Meta-evaluation for 3D Face Reconstruction Via Synthetic Data

Evangelos Sariyanidi*[1], Claudio Ferrari*[2], Stefano Berretti[3], Robert T. Schultz[1,4] and Birkan Tunc[1,4]
[1] Children's Hospital of Philadelphia [2] University of Parma
[3] University of Florence [4] University of Pennsylvania
{sariyanide,schultzrt,tuncb}@chop.edu, claudio.ferrari2@unipr.it, stefano.berretti@unifi.it

Abstract

*The standard benchmark metric for 3D face reconstruction is the geometric error between reconstructed meshes and the ground truth. Nearly all recent reconstruction methods are validated on real ground truth scans, in which case one needs to establish point correspondence prior to error computation, which is typically done with the Chamfer (*i.e.*, nearest neighbor) criterion. However, a simple yet fundamental question have not been asked: Is the Chamfer error an appropriate and fair benchmark metric for 3D face reconstruction? More generally, how can we determine which error estimator is a better benchmark metric? We present a meta-evaluation framework that uses synthetic data to evaluate the quality of a geometric error estimator as a benchmark metric for face reconstruction. Further, we use this framework to experimentally compare four geometric error estimators. Results show that the standard approach not only severely underestimates the error, but also does so inconsistently across reconstruction methods, to the point of even altering the ranking of the compared methods. Moreover, although non-rigid ICP leads to a metric with smaller estimation bias, it could still not correctly rank all compared reconstruction methods, and is significantly more time consuming than Chamfer. In sum, we show several issues present in the current benchmarking and propose a procedure using synthetic data to address these issues.*

1. Introduction

> "*Good evaluation requires that evaluation efforts themselves be evaluated.*"
>
> —Daniel Stufflebeam, "Meta-evaluation", 1974

Reconstructing the 3D shape of objects from 2D data is a long-standing and fundamental problem in computer vision [19, 27, 10]. As in any problem, defining proper benchmark metrics to accurately evaluate a reconstruction method is a paramount concern. However, measuring the similarity between a reconstructed 3D mesh and the corresponding ground truth is not trivial, as these two meshes typically differ in topology (*e.g.*, the number and density of the points), and one must establish *point correspondence* between the 3D objects before the evaluation of their similarity.

The Chamfer (*i.e.*, nearest neighbor) criterion is the standard approach for establishing point correspondence between a reconstructed mesh and its ground truth (Section 2), and it is the *de facto* benchmarking metric for face reconstruction methods. However, the geometric error computed through the Chamfer criterion must be treated as a mere estimation of the true error, since the true point correspondences are unknown when one uses real data. This raises critical questions, which proved surprisingly elusive: how can one determine which geometric error metric is more appropriate for benchmarking a face reconstruction method? Does the default (*i.e.*, Chamfer) metric allow for a fair comparison, or is it likely to under or overestimate error at different rates for different reconstruction methods? In sum, no study to our knowledge evaluated the evaluation metric itself, yet this becomes increasingly more urgent as the performance gap between reconstruction methods closes. That is, we must be able to measure small differences accurately if we are to assess the state of the art correctly.

In this paper, using synthetic data, we study the ability of geometric estimators to be used as benchmark metrics for face reconstruction. First, we highlight the pressing need for such a meta-evaluation by showing that the performance ranking of reconstruction methods can change significantly depending on how point correspondence is established (Section 4.1), even if one uses the same reference points for rigid pre-alignment. Next, we define a meta-evaluation framework that is based on using synthetic data to compare the true vs. estimated geometric error. We outline evaluation criteria in this context and define meta-metrics to quantify the degree to which these criteria are satisfied. Finally, we conduct experiments and comprehensively evaluate four existing geometric error estimators.

Our results uncover important limitations of current met-

[*] Equal contribution

rics and highlight a number of future directions. First, we show that the Chamfer estimator significantly underestimates the true error, therefore is limited in its ability to measure the metrical (*i.e.*, absolute) error of a reconstruction method, which is a recently addressed problem [50]. Second, Chamfer is limited in its ability to discern small *relative* error differences between reconstruction methods, as it may squash the performance differences or even alter the true performance ranking of the compared methods. Third, we show that, while the more time-consuming non-rigid ICP (NICP) approach can significantly reduce the estimation bias, it provides little, if any, improvement in terms of estimation variance. Fourth, NICP leads to better but not perfect ranking of compared methods. As such, the estimation of geometric error in a fair and computationally efficient manner emerges as an open problem, and one that is critical for accurately assessing the state of the art.

The contributions of this paper are as follows. First, to our knowledge, we present the first meta-evaluation framework to measure the accuracy of a geometric error estimator and asses its suitability to serve as a metric for face reconstruction. Second, we conduct comprehensive experiments on four geometric estimators in terms of the criteria put forth by our framework, namely estimation bias and variance, and dependence on reconstruction method (Section 4.3). Third, we expose how the compared metrics behave on real data with different characteristics, and which metrics behave more consistently across datasets. The code is made publicly available.[1]

2. Related Work

The standard benchmark metric to measure the geometric error in 3D face reconstruction is the Chamfer Distance (CD), which is computed by summing the squared distances between nearest neighbor correspondences of two point clouds. The major problem with this metric is that the point correspondences should be *semantically* consistent, *i.e.*, nosetips or eyes corners should be matched, which cannot be guaranteed by a nearest neighbor criterion. The standard practice is to compute CD after applying ICP or a keypoint-based rigid alignment [4, 12, 18, 31, 35, 38, 40, 41, 44, 45, 48, 50, 13], which is assumed sufficient to provide meaningful semantic correspondences. However, the suitability of this metric for comparing different reconstruction methods is questionable due to some severe limitations that have been observed in the literature. In [1], a phenomenon defined as "Chamfer blindness" was observed; it states that in case of overly high density of points, CD can lose discriminative ability as one of its two summands becomes significantly smaller than the other given the presence of a few sparsely placed points in the non populated locations. Under a different perspective, Nguyen *et al.* [28] demonstrated that, given that CD only cares about the nearest neighbour of a point rather than their local distribution, as long as the supports of the compared meshes are close, then the corresponding CD will be small and less informative, while their corresponding distributions might be different. Both works thus point to the conclusion that CD might underestimate the true error. To account for this issue, several attempts have been made to go beyond the nearest-neighbor criterion, as for example with the point-to-plane [49], or normal-ray scheme [23] variants. Despite being more accurate in some circumstances, such as in case of high curvature [23], none of the above really address the problem of semantic correspondence, which is crucial for a meaningful error estimation. Other distances were investigated to measure the geometric similarity of shapes irrespective of mesh parameterization or rotations to eliminate the need for establishing dense point correspondence, such as the varifold metric [21, 32]. However, a kernel needs to be defined to induce the metric on the shapes, which might have varying properties. Also, their robustness to real data has not been yet convincingly explored. A workaround is to extend the rigid alignment step to a non-rigid one to improve point correspondence. The most popular technique in this field is the NICP [2], which is often used to densely align raw scans to a known template in order to build a 3D morphable model (3DMM) [5]. Being itself based on a nearest-neighbor objective though, it was recently shown falling short in providing accurate correspondence even for semantically crucial points (*i.e.*, facial landmarks) [14]. Several alternatives for non-rigid surface registration have been proposed [46, 42, 14], yet they are either based on learning, or rely on specific 3DMMs. This can limit their use as a benchmark metric, as being based on specific learned models limits fair comparability. Also, techniques including several hyper-parameters or measures that are learned via neural networks [42] may behave unpredictably and inconsistently.

In order to overcome the above, Chai *et al.* [7] recently proposed an alternative benchmark, named REALY. They collected a large set of high quality 3D faces in a constrained setting, which are then all processed to be in consistent topology. To improve the point correspondence estimation between their scans and reconstructed faces, they also provided region annotations to compute per-region alignment and error. This serves also to mitigate the bias in different reconstruction methods, which tend to be more accurate on specific face areas. Whereas REALY represents a significant step toward a fairer benchmark, the evaluation scheme still relies on NICP for alignment, which is extremely time consuming and cannot guarantee accurate correspondence as discussed above. Furthermore, its applicability is limited to the released dataset only. Finally, while Chai *et al.* [7] argue that the problem with the nearest-

[1] https://github.com/sariyanidi/meta-evaluation

neighbor based error is sensitivity to reference points (for rigid alignment), we show that this criterion is problematic even rigid alignment is performed with a fixed set of points.

The literature discussed above evidences the raising awareness about the issues of the current benchmark, yet a systematic study on its limitations and the extent to which it can compromise a correct and fair comparison across reconstruction methods is missing. In this paper, we try to fill this gap by providing robust evidence that estimating the geometric reconstruction accuracy is an open problem, and that the current benchmark is not reliable to this aim. In response, we propose a meta-evaluation framework to evaluate geometric error estimators via synthetic data. The usefulness of synthetic data in the 3D face domain is not an innovative finding, and many previous works successfully employed synthetic 3D faces to, for example, improve the robustness of recognition [24, 25], reconstruction [34], tracking [26] or even generative methods [29]. However, to our knowledge, no study investigated to what extent synthetic data can serve for benchmarking purposes, or used it to systematically evaluate the ability of a benchmark metric to fairly compare reconstruction methods.

3. Background and Notation

The problem addressed in this paper is measurement of the geometric error between a reconstructed 3D face mesh and the corresponding ground truth. Suppose that the reconstructed mesh R contains N points represented as an $N \times 3$ matrix $R = (r_1^T, r_2^T, \dots, r_N^T)^T$, and that the ground truth G contains M points as $G = (g_1^T, g_2^T, \dots, g_M^T)^T$. The meshes R and G are assumed to be rigidly aligned to each other.

Typically, the points in R arise from a specific mesh topology (*e.g*., the topology of a 3DMM). Therefore, the index i of a point r_i determines the location of the point relative to other points in R, and one knows in advance the indices of the points that correspond to each facial feature (*e.g*., eyes, nose). However, this is not the case for the ground truth mesh G, when it is obtained with a 3D scanner. In this case, the indices of the points g_i provide no information about the facial feature or region to which the point belongs and, in general, $M \neq N$. Therefore, one needs to establish the *point correspondence* between R and G before measuring the error. This requires, for each point r_i, to find the index c_i of the corresponding point g_{c_i} on G.

3.1. The Chamfer approach

The standard solution to establish point correspondence for a point r_i is using the *Chamfer* approach; that is:

$$c_i = \arg\min_j ||r_i - g_j||, \tag{1}$$

where $|| \cdot ||$ is the Euclidean distance. However, the Chamfer approach underestimates the geometric error and tends to establish semantically incorrect point correspondences as illustrated in Figure 1a with an example, where the nose tip as well as other points are mismatched. The underestimation problem can also be shown formally. To this end, let us assume that for any point r_i there is a distinct corresponding point g_i' on G. Then, since Chamfer correspondence is the nearest neighbor of r_i on G, we have by definition:

$$\min_j ||r_i - g_j|| = ||r_i - g_{c_i}|| \leq ||r_i - g_i'||. \tag{2}$$

As shown in Figure 2, error underestimation by Chamfer may not be benign. In the experiments, we show that this is not an exception, and the Chamfer approach significantly underestimates the real error.

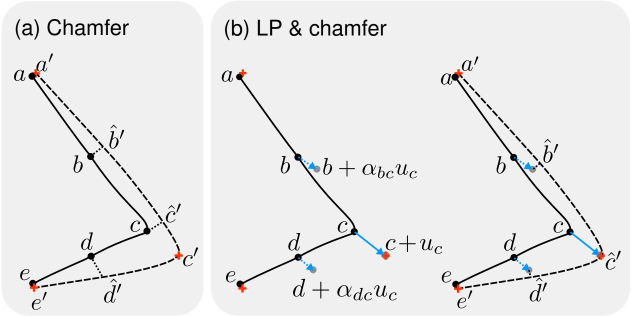

Figure 1. Establishing point correspondence between the ground truth (dashed curve) and the reconstructed mesh (solid curve) with (a) the Chamfer and (b) the LP & Chamfer approach. We focus on the part of the meshes and ground truth landmarks (red crosses a', c' and e') on the nose. The estimated nose tip with the Chamfer method, $\hat{c}'$, is far from the true nose tip c'. Also, the point b is approximately in the middle of the nose root a and tip c, but its counterpart $\hat{b}'$ estimated with the Chamfer method is closer to a' than c' (a similar observation holds for d and $\hat{d}'$). LP non-rigidly aligns the meshes by “pulling” the nose tip c towards c' with a force of u_c and dragging all the points on the mesh. Thus, LP & Chamfer finds better correspondences $\hat{b}'$ and $\hat{d}'$ as well as $\hat{c}'$.

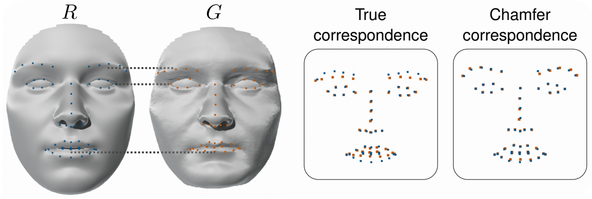

Figure 2. Underestimation of geometric error by Chamfer-based point correspondence on a real example. The reconstructed mesh R fails to correctly capture facial features like brows. While true correspondence shown for landmarks highlights this inconsistency, Chamfer correspondence misleadingly suggests that the landmarks are captured very well, leading to underestimated error.

3.2. Non-rigid ICP

NICP is an iterative approach that non-rigidly aligns the reconstructed mesh R to the ground truth G, until a conver-

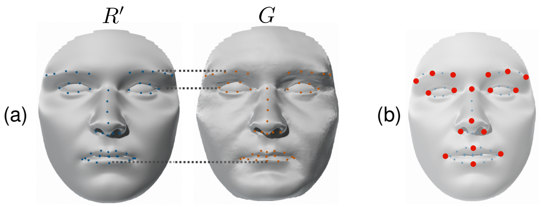

Figure 3. (a) Feature-level registration of the mesh R in Figure 2 to the G in the same figure. The registered mesh R' eliminates the misalignment of the brows and the lower lip observed in Figure 2. (b) The 18 landmarks that we use for feature-level registration.

gence criterion is met. A by-product of this process is the set of estimated point correspondences between R and G, which can be used for measuring geometric error between the (unwarped) R and G. The computational cost of NICP is significantly higher than that of Chamfer, as the former typically takes minutes on a CPU, whereas the latter takes a few seconds. Despite the added computational complexity, there is no guarantee that NICP will lead to better point correspondences than Chamfer, since the objective function to minimize is based on the latter [14]. To our knowledge, no study directly evaluated whether the NICP-based point correspondences are accurate enough to enable a fair comparison of reconstruction methods. Our evaluation framework (Section 4) and experiments address this open issue.

3.3. Landmark-based pre-alignment

If some landmark points are annotated on the ground truth mesh G, one can use them to apply a (non-rigid) landmark-based pre-alignment (LP) to the reconstructed mesh in a way that the corresponding landmarks match (Figure 3). This procedure can improve point correspondence not only for the landmarks themselves, but also for the points close to them (Figure 1b). Such strategies have been used for establishing point correspondence while constructing 3DMMs [5, 15, 8, 43] or when training 3DMM fitting methods [36]. In our study, we investigate the potential benefits of applying such a pre-alignment strategy. The LP approach that we use is outlined in Appendix A.

4. The Meta-evaluation Framework

4.1. Why Do We Need Meta-evaluation?

Table 1 compares four face reconstruction methods in terms of (estimated) geometric error on the widely used BU4DFE data [47]. Error is estimated with two different approaches in a controlled manner, as we implemented all reconstruction methods from scratch and used the same exact reconstructed meshes while computing error—the only difference between the errors reported in the two columns is the manner in which point correspondence is established.

	Estimated Error (mm)	
	Estimator 1	Estimator 2
INORig [4]	**1.72**	3.06
Deep3DFace [9]	*1.74*	*3.05*
3DI [39]	1.81	**2.80**
3DDFAv2 [18]	2.04	3.50
Mean error	1.83	3.10
Standard dev.	0.13	0.25

Table 1. Performance of four methods according to two different error estimators. Estimator 1 is the Chamfer error and Estimator 2 is LP+NICP (Section 5.1). **Bold**, *italic* and underlined text indicate first, second and third best method according to an estimator.

The results in Table 1 show a concerning outcome: The ranking of the methods depends significantly on the estimator that is used to measure error. Moreover, the range of the reported errors shows a significant difference; the mean error of compared methods is 1.83 according to one estimator, and 3.10 according to the other. Finally, the performance differences between the methods is also squashed; standard deviation of the errors estimated by the Estimator 2 is almost the double of that of Estimator 1.

In sum, it is difficult to know which reconstruction method is the best without knowing which error estimator is more reliable. Further, estimating the absolute error is also problematic: at least one estimator under or overestimates error, and we do not know which one it is. The meta-evaluation framework presented in this paper aims to fill this critical gap by presenting tools and procedures that allow us to know which estimator is more likely to predict the correct ranking or produce less estimation bias.

4.2. Comparison on Synthetic Data

The goal of our framework is to evaluate the quality of a geometric error estimator, which is essentially done by comparing the geometric error estimated by a specific method (*e.g.*, Chamfer) with the true error. However, computing the true geometric error requires the knowledge of exact point correspondence, therefore one has to use a synthetic dataset in these comparisons, as the real data collected from 3D scanners do not adhere to a pre-determined mesh topology (Section 3), and manual annotation over a dense set of points is not an option. More precisely, we synthesize data that is consistent with the mesh topology of the reconstruction method(s).

4.3. Evaluation criteria

There can be a number of criteria to evaluate an error estimator, and depending on the context and application, some criteria be more prioritized. In this paper, we focus on three evaluation criteria. The first one is *estimation bias*, which evaluates if an estimator systematically underestimates or overestimates the true error. An unbiased estimator is critical when determining whether a face reconstruction method is appropriate for an application where one, say, needs to measure the distance between facial features during online eyeglass shopping. The second criterion is *estimation variance*. A low variance is necessary when estimating error for a specific subject, or the average error of a reconstruction method on a dataset with a relatively small number of participants. The third criterion for estimated error is *dependence to reconstruction method*, which is particularly critical when one compares several 3D reconstruction methods-—performance comparison cannot be fair if error is over or underestimated at different rates for different methods.

4.4. Meta-metrics

To quantify error estimation bias and variance, we fit a linear model[2] to the true vs. estimated error values, and report the slope of the linear model and the R^2 score, as in Figure 4. The slope of the fitted line, m, determines if the estimation is biased; $m = 1$ means that the estimation is unbiased, $m < 1$ means that error is underestimated, whereas $m > 1$ means that it is overestimated. The R^2 score is (inversely) related to the variance of the estimator, as it quantifies the proximity of the estimated errors to the fitted line. It takes values in the range of $[0, 1]$—the higher the better.

To quantify the degree to which an error estimator is likely to lead to an unfair performance comparison (*i.e*., dependence to reconstruction method; see Section 4.3), we follow a similar approach based on line fitting, but we fit separate lines for the errors of different reconstruction methods, and compare whether the slopes of the lines are consistent. Suppose that we are comparing K reconstruction methods on a dataset of N subjects, and that the estimated geometric error for the ith subject with the kth method is $\hat{\varepsilon}_i^k$, whereas the true error is ε_i^k. To compute the rate of under- or over-estimation of the error for the kth method, we fit a linear model to the pairs $(\varepsilon_1^k, \hat{\varepsilon}_1^k), (\varepsilon_2^k, \hat{\varepsilon}_2^k), \dots, (\varepsilon_N^k, \hat{\varepsilon}_N^k)$, and use the slope of this model, m_k. Then, the *rate of inconsistency* (ROI) of the geometric error estimator, η, is computed as the coefficient of variation as:

$$\eta = \sigma_m / \bar{m}, \tag{3}$$

where $\bar{m}$ and σ_m are, respectively, the average and standard deviation of the values $m_1, m_2, \dots, m_K$. The lower the η the better, as this indicates that any bias in the estimation of the geometric error is consistent across reconstruction methods. On the contrary, a large ROI may lead to an incorrect ranking of the compared methods.

5. Experiments

We experimentally show the use of our meta-evaluation framework by comparing four error estimators in terms of: *(i)* prediction of point-wise errors on reconstructed meshes (Section 5.2); *(ii)* prediction of average per-subject errors and consistency across reconstruction methods (Section 5.3); and *(iii)* predicted ranking of reconstruction methods (Section 5.4).

5.1. Experimental setup

Compared geometric error estimators. We compared four geometric error estimators: *(i) Chamfer* error; *(ii) LP+Chamfer*, which is the error computed from correspondences obtained via the Chamfer criterion after LP (Section 3.3); *(iii) NICP*, which is the error computed after establishing point correspondence via NICP [2]; *(iv) LP+NICP*, which is the error computed through point correspondences based on NICP established after LP. Chamfer-based errors are computed unidirectionally.

Datasets. We conducted experiments on two synthesized datasets, as well as two real datasets (Florence [3] and BU4DFE [47]). The first synthesized dataset, referred to as *BFM-synth*, includes meshes and texture data of 100 subjects, obtained by randomly generating shape and texture coefficients using the Basel Face Model (BFM) [30]. We rendered 50 images per subject from various poses, and performed 3D reconstruction experiments using 1, 10, or 50 frames per subject. We included a second variant, the *BFM-synth-biased* dataset, to ensure that reconstructions similar to BFM mean face do not get unfair advantage. This variant is generated in a similar way to BFM-synth, except that the shape coefficients are not zero-mean—they are subject to a bias that we enforced by randomly forcing some of the generated shape coefficients to be positive. The second dataset, *FLAME-synth* was generated similarly from the FLAME model [22]. With Florence and BU4DFE datasets, we used texture and shape data to similarly render 50 images per subject. All meshes and rendered images are neutral with no added illumination variation or occlusion.

Pre-alignment. Prior to estimating 3D reconstruction error, we rigidly aligned the ground truth to the reconstructed mesh by using 5 keypoints (mouth corners, outer eye corners, nose tip). The 18 landmarks (Figure 3b) needed for the LP were computed automatically with the same strategy used by Booth *et al*. [5], which is based on rendering multiple 2D images per mesh, then detecting 2D landmarks [6], and finally re-mapping the landmarks to 3D.

Compared reconstruction methods. We implemented a total of six reconstruction methods: 3DDFAv2 [18], 3DI [39], DECA [11], Deep3DFace [9], INORig [4] and RingNet [37]. Four methods (3DDFAv2, 3DI, Deep3DFace, INORig) are based on the BFM mesh topology, and we use them to directly compare the known error on the synthesized BFM data with estimated errors. Similarly, the methods based on FLAME mesh topology are used to compare true vs. estimated error on the FLAME-

[2] We fit a model without intercept, because if true error for a point is 0, the Chamfer distance will also be 0 according to (1).

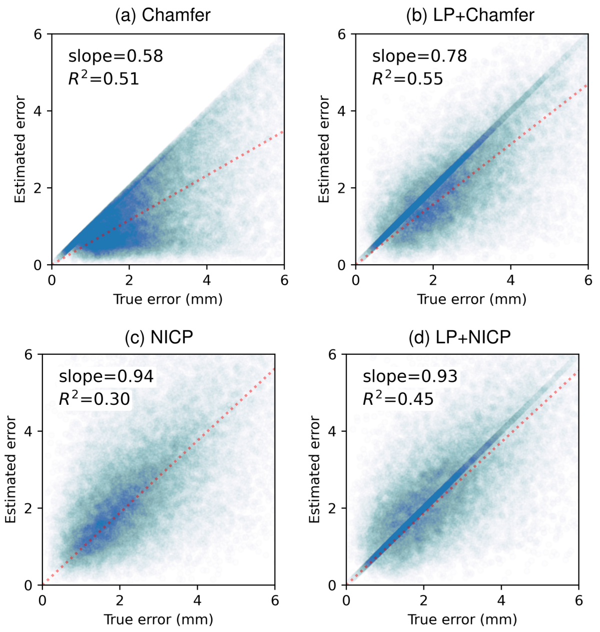

Figure 4. True vs. estimated point-wise error of 50k randomly selected points from reconstructed meshes on the BFM-synth dataset, shown for the 4 metrics in (a–d). A line (dotted) is fit to each plot and its slope and R^2 score (Section 4.4) are reported.

synth dataset. Main experiments are based on BFM-based methods, and FLAME-based methods are provided in Appendix B. For BFM-based methods, we reported results for the 23,470 mesh points that are common to the meshes produced by all BFM-based methods. For natively multi-frame methods (3DI and INORig), we performed reconstruction using F=10 and F=50 frames per subject. Single-frame methods that produce frontal and neutral meshes were extended to multi-frame by naive-averaging [17] of $F = 10, 50$ frames per subject. Unless stated otherwise, results are reported for F=50 frames.

5.2. Point-wise error prediction

We investigated how well the true geometric error of an arbitrary point on a reconstructed mesh can be predicted by the estimated error. Figure 4 compares the four benchmark metrics by plotting true vs. estimated error for points of reconstructed meshes obtained by the four BFM-based methods on the BFM-synth dataset.

Figure 4a confirms that Chamfer is only a lower bound of the true error (Section 3.1), and the slope of $m = 0.58$ (Section 4.4) suggests that the error is significantly underestimated for a large number of points. In comparison, LP+Chamfer has significantly less bias, as the slope of $m = 0.78$ is closer to 1, indicating that LP is a promising strategy. Appendix B shows that LP+Chamfer provides a similar improvement over Chamfer on the FLAME topology and data. The NICP-based approaches in Figure 4c,d have even less bias than LP+Chamfer, with slopes significantly closer to 1. However, NICP without LP has the lowest R^2 score ($R = 0.30$), indicating that a very large estimation variance is expected. The usage of LP along with NICP leads to a better estimation variance ($R = 0.45$).

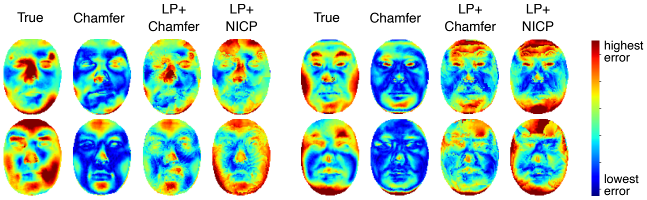

Figure 5. Point-wise error for four BFM-synth meshes visualized via heatmaps. Colors indicate amount of error for each mesh consistently across error metrics.

In sum, although the usage of NICP reduces bias significantly, none of the compared error estimators produces a notably high R^2 score, indicating that the error of an arbitrary point on a reconstructed mesh is likely to be estimated unreliably. Nevertheless, NICP-based approaches or LP+Chamfer have significantly less bias than Chamfer, since the latter largely underestimates true error, as seen also in the heatmap representation in Figure 5.

5.3. Per-subject error prediction

We also compared geometric error estimators in their ability to predict the average per-vertex error of a mesh, which is the building block of statistics that are used to compare reconstruction methods. Figure 6 plots the true vs. estimated per-subject error for all compared benchmark metrics on the BFM-synth and BFM-synth-biased datasets, along with the R^2 scores; the plotted lines depict the slopes m_k for each reconstruction method (Section 4.4). The wider these lines are spread (*i.e.*, the higher the η) for an estimator, the more likely that this estimator will be unfair to some reconstruction method(s). The R^2 score of LP+Chamfer is higher than that of Chamfer, indicating that the usage of LP improves average per-subject estimation error. However, the η value of LP+Chamfer is also higher, as the this estimator favors the 3DI method, which likely stems from the high accuracy of 3DI on landmarks (Table 2). That is, the landmark-based alignment, LP, causes relatively little warping for the 3DI method, therefore the Chamfer bias remains stronger. Overall, the lowest η is obtained by LP+NICP, therefore this estimator will likely lead to fairer comparison of reconstruction methods. The R^2 of LP+NICP is relatively low, which means that the error for a specific subject's mesh may not be estimated accurately, although the average over per-subject estimations over a dataset may be estimated more accurately given that the dataset has sufficiently many subjects.

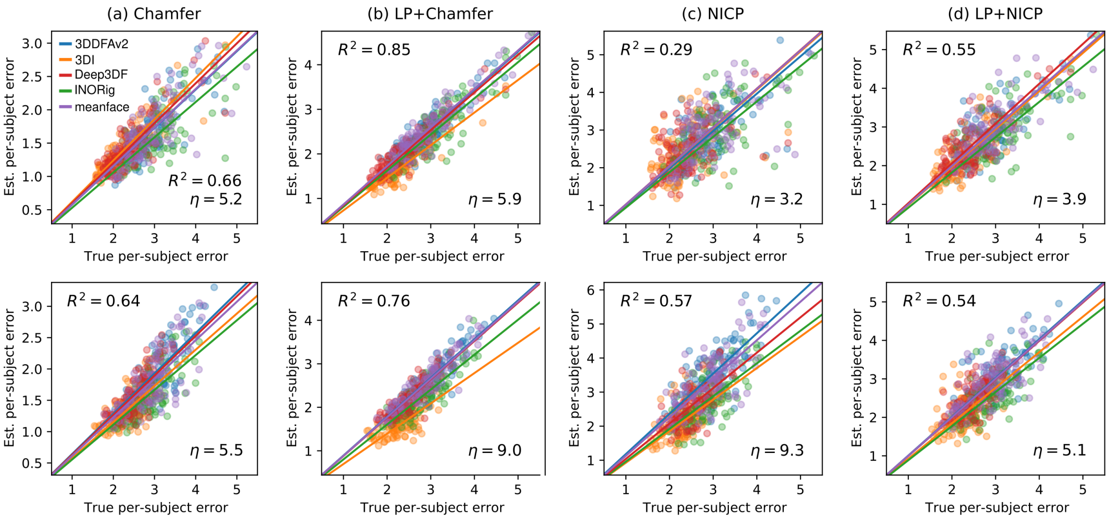

Figure 6. True vs. estimated per-subject error (in mm), shown separately for the four compared benchmark metrics; top row: results for the BFM-synth dataset: bottom row results for the BFM-synth-biased dataset. The colors of the points indicate the reconstruction method that is used, and the line (with the same color) indicates the slope of the linear model fit to the true vs. estimated errors of the same method.

Table 2. Mean landmarks error (in mm) comparison.

3DI	3DDFAv2	Deep3DFace	INORig	meanface
1.30	2.36	1.96	2.38	2.45

Table 3. True and estimated average error (in mm) of reconstruction methods on BFM-synth data. Highlighted cells indicate cases where ranking according to estimated error is inconsistent with true ranking. The subscript of reconstruction methods is the number of frames used per reconstruction. **Bold**, *italic*, and underlined text indicate the best, second and third methods, respectively.

		Estimated error			
Rec. method	True	Cham.	LP+ Cham.	NICP	LP+ NICP
3DI$_{50}$	**2.33**	*1.47*	**1.75**	**2.51**	**2.30**
Deep3DFace$_{50}$	*2.41*	**1.45**	2.04	*2.52*	2.49
Deep3DFace$_{10}$	2.42	**1.45**	2.06	2.60	2.51
3DI$_{10}$	2.47	1.54	*1.87*	2.66	*2.44*
Deep3DFace$_{1}$	2.55	1.54	2.20	2.78	2.69
3DDFAv2$_{50}$	2.85	1.64	2.45	2.86	2.85
3DDFAv2$_{1}$	2.86	1.64	2.45	2.83	2.85
Mean face$_{1}$	3.03	1.75	2.62	3.14	3.04
INORig$_{50}$	3.10	1.64	2.52	2.98	2.88
INORig$_{10}$	3.12	1.66	2.53	3.02	2.90

An interesting outcome is that the INORig method's error seems to be systematically underestimated; it has either the lowest or second lowest slope on any plot in Figure 6. The fact that error can be systematically more underestimated for a specific method is a critical issue for benchmarking, therefore we conducted further analyses to understand the reason behind it, which lead to a surprising outcome: The reconstruction method that is used has a very strong effect on a reconstructed mesh R, to the point that the meshes of two different subjects obtained by the same reconstruction method are in general more similar than the meshes of the same subject obtained by different methods (Appendix C). As such, it is not implausible that error can be under or overestimated more for a specific reconstruction method compared to others.

5.4. Predicting ranking of reconstruction methods

A fundamental use of a benchmark metric is to compare reconstruction methods. Table 3 shows the true and predicted error of nine variants of reconstruction methods on the (zero-mean) BFM-synth dataset, as well as the error obtained by simply using the mean face (of the BFM)—a practice suggested in the literature [10]. The methods are ranked based on their true error, and the highlighted cells depict the cases where the estimated errors lead to an incorrect ranking or artificially equate methods. Table 3 shows that Chamfer leads to incorrect ranking for several methods. Even if the ranking is correct, performance differences may be squashed; *e.g.*, the true error difference between Deep3DFace$_1$ and 3DDFAv2$_1$ is 0.30mm, but according to Chamfer it is 0.10mm. Overall, the NICP-based methods are the most successful in terms of recovering the correct ranking, but none of the estimators correctly rank all reconstruction methods. In particular, the INORig method seems to be consistently getting an unfair advantage.

We next evaluate how the compared geometric error estimators behave on real datasets, namely Florence and BU4DFE. Table 4 reports the (estimated) errors and ranking of the reconstruction methods on synthesized and real data. To reduce the unfair advantage that the mean face "reconstruction" can have on the synthesized data, we used the

Table 4. True and estimated average error (Chamfer or LP+NICP) on synthesized (BFM-synth-biased) and real data. Subscript F is number of frames per reconstruction. **Bold**, *italic*, and $\underline{\text{underlined}}$ text indicate the best, second and third methods, respectively.

	Synthesized data			Florence data		BU4DFE data	
Rec. Method	True	Ch.	LP+ NICP	Ch.	LP+ NICP	Ch.	LP+ NICP
3DI$_{50}$	**2.30**	**1.33**	**2.15**	**1.64**	**2.58**	1.81	**2.80**
3DI$_{10}$	*2.42*	*1.39*	*2.24*	**1.64**	*2.62*	1.85	*2.89*
Deep3DF$_{50}$	$\underline{2.49}$	$\underline{1.57}$	$\underline{2.50}$	1.83	$\underline{3.13}$	$\underline{1.74}$	$\underline{3.05}$
Deep3DF$_{10}$	2.50	1.57	2.51	1.83	3.15	*1.73*	3.06
Deep3DF$_{1}$	2.69	1.70	2.71	1.93	3.38	1.86	3.28
INORig$_{50}$	2.89	1.60	2.57	*1.77*	3.19	**1.72**	3.06
INORig$_{10}$	2.89	1.60	2.57	$\underline{1.78}$	3.20	$\underline{1.74}$	3.09
3DDFA$_{50}$	2.94	1.81	2.97	1.95	3.39	2.04	3.50
3DDFA$_{10}$	3.00	1.92	3.04	1.96	3.41	2.08	3.57
Mean face	3.04	1.95	3.07	2.07	3.56	2.19	3.77

BFM-synth-biased data in these experiments, as the mean face is the optimal constant face shape on the BFM-synth data but not on BFM-synth-biased.

A critical result in Table 4 is that the Chamfer-based ranking on real data has little consistency with the Chamfer-based ranking on synthesized data. The Chamfer-based ranking on the BU4DFE dataset is particularly inconsistent with all other rankings. This is unexpected, as we followed identical 2D image rendering procedures and used only on neutral scans/images. The Chamfer error's lack of consistency between real datasets (Florence and BU4DFE) may be due to the difference in the effective resolution of the meshes in the two respective datasets (Figure 7).

The ranking of reconstruction methods in Table 4 based on the LP+NICP metric is more consistent between the three datasets, namely the BU4DFE, Florence and BFM-synth-biased. Since we do not have access to the true error on the real datasets, we cannot reach any conclusive inference about which error estimator works better on the real data. However, it is reasonable to assume that the LP+NICP metric predicts the ranking better, as one would expect methods to be ranked similarly when all the experimental conditions (*i.e.*, image generation procedures and facial expression on meshes) are identical across datasets. Moreover, we can readily see that LP+NICP makes fewer ranking errors than the Chamfer on synthesized datasets (Table 3,4). As such, one may conjecture that the LP+NICP error predicts the ranking of compared methods better on real data.

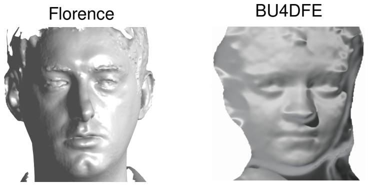

Figure 7. 3D meshes from the Florence and BU4DFE datasets. The effective resolutions of the two datasets seem to be different; the Florence mesh provides sharper and more detailed 3D data.

6. Limitations

We evaluated methods that perform reconstruction by using a 3D face template (*i.e.*, mesh topology); however, there exist approaches that do not necessarily use one [20, 16, 33]. In these cases, using synthetic data is not helpful, since the topology of the reconstruction is not known and might differ from sample to sample. We also remark that the sensitivity of a geometric error estimator to the (effective) resolution of ground truth meshes should be made a criterion for meta-evaluation, as Section 5.4 suggests that resolution may significantly alter the output of some error estimators. Thus, our meta-evaluation framework may be extended to quantify how an error estimator responds to resampling or smoothing of the ground truth meshes. Finally, while we used simple (3DMM-generated) synthetic data in this initial study, one can use synthetic 2D/3D data that is more realistic and challenging; nevertheless, the meta-evaluation procedure can be used as it is.

7. Conclusion

We presented a meta-evaluation framework to assess error estimators for benchmarking 3D face reconstruction. Specifically, we introduced a procedure to measure the bias and variance in estimated error, and the dependence of the estimator to the reconstruction method. Further, we used the framework to evaluate four approaches to geometric error estimation. Critically, results exposed limitations of the standard benchmark approach, namely the Chamfer distance, by showing that the latter largely underestimates true errors, and can also squash relative performance differences between reconstruction methods or even alter true ranking. While the usage of non-rigid ICP (NICP) with landmark-based (non-rigid) pre-alignment led to more consistent ranking and less bias, it still did not correctly rank all reconstruction methods. Moreover, the computational intensity of NICP may prove prohibitive for some applications. In sum, our study exposed the limitations in benchmarking reconstruction methods, and provided a set of tools to measure the needed progress.

Acknowledgments

This work is partially funded by the National Institutes of Health (NIH), Office of the Director (OD), National Institute of Child Health and Human Development (NICHD), and National Institute of Mental Health (NIMH) of US, under grants R01MH118327, R01MH122599, 5P50HD105354-02 and R21HD102078; and the IDDRC at CHOP/Penn.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *Int. Conf. on Machine Learning (ICML)*, pages 40–49. PMLR, 2018. 2

[2] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2, 5

[3] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2D/3D hybrid face dataset. In *Jint ACM workshop on Human Gesture and Behavior Understanding*, pages 79–80, 2011. 5

[4] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3d face reconstruction via in-network optimization. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6216–6225, June 2021. 2, 4, 5

[5] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5552, 2016. 2, 4, 5

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). 2017. 5

[7] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *European Conf. on Computer Vision (ECCV)*, pages 74–92. Springer, 2022. 2

[8] Hang Dai, Nick Pears, William AP Smith, and Christian Duncan. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE international conference on computer vision*, pages 3085–3093, 2017. 4

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 4, 5

[10] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Trans. Gr.*, 39(5):1–38, 2020. 1, 7

[11] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. on Graphics (ToG)*, 40(4):1–13, 2021. 5

[12] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *European Conf. on Computer Vision (ECCV)*, pages 534–551, Sept. 2018. 2

[13] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 780–786. IEEE, 2018. 2

[14] Claudio Ferrari, Stefano Berretti, Pietro Pala, and Alberto Del Bimbo. A sparse and locally coherent morphable face model for dense semantic correspondence across heterogeneous 3d faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 44(10):6667–6682, 2021. 2, 4

[15] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Dictionary learning based 3d morphable model construction for face recognition with varying expression and pose. In *Int. Conf. on 3D Vision*, pages 509–517. IEEE, 2015. 4

[16] Leonardo Galteri, Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Deep 3d morphable model refinement via progressive growing of conditional generative adversarial networks. *Computer Vision and Image Understanding*, 185:31–42, 2019. 8

[17] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. pages 8377–8386, 2018. 6

[18] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *European Conf. on Computer Vision (ECCV)*, 2020. 2, 4, 5

[19] Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970. 1

[20] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *IEEE Int. Conf. on Computer Vision*, pages 1031–1039, 2017. 8

[21] Irene Kaltenmark, Benjamin Charlier, and Nicolas Charon. A general framework for curve and surface comparison and registration with oriented varifolds. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3346–3355, 2017. 2

[22] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 5

[23] Feng Liu, Luan Tran, and Xiaoming Liu. 3d face modeling from diverse raw scan data. In *IEEE/CVF Int. Conf. on Computer Vision*, pages 9408–9418, 2019. 2

[24] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gerard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conf. on Computer Vision*, pages 579–596. Springer, 2016. 3

[25] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Gozde Sahin, and Gerard Medioni. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision*, 127:642–667, 2019. 3

[26] Steven McDonagh, Martin Klaudiny, Derek Bradley, Thabo Beeler, Iain Matthews, and Kenny Mitchell. Synthetic prior design for real-time face tracking. In *Int. Conf. on 3D Vision (3DV)*, pages 639–648. IEEE, 2016. 3

[27] Theo Moons, Luc Van Gool, Maarten Vergauwen, et al. 3d reconstruction from multiple images part 1: Principles. *Foundations and Trends® in Computer Graphics and Vision*, 4(4):287–404, 2010. 1

[28] Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set distances for learning representations of 3d point clouds. In *IEEE/CVF International Conf. on Computer Vision (CVPR)*, pages 10478–10487, 2021. 2

[29] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo. Generating multiple 4d expression transitions by learning face landmark trajectories. *IEEE Transactions on Affective Computing*, 2023. 3

[30] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *IEEE Conf. on Advanced Video and Signal-based Surveillance*, pages 296–301, 2009. 5

[31] Jingtan Piao, Chen Qian, and Hongsheng Li. Semi-supervised monocular 3d face reconstruction with end-to-end shape-preserved domain transfer. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Oct. 2019. 2

[32] Emery Pierson, Mohamed Daoudi, and Sylvain Arguillere. 3d shape sequence of human comparison and classification using current and varifolds. In *European Conference on Computer Vision*, pages 523–539. Springer, 2022. 2

[33] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 8

[34] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, pages 460–469. IEEE, 2016. 3

[35] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. SADRNet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Trans. on Image Processing*, 30:5793–5806, 2021. 2

[36] Augusto Salazar, Stefanie Wuhrer, Chang Shu, and Flavio Prieto. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, 25:859–879, 2014. 4

[37] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, June 2019. 5

[38] Evangelos Sariyanidi, Casey J. Zampella, Robert T. Schultz, and Birkan Tunc. Inequality-constrained and robust 3d face model fitting. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conf. on Computer Vision (ECCV)*, pages 433–449, 2020. 2

[39] Evangelos Sariyanidi, Casey J Zampella, Robert T Schultz, and Birkan Tunç. Inequality-constrained 3d morphable face model fitting. *TechRxiv*, 2022. 4, 5

[40] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conf. on Computer Vision (ECCV)*, pages 53–70, 2020. 2

[41] Xiaoguang Tu, Jian Zhao, Mei Xie, Zihang Jiang, Akshaya Balamurugan, Yao Luo, Yang Zhao, Lingxiao He, Zheng Ma, and Jiashi Feng. 3D face reconstruction from a single image assisted by 2D face images in the wild. *IEEE Trans. on Multimedia*, 23:1160–1172, 2021. 2

[42] Dahlia Urbach, Yizhak Ben-Shabat, and Michael Lindenbaum. Dpdist: Comparing point clouds using deep point cloud distance. In *European Conf. On Computer Vision (ECCV)*, pages 545–560. Springer, 2020. 2

[43] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022. 4

[44] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3d face reconstruction via conditional estimation. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 13289–13298, Oct. 2021. 2

[45] Zichun Weng, Youjun Xiang, Xianfeng Li, Juntao Liang, Wanliang Huo, and Yuli Fu. Learning semantic representations via joint 3D face reconstruction and facial attribute estimation. In *Int. Conf. on Pattern Recognition (ICPR)*, pages 9696–9702, 2021. 2

[46] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Balanced chamfer distance as a comprehensive metric for point cloud completion. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:29088–29100, 2021. 2

[47] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3D dynamic facial expression database. In *IEEE Int. Conf. and Workshops on Automatic Face and Gesture recognition (FG)*, pages 1–6. IEEE, 2013. 4, 5

[48] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3D face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1097–1106, June 2019. 2

[49] Xiangyu Zhu, Chang Yu, Di Huang, Zhen Lei, Hao Wang, and Stan Z Li. Beyond 3dmm: Learning to capture high-fidelity 3D face shape. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2022. 2

[50] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conf. on Computer Vision (ECCV)*, pages 250–269. Springer, 2022. 2