Supplementary Material for 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

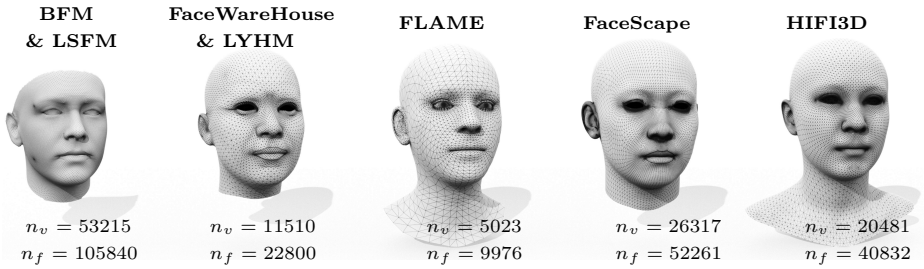

Fig. 1. Topology of different 3DMMs, where n_v and n_f represent the number of vertices and triangles respectively. In our REALY benchmark, we choose HIFI3D topology since it has better triangulation and balanced samplings with eyeballs and shoulder regions.

In this supplementary material, we provide additional technical details, qualitative examples, and discussions that could not be fitted into the main paper due of lack of space, which is organized as follows: We first give full details of how to construct our new benchmark REALY in Sec. 1. Additional experiments and results can be found in Sec. 2, where we justify the quality of REALY. In Sec. 3 we discuss the implementation details and choices of parameters. Finally, we discuss the limitation and future work in Sec. 4.

1 Details of Constructing REALY Benchmark

1.1 Preparing the Template Shape S_{temp}

We first prepare a template shape S_{temp} which is crucial for registering and retopologizing high-resolution scans from different datasets. We take the mean shape from HIFI3D [2] as our template shape that contains 20481 vertices and 40382 triangles. We then ask an *experienced* artist to label semantically meaningful and important keypoints $\mathcal{K}_{\text{temp}}$ and region masks $\mathcal{R}_{\text{temp}}$ since they play an important role in our proposed ***b*ICP** based face similarity evaluation.

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

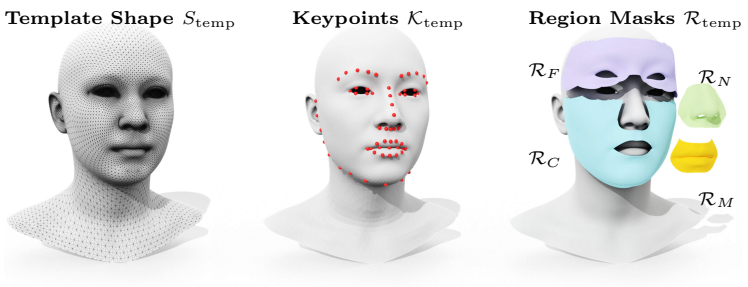

Fig. 2. We predefine 68 keypoints and 4 region masks on the template shape for constructing REALY.

HIFI3D Topology We choose HIFI3D for the following reasons: (1) BFM (LSFM) does not have edge loops to define the contours of the eyes and mouth. (2) FaceWareHouse (LYHM) has overdense samplings around the boundary of the eyes and mouth. (3) FLAME has unnatural triangulation which cannot model some realistic muscle movements such as raising the eyebrows. (4) FaceScape does not have eyeballs, interior structure of the mouth, or the shoulder region, which limits the expressiveness of different expressions. As a comparison, HIFI3D has better triangulation and balanced samplings to make realistic and nuanced expressions. Besides, HIFI3D also has independent eyeballs, interior structure of the mouth, and the shoulder region, which all benefit downstream applications such as talking head generation. Please see Fig. 1 for the topology of each 3DMM mentioned in Tab. 1 in the main paper.

Region Masks Four region masks are annotated in the S_{temp}, namely the nose region $\mathcal{R}_N$, the mouth region $\mathcal{R}_M$, the forehead region $\mathcal{R}_F$, and the cheek region $\mathcal{R}_C$. Each region mask is defined as a list of vertices and faces with smooth boundary (thanks to the good structural triangulation of HIFI3D topology). When constructing the region masks, we particularly *exclude* the ear, eyeball, nostril regions because these regions might not be reconstructed in some reconstruction methods or not considered in some 3DMMs. We also include some *overlapping* regions between two adjacent masks to avoid boundary instability during evaluation. Please see Fig. 2 for an illustration.

Keypoints We prepare three sets of keypoints on S_{temp} for different use cases: (1) **Keypoints for alignment and wrapping**. We ask experts to manually label 118 keypoints on the facial region of the template shape, including 24 keypoints on the eyebrow, 48 on the eyelids, 10 on the nose and nose bridge, 36 on the mouth. This set of keypoints is used to align and retopologize the input scans (elaborated in Sec. 1.2 and Sec. 1.4 respectively). (2) **Keypoints for evaluation**. Existing methods [16,14,12,7,23,19] usually include 68 semantically meaningful keypoints for evaluation or defining landmark loss for training. To setup a comparable setting, we also prepare 68 keypoints with the same semantic information as previous work, including 10 keypoints on the eyebrows, 12 on the

Fig. 3. *Left*: inaccurate keypoints (e.g., in the nose) provided in LYHM dataset, which are extracted using the mixture-of-trees algorithm [24]. *Right*: our high-quality keypoints obtained from a state-of-the-art landmark detector for global alignment & registration.

eyelids, 9 on the nose and nose bridge, 20 on the mouth, and 17 on the cheek contour. This set of keypoints will be transferred to the ground-truth scans and the retopologized meshes for evaluation (e.g., used in ***g*ICP**, ***r*ICP** and ***b*ICP** in our evaluation pipeline as introduced in Sec. 6 in our main paper). We particularly denote this set of keypoints as $\mathcal{K}_{\text{temp}}$. (3) **Keypoints for 3DMM fitting**. We follow [2] to prepare 86 keypoints including, 18 on the eyebrows, 16 on the eyelids, 15 on the nose and nose bridge, 20 on the mouth, and 17 on the cheek contour. This set of keypoints will be transferred to our newly introduced basis HIFI3D^{++} for 3DMM fitting.

1.2 Aligning Scans to S_{temp}

To construct our benchmark REALY and 3DMM basis HIFI3D^{++}, we need to collect and register large set of scans from different datasets [21,6,2]. However these scans are in random pose and scales. For example, the surface area of scans in LYHM [6] dataset ranges from $82,913$ mm^2 to $340,916$ mm^2, while the scans in FaceScape [21] may have opposite orientations.

Therefore, our first step is to rescale and align the input scans to the template shape S_{temp}. Specifically, for a given scan S_H, we first rescale and align it to S_{temp} using the provided keypoints from the source dataset. However, those provided keypoints are not accurate enough, as shown on the left of Fig. 3, which leads to unsatisfactory alignment. To tackle this problem, we iterate through the following steps until convergence: (1) render a frontal face image of S_H with texture (3k+ resolution) 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); (2) detect 256 2D facial keypoints on the rendered image of S_H using a state-of-the-art

landmark detector[3]; (3) project the 2D keypoints into 3D using the rendering camera pose; (4) 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}. Note that we solve for a scale factor, rotation matrix, and translation vector for the shape transformation.

1.3 Synthesizing Multi-view Images

Since all the high-resolution scans have been aligned to the template shape S_{temp}, which has known position and orientation, we can now synthesize multi-view images for each scan in a controlled setting. Specifically, We render the input scan with its corresponding texture on black background through a *perspective camera*. We fix the intrinsic parameters ($f_x = 2500, f_y = 2500, c_x = 512, c_y = 512$) of the camera and change the extrinsic parameters and lighting conditions to get a set of multi-view images in 1024×1024 resolution, including a frontal image and 4 images in *random* poses (with angles less than 20 degree). We also generate the ground-truth depth map for each image and record the ground-truth camera parameters. Our generated multi-view RGB-D image collection makes our benchmark suitable for evaluating face reconstruction methods under various input settings (i.e., single/multi-view RGB(-D) images). Fig. 4 shows some examples of the generated images in our REALY benchmark.

1.4 Retopologizing the Aligned Scans

For each scan S_H, we wrap the template shape S_{temp} to obtain S_L, a ground-truth mesh in relatively low resolution with consistent topology across different individuals. Recall that S_{temp} adopts the HIFI3D topology that contains 20481 vertices and 40832 triangles together with 3 sets of predefined keypoints and facial region masks. We follow [4] to retopologize the input scan in neutral expression via a two-step approach. (1) The facial region of S_{temp} is deformed to fit the facial region of S_H using non-rigid ICP technique [1]. The total energy on mesh deformation includes a smoothness term and a landmark loss term, where the predefined keypoints on S_{temp} (the first set of keypoints introduced in Sec. 1.1) are forced to be as close as possible to the automatically detected keypoints on S_H (introduced in Sec. 1.2). (2) We postprocess the deformed S_{temp} to remove the spikes, which come from fitting to the noisy regions in S_H. We use Laplacian-based editing operations to fix this issue and obtain our high-quality mesh S_L. See Fig. 2 in the main paper for some examples of the registered scans.

1.5 Transferring Keypoints and Region Masks

With the help of the retopologized meshes S_L, we can now transfer the keypoints and region masks defined on S_{temp} to the high-resolution scans S_H. First of all, we

[3] we only keep 118 keypoints in the facial region, which are in correspondences with the 118 keypoints defined on S_{temp}, i.e., the first set of keypoints we discussed in Sec. 1.1.

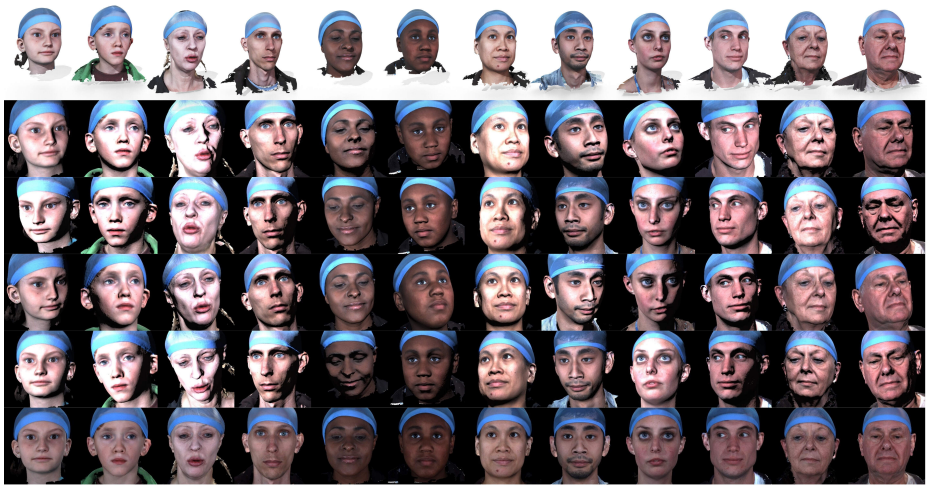

Fig. 4. Examples of our synthesized images in REALY benchmark. *First row*: aligned high-resolution scans with textures. *Second-forth rows*: multi-view images of each scan. *Fifth row*: frontal images of each scan.

can easily transfer the keypoints/region masks from S_{temp} to S_L (by vertex index) since they share the same mesh topology. Then the keypoints/region masks are transferred from S_L to S_H as follows: (1) Firstly, we traverse each point in a region $\mathcal{R}_L$ on S_L, and find its closest plane in S_H. We then collect these mapped triangles and their one-ring neighbors as the candidate corresponding region $\mathcal{R}_H$ on S_H. However, due to the significant difference in resolution between S_L and S_H, the candidate region $\mathcal{R}_H$ only contains limited and isolated vertices and triangles on S_H. (2) Secondly, we improve $\mathcal{R}_H$ by searching from the other direction, i.e., from S_H to $\mathcal{R}_L$. We find the vertices in S_H such that their nearest neighbor in S_L (in vertex-to-plane distance) lie in the region $\mathcal{R}_L$. We then include these vertices into $\mathcal{R}_H$. In this way, we get a more complete region $\mathcal{R}_H$ on S_H. Note that this step can be greatly accelerated by only considering a bounding box calculated based on step one's results instead of considering all the vertices in S_H for searching. (3) Thirdly, we filter out the vertices lie in eyeball, nostril, or mouth cavity from $\mathcal{R}_H$ since these regions might be wrongly included into $\mathcal{R}_H$ due to nearest neighbor searching. To achieve this, we construct pseudo faces in these cavity regions on the template shape and find the vertices in S_H that have nearest neighbor lying in these pseudo faces. These vertices will be excluded from $\mathcal{R}_H$. (4) We then crop a region centered at the nose tip for each scan for evaluation. Specifically, the region has a radius of $0.7 \times (d_{\text{outer\_eye}} + d_{\text{nose}})$, where $d_{\text{outer\_eye}}$ is the outer-eye-distance and d_{nose} is the distance between nose bridge and nose lower cartilage, respectively. (5) Finally, we find the maximum connected region of $\mathcal{R}_H$ and take it as the final region mask on S_H.

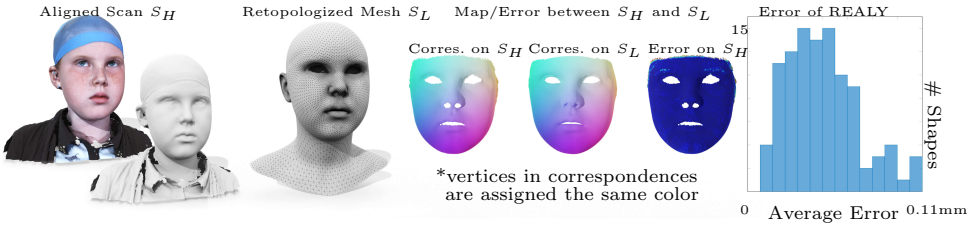

Fig. 5. Illustration of REALY quality. *Left*: samples of S_H, S_L, correspondence & error maps on S_H and S_L. *Right*: error distribution of 100 individuals in REALY, x-axis represents the error range between $0 \sim 0.11$mm, y-axis represents the shape numbers.

2 Additional Experiment and Discussion

2.1 Quality of REALY

To demonstrate the reliability of the retopologized meshes S_L, we evaluate the similarity between S_L and the corresponding scan S_H as illustrated in Fig. 5. Note that S_L and S_H are aligned and we can compute a map $T^{\text{pts}}_{h\to l}$ from S_H to S_L via nearest neighboring searing in Euclidean space. In the middle of Fig. 5 we visualize the map via color transfer. We then evaluate the shape similarity by NMSE, i.e., $e(T^{\text{pts}}_{h\to l})$. We also visualize the per-vertex error on S_H, which shows extremely small errors in the facial region. We then evaluate the shape similarity between S_H and S_L on the 100 individuals in REALY and report the average errors in the histogram in Fig. 5. Specifically, the NMSE in the facial region ranges from $0.047 \sim 0.108$ mm. The average error over all vertices in the facial region across 100 individuals is 0.070 mm. This suggests that our retopologized mesh S_L are in high-quality and guarantees the similarity to the original scans S_H.

2.2 User Study Details

We invited 70 volunteers with computer science or modeling background to conduct the user study. In every question, the user is asked to select the most similar reconstructed mesh(es) compared to the given ground-truth scan. Specifically, for each test sample, we design the following two questions:

(Q1) choose the most similar *two* (compared to the ground-truth) from nine candidate meshes reconstructed using different methods (as shown in Fig.5 in the main paper).
(Q2) choose the most similar *one* from up to three candidate meshes, which are the best reconstructions according to three different evaluation protocols (i.e., ***g*ICP** from two directions, and ours; highlighted in blue, purple, and orange boxes in the main paper).

We report the results of (Q1) in Fig. 5 in the main paper, where the best (second best) is highlighted via "$\star$" ("†"). We also show the statistics of the user study

Table 1. Detailed user study results of Fig. 5 in the main paper.

Sample	Best recon. ⋆	2nd best recon. †	Best selected by ***b*ICP**
Fig. 5 (1)	72.2% (MGCNet)	53.1% (GANFit)	85.4% (GANFit)
Fig. 5 (2)	43.9% (Deep3D)	39.0% (MGCNet)	70.7% (Deep3D)
Fig. 5 (3)	51.3% (Deep3D)	35.9% (MGCNet)	76.9% (Deep3D)

in Tab. 1. For (Q2), on average, 76.1% users agreed that our ***b*ICP** selected the best reconstructed mesh compared to ***g*ICP** evaluation protocols.

This user study shows that our ***b*ICP** indeed better aligns with human perception in measuring the similarity between the ground-truth and the reconstructed mesh. The additional results in Fig. 7 are also marked via "⋆" and "†" according to the above user study.

2.3 Comparing Different Reconstruction Methods

We visualize the error map of different methods using the standard ***g*ICP** based evaluation pipeline (evaluated on both directions between the constructed face and the original scan) or our proposed ***b*ICP** based evaluation pipeline in Fig. 6 and Fig. 7, where the errors are globally normalized across different methods, and blue (red) represents smaller (larger) error. Note that, the ***g*ICP** based errors of $e(T^{\text{pts}}_{p\to h})$ are computed on the reconstructed faces S_P (see the *third* row of each sample in Fig. 6 and Fig. 7), while our ***b*ICP** based errors are computed on the four fine-grained regions on the high-resolution scans S_H (see the *fourth* row of each sample in Fig. 6 and Fig. 7).

In some cases, the global-wise error map may exhibit misleading results mainly due to inaccurate alignment between S_P and S_H, which makes it hard to identify the best predicted face from different methods. As a comparison, our region-aware pipeline is more fair by making a comparison based on the errors defined on the same mesh S_H among different methods. Indeed, the best predicted face selected by ***b*ICP** is visually more similar to the input scan compared to the face selected by ***g*ICP** when there is a disagreement. At the same time, our ***b*ICP** can suggest which method performs the best in a particular region (see Fig. 5 and Tab. 4 in the main paper).

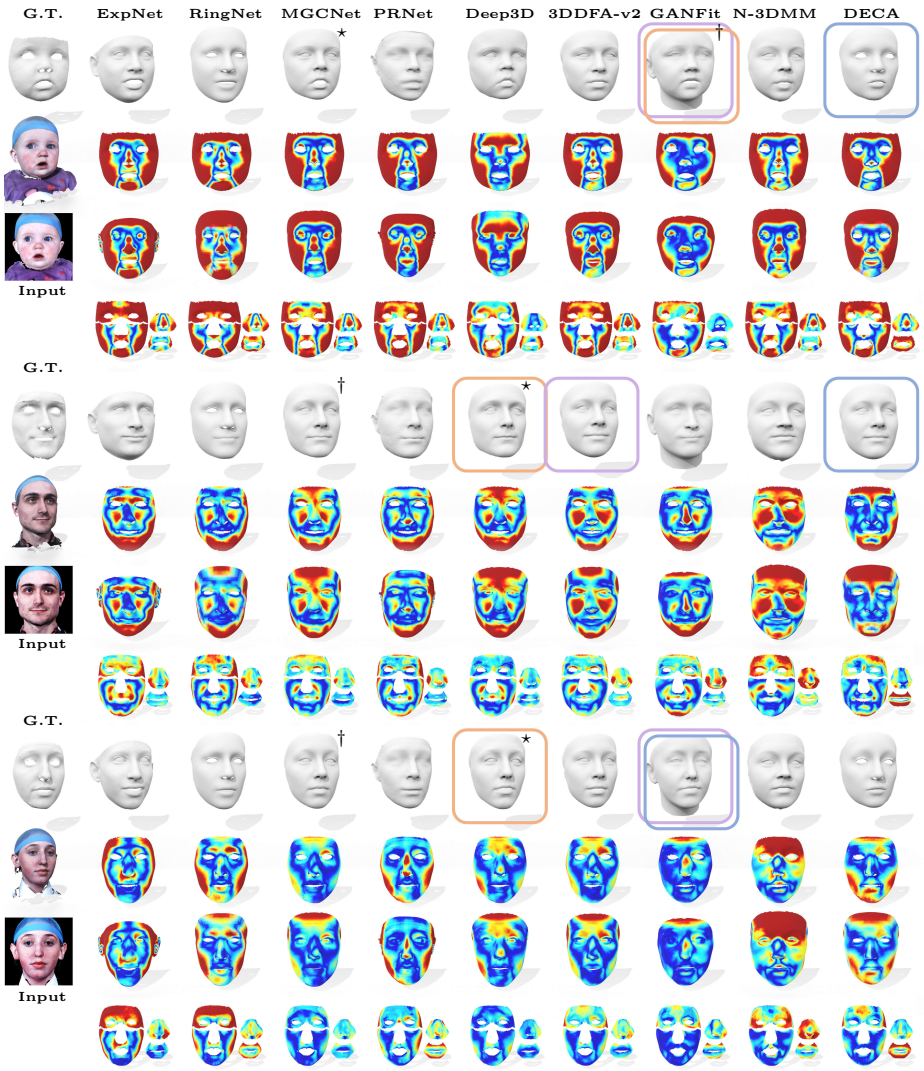

Fig. 6. Comparing different face reconstruction methods (part 1). We visualize the reconstruction error of each reconstructed face using the standard evaluation pipeline (***g*ICP**) and our novel evaluation pipeline (***b*ICP**, shown in four regions), where large (small) errors are colored in red (blue). The best reconstructed face selected using our measurement (in orange boxes) are visually closer to the ground-truth meshes than the ones selected using the standard measurements (blue boxes for $e(T^{\text{pts}}_{p\to h})$ & purple boxes for $e(T^{\text{pts}}_{h\to p})$). We also mark the best (second best) reconstructed face voted in our user study by $\star$ (†). The first row of each sample is the reconstructed shape, the second/third/fourth row of each sample is the error map of $e(T^{\text{pts}}_{h\to p}))/e(T^{\text{pts}}_{p\to h})$/ours. These three samples are illustrated in main paper and we show bigger versions for easier comparisons.

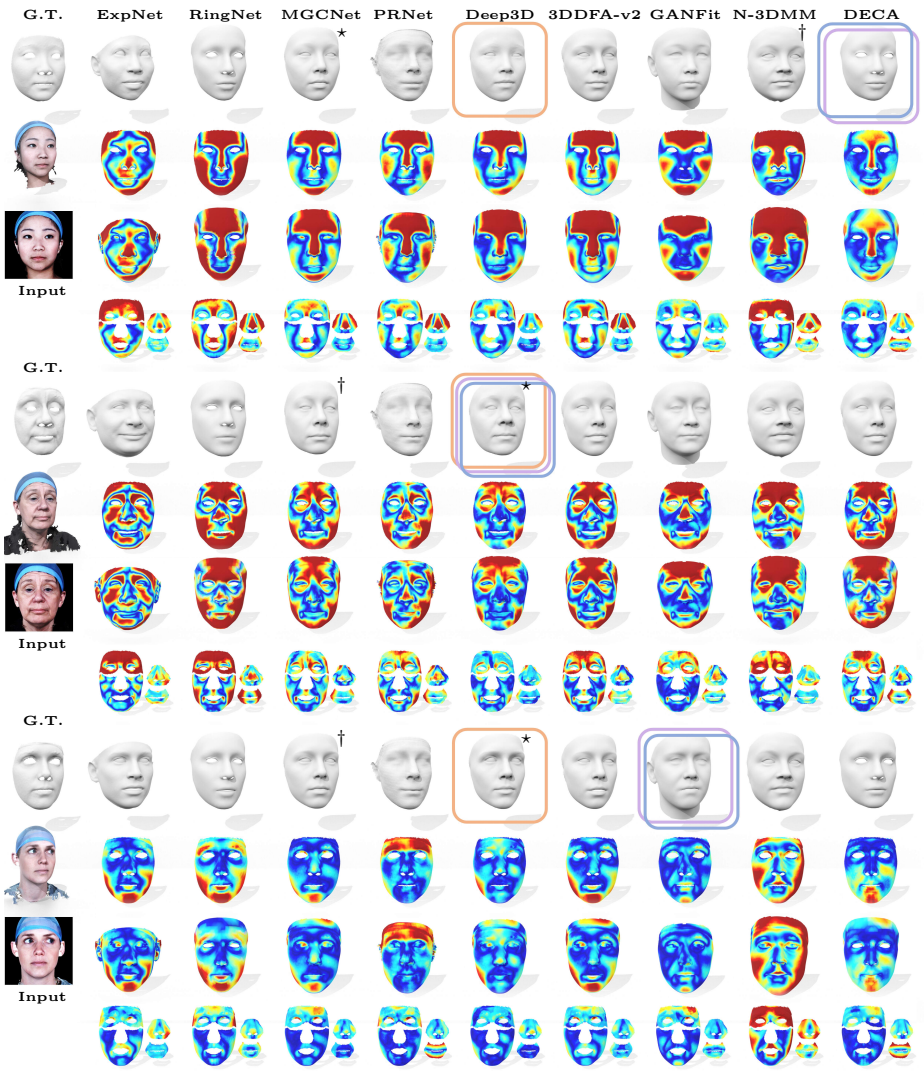

Fig. 7. Comparing different face reconstruction methods (part 2). We visualize the reconstruction error of each reconstructed face using the standard evaluation pipeline (***g*ICP**) and our novel evaluation pipeline (***b*ICP**, shown in four regions), where large (small) errors are colored in red (blue). The best reconstructed face selected using our measurement (in orange boxes) are visually closer to the ground-truth meshes than the ones selected using the standard measurement (blue boxes for $e(T^{\text{pts}}_{p\to h})$ & purple boxes for $e(T^{\text{pts}}_{h\to p})$). We also mark the best (second best) reconstructed face voted in our user study by $\star$ (†). The first row of each sample is the reconstructed shape, the second/third/fourth row of each sample is the error map of $e(T^{\text{pts}}_{h\to p}))/e(T^{\text{pts}}_{p\to h})$/ours.

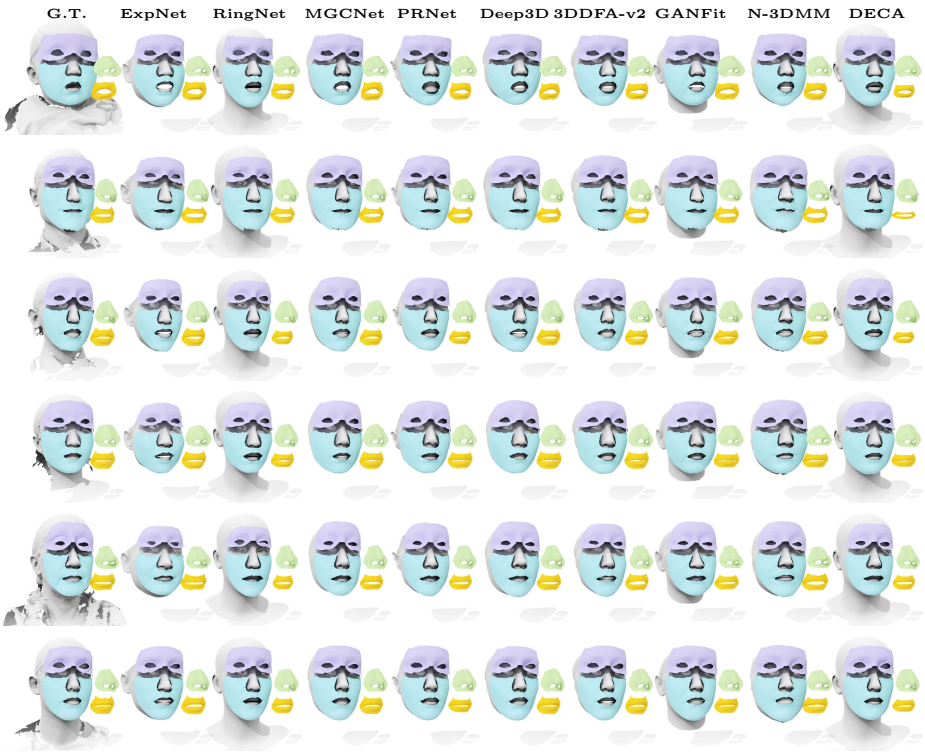

Fig. 8. Examples of the deformed regions $\mathcal{R}_H^*$ of each method. We illustrate the deformed regions $\mathcal{R}_H^*$, i.e., the intermediate results obtained after applying ***n*ICP** to deform the G.T. region $\mathcal{R}_H$ (the *first* column) to fit S_P^* in our evaluation pipeline.

As explained in Sec. 6 in the main paper, our fine-grained region-wise alignment and the two-step coarse-to-fine registration effectively helps ***n*ICP** to converge to a reasonably deformed shape $\mathcal{R}_H^*$. See Fig. 8 for such examples, where we visualize the deformed regions on top of the reconstructed faces.

2.4 Comparing Different 3DMMs

Model Variations of Different 3DMMs Fig. 9 shows the shape variations of different 3DMMs. As we discussed in the main paper, previous 3DMMs have limited shape variations because of the imbalanced ethnic scans. In contrast, HIFI3D^{++} is capable of expressing individuals in different ethnic, gender, and age groups with better generalization for downstream face reconstruction tasks.

BFM & FLAME on RGB Fitting As shown in Fig. 10, the reconstructed faces from some 3DMMs (especially BFM [16] and FLAME [14]) on RGB Fitting from a single image can be misshapen for the following reasons: (1) It has been acknowledged that 3D reconstruction from a 2D image is a severely

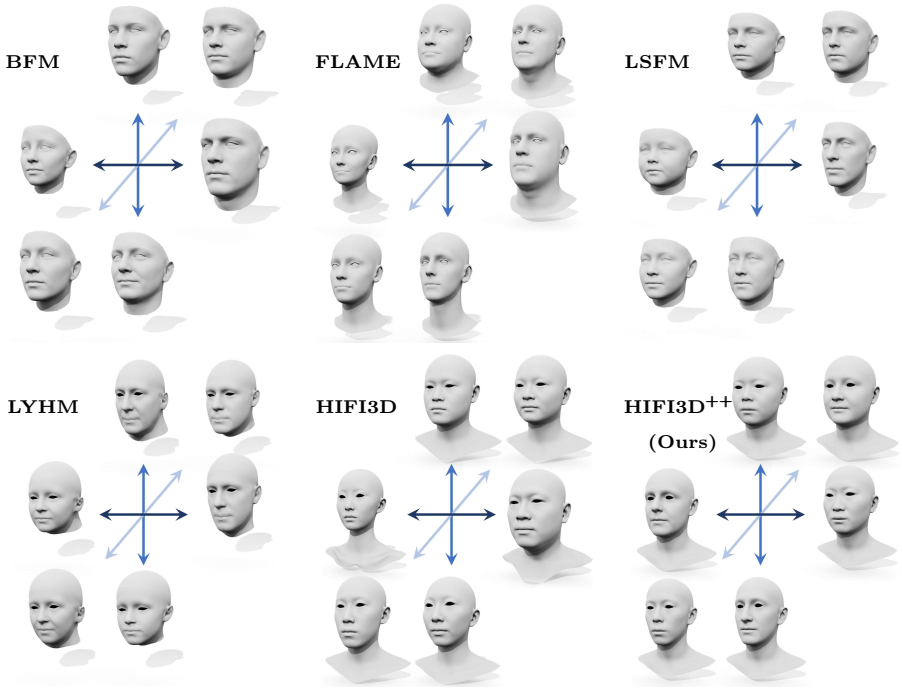

Fig. 9. Model variations of different 3DMMs. We show the shape (geometry) variation of BFM [16] (*Top left*), FLAME [14] (*Top middle*), LSFM [3] (*Top right*), LYHM [6] (*Bottom left*), HIFI3D [2] (*Bottom middle*), the proposed HIFI3D^{++} (*Bottom right*). For shape variations, the first three principal components are visualized at ± 3 standard deviations.

ill-posed. In an under-constrained setting (such as without depth information), the reconstruction quality can be poor due to the limited expressiveness of the 3DMMs (such as the FLAME results shown in Fig. 10). (2) On the other hand, the quality of the scans that are used for constructing 3DMMs can also affect the reconstruction quality. Take BFM as an example, the reconstructed example shown in Fig. 10 has unnatural noise. As a comparison, LSFM that adopts the same topology as BFM achieves less noisy results with higher quality when we assign small weights to regularization terms for fitting using these two 3DMMs, since LSFM is constructed from larger number scans with higher quality.

HIFI3D^{++} on RGB-D Fitting In order to avoid reconstructing misshapen faces, RGB fitting relies more on face prior (i.e., impose stronger regularization on the 3DMM parameter α) due to its access to only limited geometry supervision. In contrast, RGB-D fitting relies more on the denser and more informative depth supervision, which is more suitable to evaluate the expressiveness different 3DMMs. In Fig. 11, Fig. 12, and Fig. 13, we show face reconstruction results

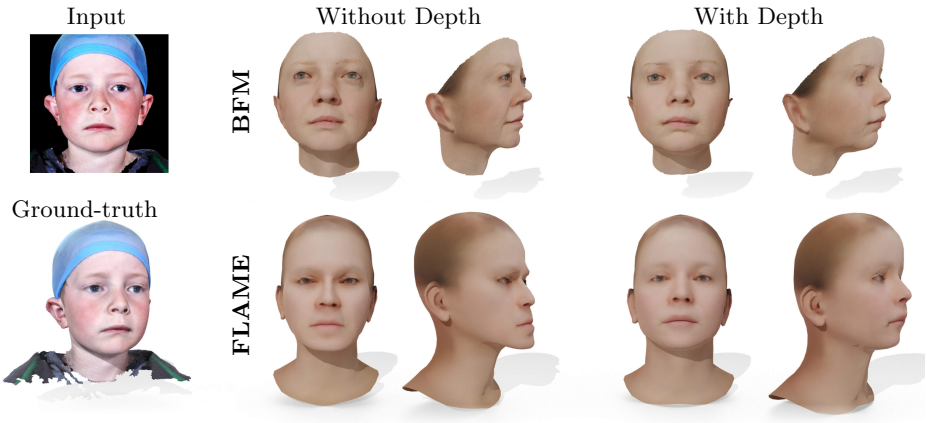

Fig. 10. Failure examples of RBG fitting and RGB-D fitting using FLAME [14] and BFM [16] 3DMMs. The reconstructed shapes may look reasonable in the frontal view but reveal their poor quality from other views. Providing depth information for fitting can significantly improve the reconstruction quality.

using different 3DMMs on RGB and RGB-D fitting. And indeed, HIFI3D^{++} outperforms the other 3DMMs on RGB-D fitting.

Conclusion Thanks to our region-based evaluation pipeline, we find that a 3DMM can have different levels of expressiveness in different regions. For example, FLAME [14] can reconstruct the nose region well, but fail to express the overall shape and the curvature in the forehead region. We believe it is promising to investigate how to construct a region-aware 3DMM in the future work, which integrates the advantages of different 3DMMs and introduces more fine-grained prior for 3D face reconstruction.

3 Implementation Details

3.1 Details of RGB(-D) Regression

For the baselines, we use the officially released 3DMMs including the shape basis and the texture basis. For our newly introduced HIFI3D^{++}, we use the same texture basis as HIFI3D and HIFI3D^{A}. Following [2], we render images with the estimated parameters (i.e., 3DMM shape & texture parameters, second-order spherical harmonics lighting parameters, and pose parameters) via a differentiable renderer [11], and adopt RGB photo loss, depth loss, identity perceptual loss, landmark loss and regularization terms to optimize these parameters, which are defined as follows.

RGB Photo Loss. The pixel-wise RGB photometric loss is defined as:

$$L_{\text{rgb}} = \left\| I_{\text{rgb}} - \hat{I}_{\text{render}} \right\|_2 \tag{1}$$

Fig. 11. Comparing different 3DMMs with RGB(-D) fitting (part 1). We highlight the best (second best) reconstructed face via red (blue) underline chosen by the proposed evaluation pipeline, and HIFI3D^{++} shows generally the most realistic face than others, quantitatively and perceptually. From left to right, LYHM [6], BFM [16], FLAME [14], LSFM [3], FaceScape [21], HIFI3D [2], HIFI3D^A [2], and the proposed HIFI3D^{++} are compared. The *first* (*second*) row of each sample shows the results of RGB (RGB-D) fitting.

where I_{rgb} is the input RGB image, $\hat{I}_{\text{render}}$ is the rendered RGB image from the differentiable renderer using the predicted parameters. We adopt $l_{2,1}$-norm for its demonstrated robustness against outliers than l_2-norm [2].

Depth Loss. The depth loss is defined as:

$$L_{\text{dep}} = \rho\Big(\big\|I_{\text{dep}} - \hat{I}_z\big\|_2^2\Big) \tag{2}$$

where $\rho(\cdot)$ defines a truncated l_2-norm that clips the per-pixel MSE, I_{dep} is the input depth image, $\hat{I}_z$ is the rendered depth image from the differentiable renderer. The truncated function is proved to be more robust to depth outliers [2].

Fig. 12. Comparing different 3DMMs with RGB(-D) fitting (part 2). We highlight the best (second best) reconstructed face via red (blue) underline chosen by the proposed evaluation pipeline, and HIFI3D^{++} shows generally the most realistic face than others, quantitatively and perceptually. From left to right, LYHM [6], BFM [16], FLAME [14], LSFM [3], FaceScape [21], HIFI3D [2], HIFI3D^{A} [2], and the proposed HIFI3D^{++} are compared. The *first* (*second*) row of each sample shows the results of RGB (RGB-D) fitting.

Identity Perceptual Loss. To capture high-level identity information, we apply the following identity perceptual loss:

$$L_{\text{id}} = \left\| \psi\left(I_{\text{rgb}}\right) - \psi\left(\hat{I}_{\text{render}}\right) \right\|_2^2 \tag{3}$$

where $\psi(\cdot)$ is the high-level identity features exacted from a pretrained face recognition model. In our experiments, we use features from the $fc7$ layer of VGGFace model [15].

Landmark loss. To achieve better fitting quality, we ask professional artist to extend the 68 keypoints defined on each 3DMM into 86[4] keypoints. Land-

[4] Corresponding to the third set of 86 keypoints we discussed in Sec. 1.1

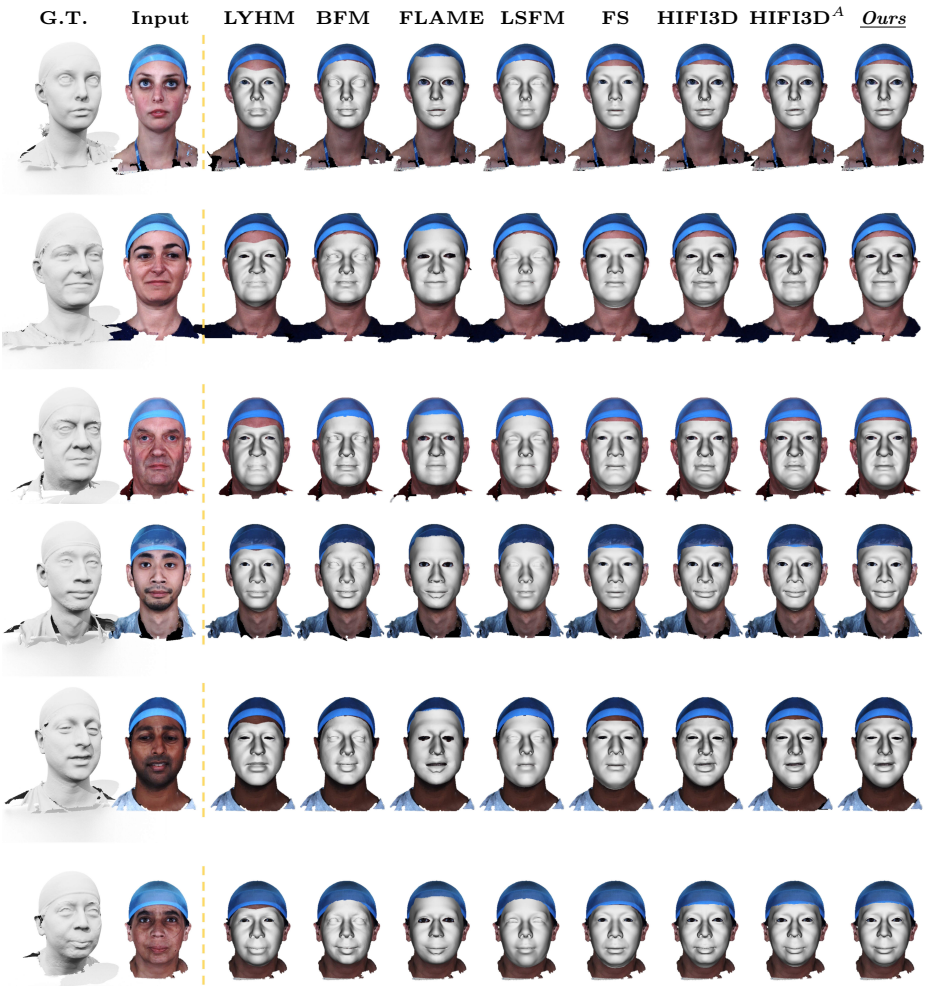

Fig. 13. Comparing different 3DMMs with RGB-D fitting (part 3). We visualize rendered shapes on input image via predicted camera parmeters. HIFI3D^{++} shows best visualized results with high fidelity features among other 3DMMs. From left to right, LYHM [6], BFM [16], FLAME [14], LSFM [3], FaceScape [21], HIFI3D [2], HIFI3D^A [2], and the proposed HIFI3D^{++} are compared.

mark loss is defined as the weighted average distances between the detected 2D landmarks and the projected landmarks from the predicted 3D model.

$$L_{\text{lmk}} = \frac{1}{\mathcal{F}} \sum_{f_j \in \mathcal{F}} \omega_j \left\| f_j - \Pi(\Phi(v_j)) \right\|_2^2 \tag{4}$$

where $f_j \in \mathcal{F}$ are the detected landmarks, $\Pi(\cdot)$ represents world-to-image plane projection with given camera parameters, $\Phi(\cdot)$ represents 6DoF head pose that rigidly rotates and translates the mesh, and v_j represent keypoints on the mesh.

Table 2. Loss parameters in Eq. (6) for RGB fitting.

3DMMs	ω_{rgb}	ω_{dep}	ω_{id}	ω_{lmk}	ω_{shp}	ω_{tex}
BFM	1000.0	0.0	4.00	5.0	5.0	1.0
FLAME	1000.0	0.0	4.00	5.0	2.0	10.0
LSFM	1000.0	0.0	4.00	5.0	2.0	1.0
FS	1000.0	0.0	0.10	5.0	50000.0*	0.0
HIFI3D	1000.0	0.0	4.00	5.0	2.0	1.0
HIFI3D^A	1000.0	0.0	4.00	5.0	2.0	1.0
LYHM	1000.0	0.0	4.00	10.0	100.0	0.01
Ours	1000.0	0.0	4.00	5.0	2.0	1.0

Table 3. Loss parameters in Eq. (6) for RGB-D fitting.

3DMMs	ω_{rgb}	ω_{dep}	ω_{id}	ω_{lmk}	ω_{shp}	ω_{tex}
BFM	1000.0	1000.0	1.00	5.0	5.0	1.0
FLAME	1000.0	1000.0	1.00	5.0	2.0	2.0
LSFM	1000.0	1000.0	1.00	5.0	2.0	1.0
FS	1000.0	1000.0	0.10	5.0	50000.0*	0.0
HIFI3D	1000.0	1000.0	1.00	5.0	2.0	1.0
HIFI3D^A	1000.0	1000.0	1.00	5.0	2.0	1.0
LYHM	1000.0	1000.0	1.00	10.0	20.0	0.2
Ours	1000.0	1000.0	1.00	5.0	2.0	1.0

*FaceScape does not converge with small ω_{shp} weight.

The weight ω_j controls the importance of each keypoint, where we set 50 for those located in eye, nose and mouth region, and 1 otherwise.

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

$$L_{\text{reg}} = \omega_{\text{shp}} \left\| \alpha_{\text{shp}} - \alpha_{\text{shp}}^{\mu} \right\|_2^2 + \omega_{\text{tex}} \left\| \alpha_{\text{tex}} - \alpha_{\text{tex}}^{\mu} \right\|_2^2 \tag{5}$$

where α_{shp} / α_{tex} and $\alpha_{\text{shp}}^{\mu}$/$\alpha_{\text{tex}}^{\mu}$ represent the predict shape/texture parameters and mean face shape/texture parameters respectively.

The final total loss function to be minimized is defined as the weighted sum of each part:

$$L_{\text{total}} = \omega_{\text{rgb}} L_{\text{rgb}} + \omega_{\text{dep}} L_{\text{dep}} + \omega_{\text{id}} L_{\text{id}} + \omega_{\text{lmk}} L_{\text{lmk}} + L_{\text{reg}} \tag{6}$$

Parameters In our experiments, we use Adam optimizer [13] in TensorFlow for 1000 iterations with a learning rate 0.05 decaying exponentially in every 50 iterations. We fine-tune the parameters $\omega_{\text{rgb}}, \omega_{\text{dep}}, \omega_{\text{id}}, \omega_{\text{lmk}}, \omega_{\text{shp}}, \omega_{tex}$ for each 3DMM and use the best combinations as reported in Tab. 2 and 3.

3.2 Parameters for the Evaluation Pipeline

We implement our evaluation pipeline in Python. It takes about 3 hours to evaluate 100 individuals in REALY on a single Intel i7-9700 CPU for each method, including both the global wise error and 4 region-aware error. Taking BFM with 35709 vertex (Deep3D) as an example, the baseline ***g*ICP** takes 0.81 minute while ours takes 1.73 minute per sample for evaluation. Specifically, for our evaluation pipeline, it takes 0.48 second for keypoint alignment, 1.28 minute for region alignment (***r*ICP**), and 0.45 minute for error computation (***b*ICP**).

In ***r*ICP**, we set the maximum number of iterations to 100 and stop early if the change of matching error is less than 10^{-6}. The weight $w_{\mathcal{K}}$ (in Algo. 1 in the main text) is set to the ratio between the number of vertex in $\mathcal{R}_H$ and in $\mathcal{K}_P$. We use a two-stage ***n*ICP** approach (Algo. 2, step 3 in the main text)

to avoid unsatisfactory deformation results in ***b*ICP**. At the first stage, we only include the landmark term and the stiffness term for initial deformation with weight 50 and 150 respectively. At the second stage we include all the terms including distance term, landmark term, and stiffness term with weight 1, 5, 50 respectively. In both stages, the weight of the stiffness term gradually decays to allow for more localized deformations.

3.3 Experiments Setup

Preprocess Input Image. For learning-based face reconstruction, we apply MTCNN [22] to detect and crop the face region of the input frontal images provided by our REALY, and resize them into a resolution of 300×300. As for RGB(-D) fitting, we resize input images to 512×512 without cropping.

Baselines for Face Reconstruction We use the officially pretrained model released by the previous work [5,17,18,9,7,12,10,20,8] for the face reconstruction experiments. We make sure that none of these methods have seen or fine-tuned on the tested images in REALY. We compare different methods by evaluating the similarity between the reconstructed face and the ground-truth from REALY using our evaluation pipeline. Note that only the facial region of each method is considered during the evaluation phase. Moreover, we also reorder the detected 68 keypoints of each method such that they are in correspondences with the 68 keypoints on the ground-truth for proper evaluation. Recall that our evaluation pipeline is based on predefined region masks where the eyeballs, nostrils, and mouth cavity are not considered for more fair comparisons since some reconstructed faces (e.g., FLAME) do not have eyeballs or nostrils or mouth cavity.

Error in Metric Units During the evaluating, we rescale the S_H in REALY and the predicted shapes S_P back to its original size in LYHM [6] such that the shape difference between S_H and S_P is measured in proper metric units and reflects real-world difference.

4 Limitation & Future Work

Our work still has some limitations. Although using in-the-lab images produced with controlled configurations can faithfully reflect the reconstruction ability of existing methods, the robustness of different methods is not investigated. We leave it for future work by generating more challenging images with different variations (e.g., expressions, backgrounds and occlusions) and extending our REALY benchmark to in-the-wild environment. Besides, our evaluation pipeline is computationally expensive since it requires alignment and deformation of each of the 4 regions. In the future, we would like to investigate more powerful and more efficient evaluation approaches.

References

1. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid ICP algorithms for surface registration. In: CVPR (2007)
2. 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)
3. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: CVPR (2016)
4. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. TVCG (2014)
5. Chang, F., Tran, A.T., Hassner, T., Masi, I., Nevatia, R., Medioni, G.G.: Expnet: Landmark-free, deep, 3d facial expressions. In: FG (2018)
6. Dai, H., Pears, N.E., Smith, W.A.P., Duncan, C.: Statistical modeling of craniofacial shape and texture. IJCV (2020)
7. 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)
8. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. SIGGRAPH (2021)
9. 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)
10. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: GANFIT: generative adversarial network fitting for high fidelity 3d face reconstruction. In: CVPR (2019)
11. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: CVPR (2018)
12. 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)
13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. TOG (2017)
15. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC (2015)
16. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: AVSS (2009)
17. 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)
18. 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)
19. 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)
20. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: CVPR (2018)
21. 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)
22. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multi-task cascaded convolutional networks. CoRR, abs/1604.02878 (2016)

23. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: A 3d total solution. TPAMI (2019)
24. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR (2012)