_\mathcal{K}\ \boldsymbol{n}\mathbf{ICP}\Big(\mathcal{R}_H \,\big|\, \mathcal{K}_H \to \mathcal{K}_P\Big)$ *% where* $\boldsymbol{n}\mathbf{ICP}(S \,|\, X \to Y)$ *applies non-rigid ICP to deform* S *such that the points* X *on* S *are expected to be mapped to new positions* Y

6 A Novel Evaluation Pipeline

To evaluate the quality of a reconstructed or predicted face S_P, the standard pipeline first globally aligns S_P with the ground-truth high-resolution mesh S_H to find the nearest-neighbor map $T_{p\to h}$ (or $T_{h\to p}$). The similarity or reconstruction error is then measured by the NMSE error $e(T_{p\to h})$ (or $e(T_{h\to p})$). We propose a new evaluation pipeline based on a region-aware and bidirectional alignment.

We focus on how to accurately establish correspondences between S_P and a particular region $\mathcal{R}_H$ in the ground-truth shape S_H (denoted as $S_H@\mathcal{R}_H$ for short), which consists of two main steps: (1) $\boldsymbol{r}\mathbf{ICP}$ (region-aware ICP): we first get the rigidly transformed shape S_P^* from S_P by aligning S_P to $\mathcal{R}_H$ such that the corresponding *region* on S_P is well aligned to $\mathcal{R}_H$ without taking the rest part of the face into consideration (see Algo. 1). (2) $\boldsymbol{b}\mathbf{ICP}$ (non-rigid and bidirectional ICP): with the above established correspondences between S_P^* and $\mathcal{R}_H$ as initialization, we further refine the correspondences by applying non-rigid ICP ($\boldsymbol{n}\mathbf{ICP}$) [1] to deform $\mathcal{R}_H$ to fit S_P^* (see Algo. 2). This step yields a deformed shape $\mathcal{R}_H^*$ from $\mathcal{R}_H$, as well as the correspondences between $\mathcal{R}_H$ and S_P^* induced from $\mathcal{R}_H^*$ (using vertex-to-surface projection). Our region-wise alignment and the two step coarse-to-fine registration effectively guarantee that $\boldsymbol{n}\mathbf{ICP}$ can converge to a reasonable deformed shape $\mathcal{R}_H^*$ (see supplementary for details). The resulting correspondences are used for computing errors between $\mathcal{R}_H$ and S_P^*. Note that the second alignment step can be regarded as upsampling S_P^* such that it has a similar resolution to the dense ground-truth scan $S_H@\mathcal{R}_H$, which makes the evaluation of reconstructed meshes in different resolutions easier.

We observe that the correspondences established between the rigidly transformed shape S_P^* and deformed region $\mathcal{R}_H^*$ are more accurate than the corre-

spondences detected between the original shapes S_P and S_H for evaluating the similarity between the two shapes in region $\mathcal{R}_H$. Fig. 3 shows such an example. We visualize the corresponding points on S_H of the mouth keypoints on S_P in blue/green/yellow established by ***g*ICP**/***r*ICP**/***b*ICP** respectively, where the red points are the ground-truth mouth keypoints on S_H.

We can see that ***b*ICP** gives more accurate correspondences since it focus on the mouth region and considers correspondences from both directions. As a result, y_3 obtained by ***b*ICP** is closer to the ground-truth y, while y_1 obtained by ***g*ICP** is far from y. We validate the above analysis with detailed experiments in Sec. 7.1. We use the above established correspondences to measure the NMSE error of the region-wise aligned S_P^* compared to ground-truth S_H on four predefined regions including nose, mouth, forehead, and cheek, respectively (see Fig. 1d). The NMSE error is then transformed back to the physical scale of the raw scan in millimeters (note that each ground-truth S_H in our benchmark is rigidly transformed from raw scan, see Sec. 5). Evaluation on each region individually provides us fine-grained understandings of the qualities of the reconstructed meshes. We present extensive experiments for benchmarking state-of-the-art single-image 3D face reconstruction methods in Sec. 7.2.

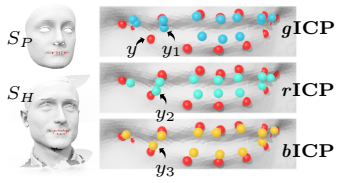


Fig. 3. Correspondence established by different ICPs and the GT correspondences (red).

7 Experiment

In this section we first demonstrate the effectiveness of our ***b*ICP** which can establish more accurate correspondences than standard ICP for reliably evaluating 3D face reconstructions. We then compare different face reconstruction methods and 3DMMs using our evaluation protocol to investigate fine-grained shape differences based on regions in a systematically way.

7.1 Ablation Study: *b*ICP v.s. *g*ICP

To demonstrate the advantages of our region-based and bidirectional evaluation protocol over the standard one, we design a controlled experiment illustrated in Fig. 4 and Tab. 2 and 3. Specifically, we carefully construct four predicted shapes S_P by modifying the same ground-truth mesh S_L such that the ground-truth correspondences between S_P and S_L are known for reference.

As illustrated in Fig. 4, we replace the nose/mouth/forehead/cheek region of shape S_L with the corresponding region from four different reference shapes $S_i (i = 1, 2, 3, 4)$ respectively. We then visualize the shape differences computed using ***b*ICP** (w.r.t. the four specified regions) and ***g*ICP** (w.r.t. the complete face) in Fig. 4 where large (small) errors are colored in red (blue).

We compare our evaluation pipeline to the standard one in two-fold: (1) we report the alignment error, the distance between the aligned S_P and the

Table 2. The distance $e(\cdot)$ between S_P (aligned by ***r*ICP**/***g*ICP**, or with G.T. alignment) and S_L of the four examples in Fig. 4.

$\mathcal{R}^{\text{rm}}$ / e(mm/10)	***r*ICP (ours)** @nose	@mouth	@forehead	@cheek	all	***g*ICP** $e(T^{\text{vtx}}_{p\to l})$	**G.T.** $e(T^{\text{vtx}}_{p\to l})$
nose	**8.376**	0.331	1.200	1.484	11.392	40.490	11.972
mouth	0.550	**4.372**	1.164	3.053	9.139	20.889	5.195
forehead	0.636	0.165	**7.107**	0.630	8.537	12.695	6.463
cheek	1.417	0.397	0.547	**3.631**	5.992	18.754	3.943

Table 3. Error of the nn-map obtained via different ICPs (Fig. 4).

$\mathcal{R}^{\text{rm}}$ / e(mm)	***g*ICP**	***r*ICP**	***b*ICP**
nose	2.882	0.706	**0.670**
mouth	1.699	0.459	**0.407**
forehead	0.707	0.581	**0.520**
cheek	1.219	0.107	**0.105**

input S_L using ground-truth correspondences in Tab. 2, where the alignment is computed using our region-wise ***r*ICP** and the global-wise ***g*ICP**. (2) we report the accuracy of correspondences established via ***g*ICP**, ***r*ICP** (our intermediate step), and ***b*ICP** in Tab. 3 compared to the ground-truth correspondences.

We can see that ***g*ICP** is extremely sensitive to local changes. For example, replacing the nose only can lead to errors in the complete face (first row in Fig. 4) and replacing the mouth can lead to errors even in forehead (second row) according to ***g*ICP**. As a comparison, our ***b*ICP** can correctly localizes the shape differences in the modified regions and lead to more accurate alignment than ***g*ICP** as shown in Tab. 2. This suggests that region-based alignment can better quantify shape differences especially in the case of local or subtle changes. Moreover, Tab. 3 shows that both our ***r*ICP** and ***b*ICP** help to find more accurate correspondences to evaluate shape differences.

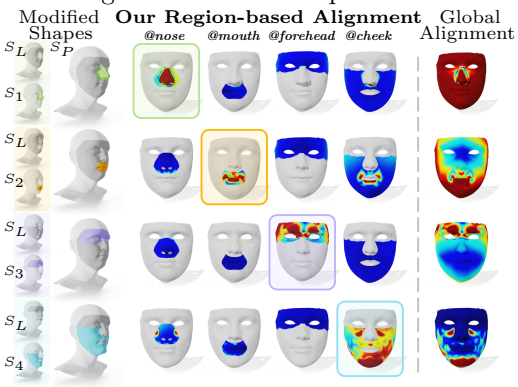


Fig. 4. We visualize the shape difference between S_P and S_L using our ***b*ICP** (shown on separate regions) and ***g*ICP**, where S_P is constructed by replacing S_L's nose/mouth/forehead/cheek region with the corresponding region from $S_1/S_2/S_3/S_4$.

7.2 Evaluating Face Reconstruction Methods

We compare recent state-of-the-art face reconstruction methods on our REALY benchmark using our new evaluation protocol including: (1) linear-3DMM based methods: ExpNet [59,14,15], RingNet [52], MGCNet [54], Deep3D [19], 3DDFA-v2 [29], GANFit [26,27], DECA-coarse [22], and (2) non(linear)-3DMM methods: PRNet [23], Nonlinear 3DMM (N-3DMM) [61,62,60].

We report the statistics (mean and std.) of errors over 100 shapes in REALY using our ***b*ICP** (in each separate region and complete face) and using standard ***g*ICP** (in complete face) in Tab. 4. ***g*ICP** suggests that DECA and Deep3D are the best among the tested methods for face reconstruction. However, our ***b*ICP** suggests that DECA models the nose region in a much better and more accurate way than others, but it obtains less satisfactory result in the mouth region. Fig. 5

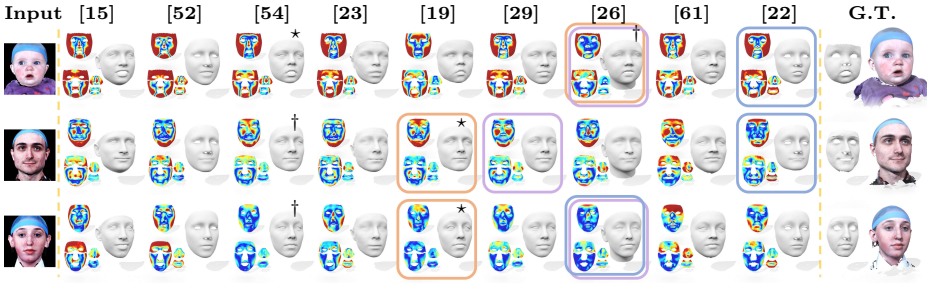


Fig. 5. Comparing different face reconstruction methods. We visualize the reconstruction error of each face using the standard evaluation pipeline (top left) and our novel evaluation pipeline (bottom left, shown in four regions), where large (small) errors are colored in red (blue). Note only the cropped region shown in G.T. is counted for evaluation. The best reconstructed face selected by our measurement (orange boxes) is visually closer to the ground-truth meshes than the ones selected using the standard measurements (blue & purple boxes). See more examples in our supplementary.

Table 4. Comparing different face reconstruction methods on REALY. We report the statistics of errors measured using our new pipeline (***b*ICP**) and the standard one (***g*ICP** from both directions). The best (second best) method w.r.t. average error is highlighted in red (blue).

methods / e (mm)	***b*ICP** (ours)									***g*ICP**			
	$@\mathcal{R}_N$(nose)		$@\mathcal{R}_M$(mouth)		$@\mathcal{R}_F$(forehead)		$@\mathcal{R}_C$(cheek)		all	$e(T^{pts}_{p\to h})$		$e(T^{pts}_{h\to p})$	
	avg.	std.	avg.	std.	avg.	std.	avg.	std.	avg.	avg.	std.	avg.	std.
ExpNet [15]	2.509	0.486	1.912	0.450	3.084	1.005	1.717	0.590	2.306	2.650	0.549	2.297	0.616
RingNet [52]	1.934	0.458	2.074	0.616	2.995	0.908	2.028	0.720	2.258	2.016	0.489	2.762	1.016
MGCNet [54]	1.771	0.380	**1.417**	0.409	**2.268**	0.503	1.639	0.650	**1.774**	2.388	0.865	2.094	0.750
PRNet [23]	1.923	0.518	1.838	0.637	2.429	0.588	1.863	0.698	2.013	3.036	0.933	2.302	0.747
Deep3D [19]	**1.719**	0.354	**1.368**	0.439	**2.015**	0.449	1.528	0.501	**1.657**	2.142	0.651	**1.908**	0.553
3DDFA-v2 [29]	1.903	0.517	1.597	0.478	2.477	0.647	1.757	0.642	1.926	2.788	0.951	2.279	0.765
GANFit [26]	1.928	0.490	1.812	0.544	2.402	0.545	**1.329**	0.504	1.868	**1.899**	0.730	**1.999**	0.748
N-3DMM [61]	2.936	0.810	2.375	0.599	4.582	1.448	1.918	0.801	2.953	3.681	1.566	3.252	1.198
DECA-c [22]	**1.697**	0.355	2.516	0.839	2.349	0.576	**1.479**	0.535	2.010	**1.698**	0.397	2.183	0.798

shows some qualitative results, where the best reconstructed face selected using ***b*ICP** (***g*ICP**) is highlighted in orange (blue&purple) box. We also conduct user study to ask people to vote for the best (labeled $\star$) and second best (labeled †) face. We can see that the faces selected by ***b*ICP** are indeed visually more similar (validated by our user study) to the G.T. shapes than those selected by ***g*ICP**.

Moreover, another advantage of our ***b*ICP** evaluation protocol is its region-aware nature, which allows us to compare different methods in some particular region. Fig. 6 shows such an example, where we select the best matched region from different methods according to our region-aware ***b*ICP** and merge them into a new face. Perceptually, the merged face is clearly better than the face reconstructed using DECA (selected by ***g*ICP**).

7.3 Evaluating Different 3DMMs

We use our new evaluation approach to compare different 3DMMs on REALY including: LYHM [18], BFM [45], FLAME [35], LSFM [8], FaceScape basis

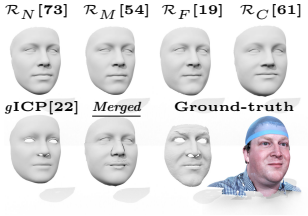


Fig. 6. We use our region-aware metric to find the best matched nose $\mathcal{R}_N$, mouth $\mathcal{R}_M$, cheek $\mathcal{R}_C$, and forehead $\mathcal{R}_F$ from existing methods, and merge them into a new face. Our merged face is more similar to the G.T. shape than DECA, the best reconstructed face voted by ***g*ICP**.

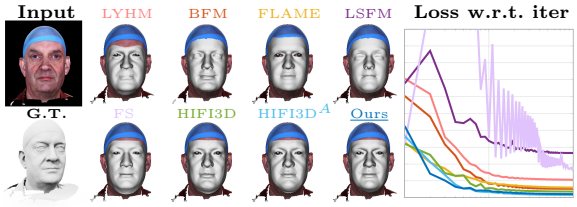


Fig. 7. Comparing the fitting errors of different 3DMMs. We visualize an example of the RGB-D fitting errors during the optimization iterations using different 3DMMs and the corresponding reconstructed faces. We observe that HIFI3D^A and HIFI3D^{++} give the most realistic reconstructed faces (especially in the mouth region) and show superior converging rate with smallest converged loss among the tested 3DMMs. As a comparison, FS and LSFM show instability during the fitting process and lead to erroneous reconstruction results.

Table 5. Comparing different 3DMM basis on REALY. The best (second best) methods w.r.t. average error are highlighted in red (blue).

3DMMs / e (mm)	*RGB Fitting*									*RGB-D Fitting*								
	@$\mathcal{R}_N$		@$\mathcal{R}_M$		@$\mathcal{R}_F$		@$\mathcal{R}_C$		all	@$\mathcal{R}_N$		@$\mathcal{R}_M$		@$\mathcal{R}_F$		@$\mathcal{R}_C$		all
	avg.	std.	avg.	std.	avg.	std.	avg.	std.	avg.	avg.	std.	avg.	std.	avg.	std.	avg.	std.	avg.
BFM [45]	2.925	0.704	2.175	0.550	3.359	0.660	1.742	0.410	2.550	1.700	0.277	1.170	0.355	2.308	0.501	0.587	0.100	1.441
FLAME [35]	**2.700**	0.543	2.616	0.476	3.891	0.786	2.737	0.687	2.986	1.687	0.232	1.397	0.354	2.178	0.609	0.495	0.125	1.439
LSFM [8]	**2.455**	0.666	2.446	0.768	4.062	0.807	3.756	1.292	3.180	1.727	0.320	1.906	0.638	2.370	0.612	0.869	0.218	1.718
FS [68]	2.852	0.776	2.524	0.827	**2.430**	0.613	1.739	0.450	2.386	2.181	0.494	2.468	0.866	2.057	0.597	1.003	0.208	1.927
HIFI3D [4]	2.974	0.752	**1.285**	0.364	2.519	0.490	2.070	0.533	2.212	**1.653**	0.258	0.909	0.332	1.343	0.366	0.468	0.121	1.093
HIFI3D^A [4]	3.076	0.709	**1.201**	0.399	2.527	0.561	1.866	0.566	**2.167**	1.746	0.271	**0.607**	0.338	**1.235**	0.363	**0.302**	0.084	**0.972**
LYHM* [18]	2.723	0.578	1.988	0.556	3.752	0.716	**1.475**	0.439	2.485	2.144	0.331	1.654	0.520	3.174	0.676	0.673	0.155	1.911
Ours*	2.898	0.732	1.288	0.408	**2.216**	0.612	**1.599**	0.537	**2.000**	**1.542**	0.258	**0.621**	0.341	**1.085**	0.359	**0.265**	0.080	**0.878**

*Some of the test shapes in REALY are used to construct LYHM and our 3DMM.

(FS) [68], HIFI3D and HIFI3D^A [4]. In this test, we use different basis and run standard RGB(-D) fitting algorithm [4] using photo loss (with/without depth loss), ID loss, landmark loss and regularization loss to regress the 3DMM coefficients from the given 2D images provided by REALY. For RGB fitting, we use a frontal face image of each individual as input. For RGB-D fitting, a frontal rendered depth image in addition to the RGB image is used. Fig. 7 shows the fitting errors over iterations using different 3DMMs. The qualitative and quantitative comparisons are presented in Fig. 8 and Tab. 5, respectively. Note that both LYHM and our 3DMM basis in Tab. 5 use some test shapes in REALY to construct the 3DMM. They are listed in the table only for reference.

As shown in Tab. 5, HIFI3D^A, HIFI3D, and FS achieve similar performance in RGB fitting, while HIFI3D^A and HIFI3D perform much better than the other 3DMMs in RGB-D fitting, indicating that HIFI3D^A and HIFI3D are more expressive and can better fit the geometry especially in RGB-D fitting with extra depth information. To directly justify the expressive power of our new 3DMM basis HIFI3D^{++}, we compare HIFI3D^{++} to HIFI3D/HIFI3D^A over 3 shapes that are unused by all three 3DMMs. We fit the 3D ground-truth scans using

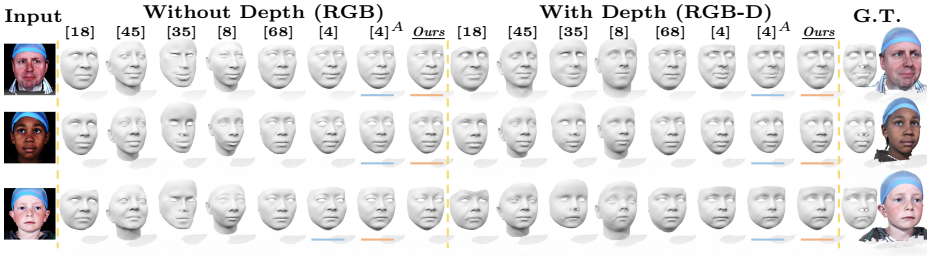


Fig. 8. Comparing different 3DMMs. From *Left to right*: LYHM [18], BFM [45], FLAME [35], LSFM [8], FS [68], HIFI3D [4], HIFI3D^A [4], and HIFI3D^{++}. We highlight the best (second best) reconstructed face via red (blue) underline. Only the cropped region shown in G.T. is counted for evaluation. See more examples in supplementary.

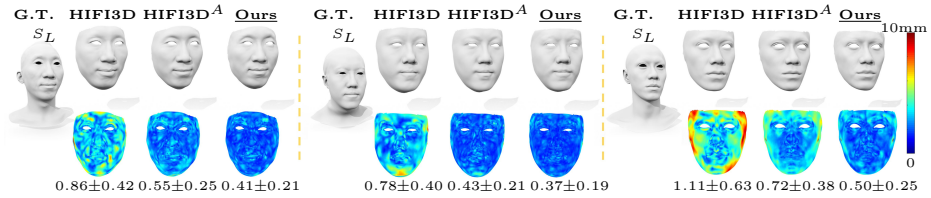


Fig. 9. Comparing our 3DMM to HIFI3D and HIFI3D^A. *Top*: Fitted faces S_P, *Bottom*: Fitting errors, where large (small) errors are colored in red (blue).

the 3DMMs. Fig. 9 shows the error heatmaps. The average error with our new basis decreases over 20% compared to HIFI3D^A.

Finally, an important observation from Tabs. 4-5 is that, while top performing single image reconstruction methods achieve results of $1.657 \sim 2.953$, the best RGB-D fitting records 0.878, which is far better than them. It implies that there is still much room for improvements in single image reconstruction methods.

8 Conclusions

In this work, we introduce a new benchmark REALY for 3D face reconstruction that provides accurate and consistent facial keypoints, region masks on the scans, and consistently retopologized meshes. During the construction procedure of the benchmark, we also derive a new powerful 3DMM basis HIFI3D^{++}. The benchmark allows us to design a novel region-aware and bidirectional evaluation pipeline to measure shape similarity, which is justified to be more reliable than the standard evaluation pipeline based on global alignment. Furthermore, we compare and analyse existing single-image face reconstruction methods and state-of-the-art 3DMM basis using our new evaluation approach on REALY, which is the first to obtain fine-grained region-wise analyses in the 3D face community. Moreover, it would be interesting to research how our benchmark can be used for supervised learning of face reconstruction.

Acknowledgment. This work was supported by SZSTC Grant No. JCYJ20190809172201639 and WDZC20200820200655001, Shenzhen Key Laboratory ZDSYS20210623092001004.

References

1. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid ICP algorithms for surface registration. In: CVPR (2007)
2. Bagdanov, A.D., Bimbo, A.D., Masi, I.: The florence 2d/3d hybrid face dataset. In: J-HGBU@MM (2011)
3. Bai, Z., Cui, Z., Liu, X., Tan, P.: Riggable 3d face reconstruction via in-network optimization. In: CVPR (2021)
4. Bao, L., Lin, X., Chen, Y., Zhang, H., Wang, S., Zhe, X., Kang, D., Huang, H., Jiang, X., Wang, J., Yu, D., Zhang, Z.: High-fidelity 3d digital human head creation from rgb-d selfies. TOG (2021)
5. Besl, P.J., McKay, N.D.: A method for registration of 3d shapes. TPAMI (1992)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)
7. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. TPAMI (2003)
8. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: CVPR (2016)
9. Brunton, A., Salazar, A., Bolkart, T., Wuhrer, S.: Review of statistical shape spaces for 3d data with comparative analysis for human faces. CVIU (2014)
10. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. TOG (2013)
11. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. TVCG (2014)
12. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. TOG (2016)
13. Cao, K., Rong, Y., Li, C., Tang, X., Loy, C.C.: Pose-robust face recognition via deep residual equivariant mapping. In: CVPR (2018)
14. Chang, F., Tran, A.T., Hassner, T., Masi, I., Nevatia, R., Medioni, G.G.: Faceposenet: Making a case for landmark-free face alignment. In: ICCV Workshops (2017)
15. Chang, F., Tran, A.T., Hassner, T., Masi, I., Nevatia, R., Medioni, G.G.: Expnet: Landmark-free, deep, 3d facial expressions. In: FG (2018)
16. Chaudhuri, B., Vesdapunt, N., Shapiro, L.G., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting. In: ECCV (2020)
17. Chen, Y., Wu, F., Wang, Z., Song, Y., Ling, Y., Bao, L.: Self-supervised learning of detailed 3d face reconstruction. TIP (2020)
18. Dai, H., Pears, N.E., Smith, W.A.P., Duncan, C.: Statistical modeling of craniofacial shape and texture. IJCV (2020)
19. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: CVPR Workshops (2019)
20. Dib, A., Thebault, C., Ahn, J., Gosselin, P., Theobalt, C., Chevallier, L.: Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In: ICCV (2021)
21. Egger, B., Smith, W.A.P., Tewari, A., Wuhrer, S., Zollhöfer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3d morphable face models - past, present, and future. TOG (2020)
22. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. SIGGRAPH (2021)

23. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV (2018)
24. Feng, Z., Huber, P., Kittler, J., Hancock, P., Wu, X., Zhao, Q., Koppen, P., Rätsch, M.: Evaluation of dense 3d reconstruction from 2d face images in the wild. In: FG (2018)
25. Gao, Z., Zhang, J., Guo, Y., Ma, C., Zhai, G., Yang, X.: Semi-supervised 3d face representation learning from unconstrained photo collections. In: CVPR Workshops (2020)
26. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: GANFIT: generative adversarial network fitting for high fidelity 3d face reconstruction. In: CVPR (2019)
27. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. TPAMI (2021)
28. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: CVPR (2018)
29. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: ECCV (2020)
30. Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y., Li, H.: Avatar digitization from a single image for real-time rendering. TOG (2017)
31. Jiang, D., Jin, Y., Deng, R., Tong, R., Zhang, F., Yai, Y., Tang, M.: Reconstructing recognizable 3d face shapes based on 3d morphable models. CoRR, abs/2104.03515 (2021)
32. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
33. Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., Zafeiriou, S.: Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In: CVPR (2020)
34. Lee, G., Lee, S.: Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In: CVPR (2020)
35. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. TOG (2017)
36. Lin, J., Yuan, Y., Shao, T., Zhou, K.: Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In: CVPR (2020)
37. Lin, J., Yuan, Y., Zou, Z.: Meingame: Create a game character face from a single portrait. In: AAAI (2021)
38. Liu, F., Zhu, R., Zeng, D., Zhao, Q., Liu, X.: Disentangling features in 3d face shapes for joint face reconstruction and recognition. In: CVPR (2018)
39. Liu, P., Han, X., Lyu, M.R., King, I., Xu, J.: Learning 3d face reconstruction with a pose guidance network. In: ACCV (2020)
40. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
41. Luo, H., Nagano, K., Kung, H., Xu, Q., Wang, Z., Wei, L., Hu, L., Li, H.: Normalized avatar synthesis using stylegan and perceptual refinement. In: CVPR (2021)
42. Lyu, J., Li, X., Zhu, X., Cheng, C.: Pixel-face: A large-scale, high-resolution benchmark for 3d face reconstruction. arXiv preprint arXiv:2008.12444 (2020)
43. Ma, S., Simon, T., Saragih, J.M., Wang, D., Li, Y., la Torre, F.D., Sheikh, Y.: Pixel codec avatars. In: CVPR (2021)
44. Pan, X., Dai, B., Liu, Z., Chen, C.L., Luo, P.: Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In: ICLR (2021)
45. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: AVSS (2009)

46. Phillips, P.J., Flynn, P.J., Scruggs, W.T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.J.: Overview of the face recognition grand challenge. In: CVPR (2005)
47. Piao, J., Sun, K., Wang, Q., Lin, K., Li, H.: Inverting generative adversarial renderer for face reconstruction. In: CVPR (2021)
48. Ploumpis, S., Wang, H., Pears, N.E., Smith, W.A.P., Zafeiriou, S.: Combining 3d morphable models: A large scale face-and-head model. In: CVPR (2019)
49. R, M.B., Tewari, A., Seidel, H., Elgharib, M., Theobalt, C.: Learning complete 3d morphable face models from images and videos. In: CVPR (2021)
50. Ramon, E., Triginer, G., Escur, J., Pumarola, A., Garcia, J., i Nieto, X.G., Moreno-Noguer, F.: H3d-net: Few-shot high-fidelity 3d head reconstruction. In: ICCV (2021)
51. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 3DV (2016)
52. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: CVPR (2019)
53. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: ICCV (2017)
54. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: ECCV (2020)
55. Smith, W.A.P., Seck, A., Dee, H., Tiddeman, B., Tenenbaum, J.B., Egger, B.: A morphable face albedo model. In: CVPR (2020)
56. Stratou, G., Ghosh, A., Debevec, P.E., Morency, L.P.: Effect of illumination on automatic expression recognition: A novel 3d relightable facial database. In: FG (2011)
57. Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: ICCV (2017)
58. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of RGB videos. In: CVPR (2016)
59. Tran, A.T., Hassner, T., Masi, I., Medioni, G.G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In: CVPR (2017)
60. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: CVPR (2019)
61. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: CVPR (2018)
62. Tran, L., Liu, X.: On learning 3d face morphable model from in-the-wild images. TPAMI (2021)
63. Wen, Y., Liu, W., Raj, B., Singh, R.: Self-supervised 3d face reconstruction via conditional estimation. In: ICCV (2021)
64. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. TPAMI (2009)
65. Wu, F., Bao, L., Chen, Y., Ling, Y., Song, Y., Li, S., Ngan, K., Liu, W.: Mvf-net: Multi-view 3d face morphable model regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 959–968 (2019)
66. Wu, S., Rupprecht, C., Vedaldi, A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: CVPR (2020)
67. Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S., Li, H.: High-fidelity facial reflectance and geometry inference from an unconstrained image. TOG (2018)

68. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: CVPR (2020)
69. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: CVPR (2021)
70. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: FG (2006)
71. Zeng, X., Peng, X., Qiao, Y.: Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In: ICCV (2019)
72. Zhang, Z., Ge, Y., Chen, R., Tai, Y., Yan, Y., Yang, J., Wang, C., Li, J., Huang, F.: Learning to aggregate and personalize 3d face from in-the-wild photo collection. In: CVPR (2021)
73. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: A 3d total solution. TPAMI (2019)
74. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR (2012)
75. Zhu, X., Yang, F., Huang, D., Yu, C., Wang, H., Guo, J., Lei, Z., Li, S.Z.: Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In: ECCV (2020)
76. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. CGF (2018)