How much does input data type impact final face model accuracy?

Jiahao Luo[1] Fahim Hasan Khan[1] Issei Mori[2] Akila de Silva[1] Eric Sandoval Ruezga[1]
Minghao Liu[1] Alex Pang[1] James Davis[1]
[1]University of California, Santa Cruz; [2]University of California, San Diego
{jluo53,davisje}@ucsc.edu

Abstract

Face models are widely used in image processing and other domains. The input data to create a 3D face model ranges from accurate laser scans to simple 2D RGB photographs. These input data types are typically deficient either due to missing regions, or because they are under-constrained. As a result, reconstruction methods include embedded priors encoding the valid domain of faces. System designers must choose a source of input data and then choose a reconstruction method to obtain a usable 3D face. If a particular application domain requires accuracy X, which kinds of input data are suitable? Does the input data need to be 3D, or will 2D data suffice? This paper takes a step toward answering these questions using synthetic data. A ground truth dataset is used to analyze accuracy obtainable from 2D landmarks, 3D landmarks, low quality 3D, high quality 3D, texture color, normals, dense 2D image data, and when regions of the face are missing. Since the data is synthetic it can be analyzed both with and without measurement error. This idealized synthetic analysis is then compared to real results from several methods for constructing 3D faces from 2D photographs. The experimental results suggest that accuracy is severely limited when only 2D raw input data exists.

1. Introduction

Three dimensional face models are used in a diverse set of application domains, including biometric identification [10], 3D avatars [54], social media filters, and photo editing [46]. All of these applications benefit from high accuracy 3D models, however the degree of accuracy required varies between a realistically rendered movie and a fun mobile app with cartoon ears.

The input data to create a 3D face model has a wide variation, ranging from 0.1 mm accurate laser scans to simple 2D RGB photographs. Raw data is usually not directly usable, whether due to holes in scanned meshes or because 2D data is insufficiently constrained to determine 3D. A common solution is to apply a model to constrain or clean up the raw data. The model can take many forms, for example it could be an explicit prior applied to the distribution of a parameter space, or implied by a specific neural network architecture and the trained node weights. In either case the goal is to best use the limited input data to predict a valid and complete 3D face.

System designers must choose a source of input data and then choose a model and a method of fitting to obtain a usable 3D face. The vast majority of research has focused on novel sensor designs to improve the raw data, or novel reconstruction methods and prior models which do a better job producing faces from under-constrained data. In contrast, this paper seeks to answer the system designers question - If my application domain requires accuracy X, how good does my input data need to be? Must I start from 3D data, or can I start from a photograph and predict a 3D model which is good enough? If my user is wearing glasses and my data is missing eyes, can I predict something based on observation of the rest of the face? What about if there is a mask blocking the mouth region? How much accuracy will that cost me? In order to provide insight into questions of this kind, we start with an existing ground truth dataset, use an existing prior model, and then evaluate the accuracy obtainable from different raw data types.

Our input data is derived from a publicly available dataset which has 100s of 3D face scans in correspondence, as well as multi-view 2D images of the same subjects [47]. This allows us to compare final 3D accuracy when raw input data is 2D feature points, 2D photographs, low quality 3D, 3D feature points, high quality 3D with partial missing data, and complete high quality 3D.

Our primary analysis makes use of "synthetic" experiments. This allows extensive analysis both with and without noise. Input data is constructed directly from ground truth by removing data to mimic measurement of only landmarks, or only low resolution 3D, or missing eyes. These experiments test whether the prior model can predict accurate results for the unobserved portion of the face. We check that the synthetic results generalize by conducting "real" experiments using only RGB photographs as input and testing

the behavior of several published methods on this very limited input data type.

Hundreds of papers exist proposing prior models ranging from simple interpolation to modern deep learning. Reconstruction accuracy is necessarily tied to the prior model chosen, and this year's state of the art will not be as good as next year's. We do not attempt to provide an indication of the absolute accuracy obtainable, but rather provide a comparison of error magnitudes when starting from different raw inputs. For our experiments, we choose a Morphable Face Model because it is one of the most widely used models, it has existed for 20 years [9], surveys exist [18], and it has some mathematical similarities to blendshape models that are industry standard in animation [27]. This model is a simple linear system constructed by finding the principal components of a training set, and fitting data to a linear sum of these components. Since this is perhaps the simplest possible model (linear), it is well understood by many researchers, and thus we hope allows easy interpretation of our results.

The contribution of this paper is a careful analysis of 3D facial reconstruction accuracy when starting from input data with various levels of completeness.

2. Related Work

3D Shape Scanning: Acquiring 3D geometry is not restricted to faces, has a long history, and many surveys exist [7, 12]. Methods range from laser scanning [1] to shape from shading [51]. Modern consumer level 3D sensors include active stereo (Intel RealSense) [25], time-of-flight (PMD, Kinect.v2) [26], and structured light (Kinect.v1, iPhoneX) [52]. To the extent these methods employ a prior on shape, it is typically restricted to smoothness, continuity, and other local surface constraints. High quality face capture often uses these same 3D acquisition methods, but with special purpose capture gantries that deliver higher accuracy than consumer devices [6, 41]. The data used in this study was acquired using a multiview stereo system with 68 cameras [47].

Morphable face models: The label Morphable Face Model was popularized in Blanz and Vetter's seminal paper [9], however the general framework of representing face variation as a linear combination of principle components has clear roots in EigenFaces in the image domain [38], and Active Appearance Models which encode 2D shape landmarks and appearance [13]. The variations proposed in the last two decades are too numerous to list exhaustively and we refer the reader to an excellent survey [18]. Most models include variation in both identity and expression, with these factors being combined additively [9], multiplicatively [39], or non-linearly [28, 29, 37]. Some methods encode the shape deformation globally [9], some subdivide into local parts [35], and some combine with a muscle or other physical model [23]. In this paper we use the simplest variant as the prior in our synthetic experiments.

Blendshapes: Animation tools such as those used in movie and game studios often represent facial deformation as blendshapes or morph-targets [21, 34, 36]. These are similar to morphable models in using a linear basis to represent shape. However instead of a orthogonal basis of principle components, blendshapes use a non-orthogonal basis of semantically meaningful exemplar facial expressions [27].

Other face reconstruction priors: Other underlying 3D face model priors have been used. A single template mesh can be used as a prior and warped to match features in a 2D image [31]. Convolutional networks have been used directly on meshes [30]. When no 3D model is available at all, a collection of images can be used to train a deep neural network to act as a prior on 3D reconstruction [4, 20, 33, 43, 45]. Zollhöfer et al. provide a survey article discussing many of these [55].

3D Datasets: There are a wide variety of 3D face datasets available. Some contain only a few individuals in animated sequences [14], and some a large number of individuals in only a single pose [16]. Some contain color and texture information [15], while others do not [50]. We choose a dataset derived from high accuracy scans, which contains hundreds of individuals.

Accuracy comparison: Most 3D face reconstruction papers provide evidence of reconstruction accuracy, and comparison with prior reconstruction methods using the same data type is common. Comparative analysis of different models exist [11]. However, no existing work has quantified error as a function of source data type, the goal of this paper.

3. Synthetic Data Analysis

3.1. Method

Choice of input data impacts eventual 3D reconstruction accuracy. Starting from our test data which contains complete 3D faces, we remove part of the data to simulate common capture conditions. For each condition we predict a full 3D face, and evaluate the accuracy of reconstructed vertices against known ground truth.

We take advantage of the fact that our test data is in known correspondence. In real applications, a reconstruction algorithm needs to both find the best fit parameters of the model and find the correspondence between captured data points and model data points. Establishing correspondence is itself a difficult process [2]. In these experiments we assume perfect correspondence, to allow an evaluation of just the reconstruction errors due to limited input data.

Dataset: Our training and testing data is derived from high resolution face scans of 766 people and is reported by the

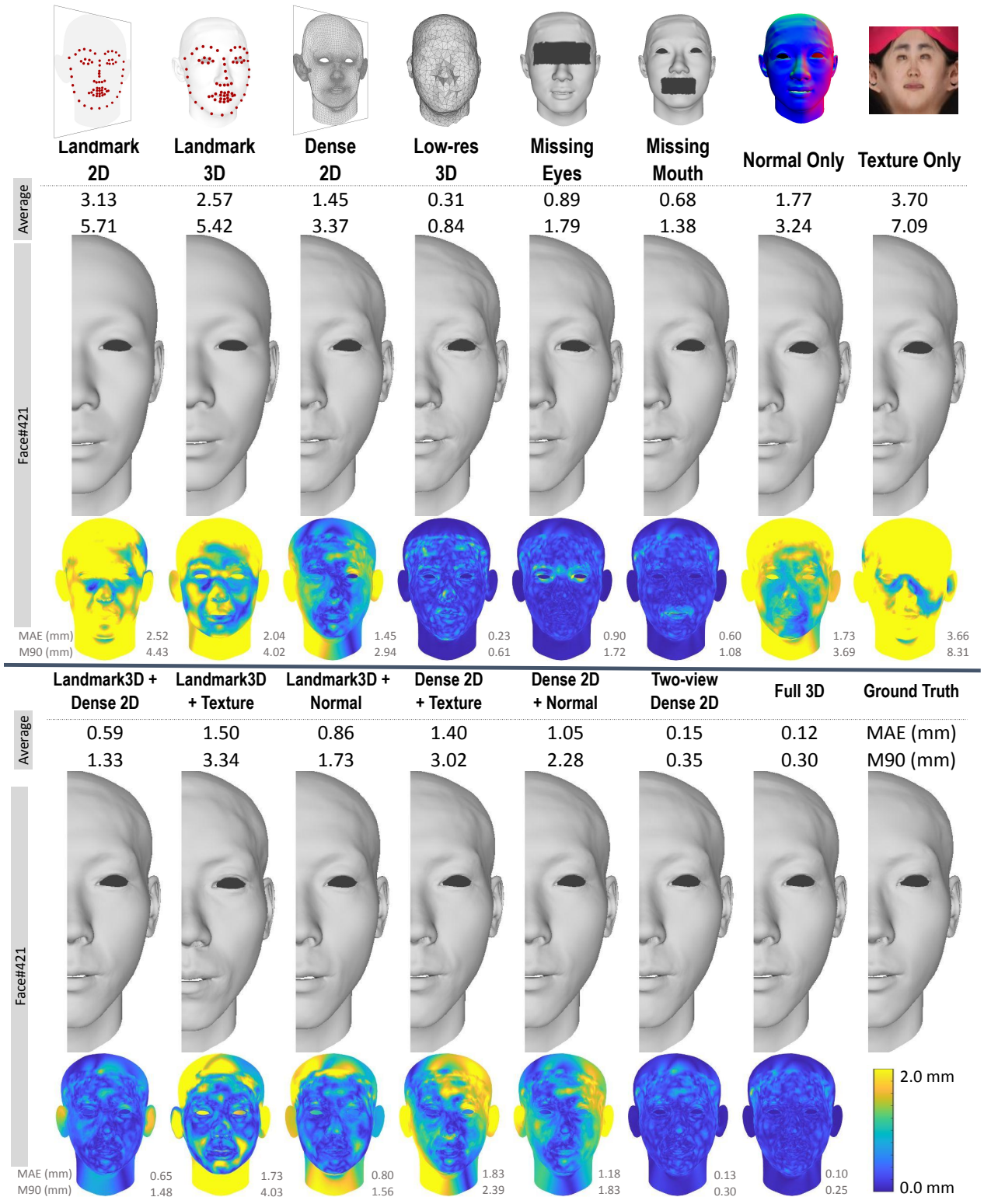

Figure 1. Comparison of reconstruction error, when using a variety of source data types. Numeric results are averages over the test dataset, reported for both mean absolute error (MAE) and max error after rejecting 10% outliers (M90). Also shown is one specific example face with corresponding colored error maps, MAE, and M90. The prior model is sufficiently powerful that any source data type can produce a reasonable face, even when sections of the face are missing entirely. Drastically reduced input data types such as Landmarks result in overly smooth meshes with high error. 2D data performs noticeably worse than 3D, even when 3D data is low resolution. Two viewpoints of 2D data performs much better than one viewpoint, and provides excellent accuracy. Since we are most interested in comparing predicted shape from limited data, we calculate error over only the missing data when applicable.

dataset creators to be 0.3mm accurate [47]. The data is provided in correspondence, meaning that the data is in a shared 3D mesh topology. All faces have the same number of vertices (26,317), and the same triangulation. Each 3D face, f_i, can thus be represented as a single vector containing vertex locations, i.e. $f_i = [x_1, y_1, z_1, ...x_n, y_n, z_n]^T$. When using additional data such as texture, we append the color values to f_i. We segment this dataset randomly into 677 training faces and 89 testing faces.

Prior: The prior for faces in our synthetic analysis is a Morphable Model. We intentionally choose the simplest variation and introduce it only briefly, since it is widely understood [9].

A morphable model is created as a linear combination of faces. We subtract the mean face, f_M, from each face to produce vectors that encode only variation from the mean, $\hat{f}_i = f_i - f_M$. All faces, including both males and females, are stacked into a single matrix $\hat{F} = [\hat{f}_1, \hat{f}_2..\hat{f}_{677}]$. We use Principal Component Analysis (PCA) on $\hat{F}$ to find principal components, C. A new face, $\hat{f}_x = C * w$, can be created as a linear combination of these components, where $w = [w_1, w_2, ...w_i]^T$ are the relative weights of the first i components. The number of components to retain is determined experimentally.

A newly observed face can be approximately represented as a least squares fit, solving for weights using the pseudo-inverse of C, $w = C^{-1} * \hat{f}_x$. There will frequently be missing or additional data in the vector $\hat{f}_x$, for example because only some of the 3D points are available, or data such as texture is appended. In these cases an optimization of the correct weights, $\widetilde{w}$, can still be obtained as a least squares fit to all the data which does exist. A prediction of the full 3D face is obtainable with C and $\widetilde{w}$.

Input data types:

We consider data types which match the options practically available in real systems.

Full 3D: We suppose a 3D scanner returns a complete head scan, with no missing data. We fit all 3D data to the model. Errors are due to the limited representational ability of the model itself. This is the minimum achievable error given a specific prior model.

Missing eyes, missing mouth: A common scenario is users wearing glasses/mask. This results in missing data near the eyes/mouth. We remove these datapoints to understand how well the model can predict this information from the context of the rest of the face.

Low-res-3D: The 3D cameras built into consumer devices are much lower resolution than our dataset. In order to evaluate whether this lower resolution data is sufficient to predict high accuracy faces we randomly select 800 data points, and use this lower resolution model to predict a complete face.

Landmark-2D, Landmark-3D: Data in many published methods is restricted to an image from which facial landmarks are extracted. In the 3D case this data might come from a very limited 3D camera. In the 2D case, we factor out camera viewpoint by assuming a frontal image. We keep only the 2D projection of 68 landmarks, discarding the depth.

Texture, Dense2D, Normal: In the case of a 2D RGB image there are several sources of data. If a model is used to render a synthetic face from the camera viewpoint then there is an implied correspondence between model vertices and pixel information. The information at each point includes both texture color and the 2D position of the pixel itself. If we consider photometric stereo and shape-from-shading methods then the surface normal may be available as well. Many published methods assume dense 2D color and positional data is enough to find correct model parameters. We separate these sources of data to better understand their contributions to recovering face shape. For Texture we use only the color in a normalized texture space, discarding the positional data at each vertex. For Dense2D we project all the vertices to a frontal plane, and keep only the information that is contained in an image, the 2D component of position. For Normals, we retain only the orientation at each vertex.

Dense2D-Two-Views: We include a special data type meant to simulate two images, which we encode as 2D positional data from two viewpoints. In principle this implicitly encodes 3D information, although "two images" is normally considered as a separate data type from true 3D data in the literature.

Combinations: The above fundamental data sources can be used in every possible combination. We select a few combinations to report on in the paper text, and included more in the supplemental materials.

3.2. Analysis Without Noise

Error is calculated for each evaluated input data condition over all 89 faces in the test dataset. In order to focus on error related to the choice of datatype, we factor out other sources of noise, assuming no measurement noise and perfect correspondence between data and model points.

Figure 1 shows one example face from our test dataset as well as aggregate error measures across the entire dataset. More face examples are available in supplemental materials. We report mean absolute error (MAE) of predicted points as well as maximum error after accounting for outliers by using the 90^{th} percentile value rather than absolute max (M90).

We start by questioning the validity of the entire experiment. Does a simple linear model have sufficient representational ability for this analysis? Notice that when Full

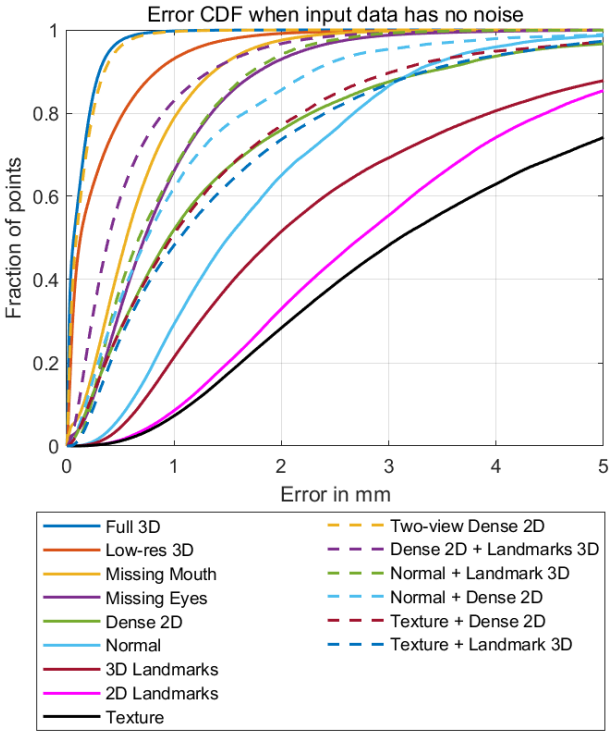

Figure 2. Cumulative error distribution, showing the fraction of reconstructed vertices with error less than a given threshold. Each solid line represents reconstruction from one type of raw input data. The dashed lines are combinations of different data types. Notice that very sparse data such as 2D and 3D Landmarks can not make accurate predictions. The highest accuracy comes from 3D data, including the implicit 3D data encoded in Two-View Dense2D.

3D data is available for fitting, M90 error is 0.3mm, matching the reported accuracy of the data itself. While there is no question that more complex models such as CoMA and FLAME are better [28, 30], this indicates that our simple model is *capable* of representing faces accurately, and thus sufficient for this analysis.

Next we consider occluded regions. Experts in 3D scanning who are accustomed to highly controlled data capture often suggest that it's necessary to scan all surfaces. However, when large regions of the face are occluded, such as eyes or mouth, the model does a surprisingly good job of predicting these regions, with correct visual appearance and maximum error staying below 2mm. The reconstruction is near perfect in the observed regions, so we report error numbers only on the predicted vertices. Notice in the color coded error map that it is possible to see the region that was reconstructed without data.

Experts in 3D scanning often dismiss consumer grade 3D as insufficient for high quality scans. However the Low-res-3D data type contains only 800 datapoints, and the resulting face prediction is high accuracy, with error less than 1mm.

We turn next to the data present in a single 2D image. Based on the number of papers published each year introducing new methods for predicting 3D faces from single images [19, 20, 22, 32, 33, 44, 49, 53], many researchers believe this data to be sufficient. Unfortunately none of the data types available in a single image do a good job in our analysis. Dense2D contains thousands of positional data points, but has M90 error above 3mm, and Normals are similar. Texture used alone is a terrible predictor of 3D shape with error above 6mm, suggesting that texture is best used as a tool for establishing correspondence between model and image, not as a direct predictor of shape. Combinations of this data such as Dense2D+Texture and Dense2D+Normal do better, suggesting that it is important to use all the information in the image, but these still have max error (M90) above 2mm. This analysis suggests that a single 2D image isn't a good choice of data type if you have any other option.

Some published methods include the use of very sparse data like 2D and 3D Landmarks [5, 17, 24, 28, 40, 48]. In our experiment, used alone these data sources produce very poor accuracy with M90 error above 5mm. The data is just too limited to constrain the search for the right face parameters. Interestingly, 3D Landmarks when used in combination with Dense2D provides accuracy of 1.3mm (M90), more than twice as good as either data source used alone. We hypothesize that the dense data from 2D images contains the details of face shape, but used alone without depth information can't distinguish between a shallow face and a deep face. If this is true, even the small number of 3D points available in 3D Landmarks is enough to constrain the overall low frequency face shape, resulting in good results when combined. This is interesting because it means that while a single image isn't a sufficient data source, even a tiny amount of 3D, such as what might be obtained from finding landmarks in both views on a dual-camera mobile phone, could be enough to substantially improve reconstruction algorithms.

Given the poor performance of 2D image information as a data type, we included a test of Dense2D positional data from two viewpoints. If we think like a stereo vision researcher then this should be equivalent to 3D data and provide very high accuracy, but if we think of this as just a second image input to a deep learning model then many existing papers find only a marginal effect on accuracy. In our results, the linear model is able to untangle the relationship between the two sets of 2D information and provides accuracy of 0.35mm M90 error, nearly as good as having full 3D information. This implies that giving single image reconstruction algorithms access to a second image should in principle produce substantially better results. Since a second image is usually easy to obtain, this seems like a

promising avenue for future exploration.

One last thing to notice about Figure 1 has nothing to do with the error numbers themselves. All the rendered faces look nearly visually identical. Only in the case of Landmarks can the authors tell the difference from ground truth. On the one hand this is excellent news. If the goal is to produce a 3D face that looks right, then almost any input data coupled with even a very simple model is good enough. On the other hand, rendered models are perhaps the most commonly presented evaluation metric in published papers, but these results suggest that visual inspection is a terrible way to judge real 3D accuracy, since these visually identical results have M90 errors ranging more than an order of magnitude from only 0.3mm to over 5mm.

We designed a user study to test this observation in which we asked Mechanical Turk users to choose which of eight face reconstructions was closest to a comparison ground truth image. One option had very low error ($<$ 0.3mm MAE) while the remaining seven had high error ($>$ 1.0mm MAE). We collected 100 trials (10 different faces each with 10 different users). When the user was presented with the same face options in three re-shuffled positions on the page, they were able to consistently make the correct choice only 3% of the time. Viewers indeed found it difficult to judge accuracy from rendered images alone. In comparison, when presented with color coded error maps, users correctly identified the low error model 99% of the time.

Any single measure of error, such as MAE or M90, only tells part of the story. We also evaluated median, MSE, several other outlier ratios, and true max. We chose MAE and M90 as the most representative in the analysis above. Figure 2 shows error for each datatype as a cumulative distribution function, which provides a more complete picture. A horizontal line through the plot at 0.9 is identical to the M90 error metric reported above. However this plot can be read in the other direction as well. Suppose a application requires error below 2mm. We can see that only 32% of the points derived from 2D landmarks are within this tolerance, while 95% of those from low resolution 3D are.

Evaluating error across data types is a challenge. It is insufficient to use exactly the same model in all cases. Most prior models contain thresholds and hyperparameters that require tuning, and a fair comparison demands tuning for each scenario. Importantly, we want to use a model which allows maximum face shape variability, without overfitting the available data. There are multiple ways to constrain variability including a prior on the allowed distribution of model parameters. In the case of morphable models, heavily regularizing higher dimensional eigenvectors forces low variability in these components limiting the models range of expressiveness and thus tendency to overfit. To produce as fair a comparison as possible, we started with 600 eigenvectors and then checked for overfitting using each input datatype. We reduced expressiveness of the model when needed by exhaustively searching for the optimum number of eigenvectors which minimized MAE error.

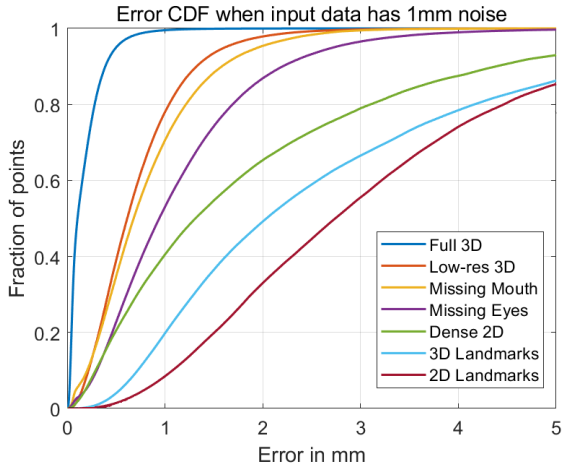

Figure 3. In order to evaluate whether our analysis generalizes when noise is present, we plot the cumulative error distribution when input has 1mm Gaussian noise added, showing the fraction of reconstructed vertices with less than a given threshold error. The overall relationship between datatypes is the same as in the no noise analysis. Median error increased in each case, but by less than the 1mm noise that was added.

3.3. Analysis With Noise

One possible limitation of our analysis is that real data is corrupted with noise caused by poor image sensors, failed landmark detectors, and incorrect correspondence. In our primary analysis we intentionally removed these factors to understand the data types under optimal conditions. In order to investigate the affects of noise, in this section we treat these sources together and add Gaussian random noise to each "measured" datapoint.

Figure 3 shows a cumulative distribution plot in the presence of Gaussian noise with 1.0mm standard deviation. Other choices of noise distribution and magnitude are reported in the supplemental material. We include only primary data types, not combinations, and leave out Texture and Normals since it's not possible to directly compare color and orientation noise with positional noise. Notice that the overall relationship between datatypes is similar to the no-noise case. The error has somewhat increased, but the conclusions of the synthetic analysis appear to generalize when noise is present.

3.4. Sensitivity to Structured Errors

Real errors in measurement are often structured as opposed to Gaussian. To investigate the possible effects of

structured bias we consider the case of Two View Dense 2D. In the primary analysis we assumed perfect knowledge of camera calibration between the two viewpoints. A fair criticism would be that unlike our analysis, real methods must simultaneously estimate camera viewpoint and face shape. We thus introduce a structured error in terms of incorrectly estimated camera viewpoint. Figure 4 shows MAE as a function of angular error in estimating camera viewpoint. Although MAE does increase with viewpoint estimation error, note that sensitivity is relatively low. Within a relatively large range of ± 10 degrees of viewpoint estimation error, MAE stays below 0.75mm, better than all other combinations of 2D image derived data types. Thus in this case, the conclusions of the primary noise-free analysis appear to generalize when structured errors are present.

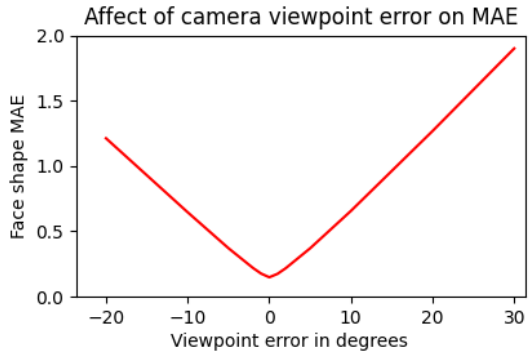

Figure 4. Reconstruction is performed using Two-view Dense2D data with varying amounts of error in the assumed camera viewpoint. Incorrect camera viewpoint increases reconstruction error, but even with 10 degree angular error in camera viewpoint, reconstruction MAE remains below 0.75mm.

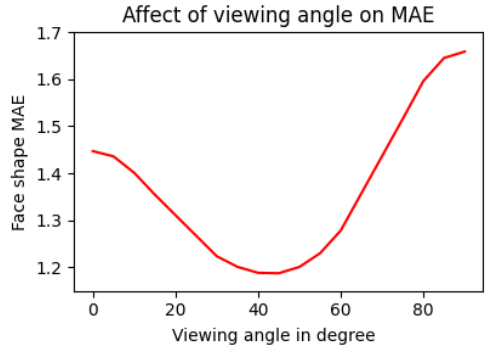

Figure 5. Changes to input data, even within a single data type matter. We show the effect of viewing angle on reconstruction MAE when using single view Dense2D positional data to reconstruct shape. A frontal face is at 0 degrees, while a profile view is 90 degrees. Note that error is substantially lower when the face is viewed at a 45 degree angle, rather than frontally at 0 degrees.

3.5. Within Data Type Variation

The primary analysis of this paper investigates how much changing input data type impacts accuracy. However, even within a single data type seemingly simple changes can have an effect on accuracy. As an example, consider Dense2D data available from a 2D image. Does the angle of viewing matter?

We rotated our training and test data to simulate different viewing angles and trained a linear model for each angle. Using the matching Dense2D test data, always in perfect correspondence with correct view angle, we reconstructed 3D faces. A plot of MAE as a function of viewing angle is provided in Figure 5. Note that error *is* strongly affected by view angle, and minimized when the face is viewed at a 45 degree angle. The change in error between frontal (1.45mm) and 45 degrees (1.2mm) is 20%. This is a significant change in accuracy, especially considering that many papers introducing new models report 5-10% improvements over prior work. Input data choices clearly matter in non-obvious and subtle ways, and carefully investigating the properties of input data is thus worth research attention.

3.6. Generalization to Modern Methods

Modern face reconstruction methods are substantially more sophisticated than a simple linear model. Our synthetic analysis suggested that high quality reconstruction is not possible using only 2D data. However it is possible that our analysis does not generalize, and modern computer vision models perform better. In order to investigate this possibility, we try nine existing methods [3,4,6,20,22,32,42,44,49]. Each is provided with example photographs and ground truth 3D from our test dataset.

We derive a 3D mesh in each case using author supplied code without modification. The 3D mesh is scaled and aligned to ground truth using the Iterative Closest Point (ICP) algorithm [8]. Each vertex in the ground truth mesh is assumed to be in correspondence with the closest point on the reconstructed mesh for purposes of computing error. The tested methods often produce an incomplete face due to visibility in a single image, and the coverage of the face is inconsistent. Thus we limit error computation to "valid" regions by manually cropping the ground truth mesh to match each specific reconstructed example.

We chose five recent methods which take a single 2D image as input, and three recent methods which allow three images as input. We also test two-view passive stereo, because Two-View Dense2D produced excellent results in our analysis and because multiview stereo is the approach used in high accuracy face scanners [6].

All methods were given access to high resolution images as input, although some methods used downsampled data for reasons of efficiency. Thus the comparison provided is not completely fair. The multiview methods including

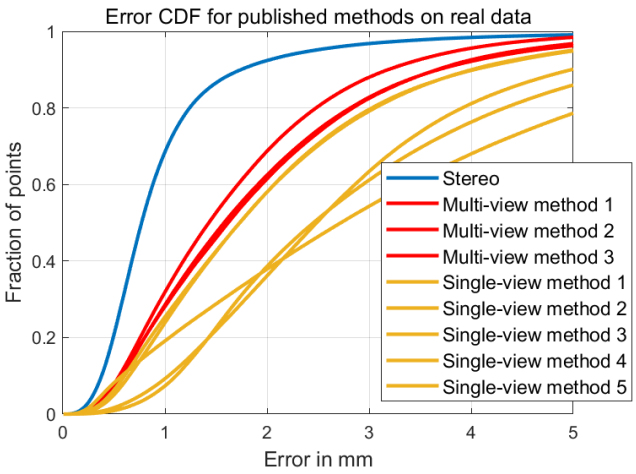

Figure 6. When reconstructing 3D faces using published methods on real 2D photographs, single view 2D methods have low accuracy. Multi-view 2D methods have access to implicit 3D information, and perform better. Stereo does not make use of any face prior, but explicitly reconstructs 3D and shows the highest accuracy. This improvement from adding a second viewpoint is significant. For example, if 2mm accuracy is required, the single view methods provide less than 60% of data within this tolerance, while Stereo provides 90%. These experiments on real data support the conclusions of our primary analysis on synthetic data.

stereo had access only to images, not camera information.

Figure 6 provides cumulative error curves for each test. We have intentionally avoided labeling curves with the precise paper citation because we want the focus to be on input data type, not the specific model. All of the papers we selected are highly regarded, including a CVPR best paper award and multiple highly ranked methods from the NoW Challenge. We want to caution against inadvertently assuming poor performance of specific methods, because we did not adjust model parameters perfectly and because our ICP alignment was not as carefully tweaked as it could be. Instead we want to focus on the aggregate observations across methods, and the relative performance of different data types.

The single view methods that have access to only 2D information show the least overall accuracy. The multiview methods perform better, presumably because they have access to 3D information encoded implicitly in the multiple views. Two-view stereo reconstruction explicitly estimated camera pose, used the multiview information to compute 3D, and performed well. The cumulative error curve for Stereo shows that the majority of datapoints are below 1mm error, and 90% of the datapoints are within 2mm of ground truth. The overall trends match our synthetic results using a simple linear model, and seem to suggest that the results in this paper generalize to modern models as well as real image conditions.

4. Discussion

This work arose from a query by a company shipping a 3D face scanning product. The question posed to us as academic researchers was ”Can we replace the 3D camera on mobile phones with a regular 2D camera and still get good results?” Unfortunately, the company couldn’t get a consistent answer from academics. When you ask this question to someone who designs 3D cameras they will say “No! You must have a 3D camera for high accuracy”. In contrast when you ask this question to someone who works on Machine Learning they will say ”There are lots of recent papers that get great results using only 2D images. There is no longer a need for 3D cameras”. Both of these positions are partially correct, and we could find no prior work rigorously comparing input data types, so we ran this study.

This paper provides evidence that given a well trained prior model, almost any input data is sufficient to recover a 3D face which visually looks correct. This includes filling in missing data which wasn’t observed by the sensor. For applications which just need to look right, 2D input data is very likely sufficient. However 2D input data, even dense 2D data, seems insufficient for applications intended for high accuracy 3D use such as measuring physical distances on the face. When this is the application need, it appears that 3D input data is required.

Our experimental results that 3D raw data contains more information than 2D data is not itself surprising, and certainly other researchers have noted the limitations of geometry from 2D images [5, 18, 32]. However, this study provides a numerical comparison across many input datatypes which has not previously existed.

As an example of how this analysis might influence research choices, many researchers are actively working on 3D faces from 2D RGB photographs. The analysis in this paper suggests when the goal is accuracy, increasing the input data to include low resolution 3D, or to include two image viewpoints instead of one, may result in substantial accuracy gains. Since many modern mobile phones contain low resolution 3D sensors, and dual cameras, this seems like a promising avenue for increased research attention.

There are limitations to this study of course. There are many possible variations of data, model, error metric and existing methods. It’s unlikely that we happened to choose the reader’s preferred combination and we hope followup papers will address more possibilities. In addition, many published papers seek to deal with facial expressions, a factor completely ignored in this study.

This paper provides a careful evaluation of *relative* 3D facial reconstruction accuracy while varying the input data type. We hope this work inspires both additional research on the merits of various input data types, as well as encouraging researchers to consider selection of data as carefully as they consider selection of model.

References

[1] Gerald J. Agin and Thomas O. Binford. Computer description of curved objects. *IEEE Computer Architecture Letters*, 25(04), 1976. 2

[2] Dragomir Anguelov, Praveen Srinivasan, Hoi-Cheung Pang, Daphne Koller, Sebastian Thrun, and James Davis. The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. *Advances in neural information processing systems*, 17, 2005. 2

[3] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3d face reconstruction via in-network optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 7

[4] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep Facial Non-Rigid Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5850–5860, 2020. 2, 7

[5] Anil Bas and William A. P. Smith. What does 2D geometric information really tell us about 3D face shape? In *International Journal of Computer Vision*, 2019. 5, 8

[6] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9. 2010. 2, 7

[7] Paul J. Besl. Active optical range imaging sensors. In *Advances in machine vision*, pages 1–63. Springer, 1989. 2

[8] Paul J. Besl and Neil D. McKay. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 7

[9] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *ACM SIGGRAPH*, 1999. 2, 4

[10] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9), 2003. 1

[11] Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. Review of statistical shape spaces for 3D data with comparative analysis for human faces. *Computer Vision and Image Understanding*, 128:1–17, 2014. 2

[12] Fang Chen, Gordon M. Brown, and Mumin Song. Overview of 3-D shape measurement using optical methods. *Optical Engineering*, 39(1), 2000. 2

[13] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *European conference on computer vision*, pages 484–498. Springer, 1998. 2

[14] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[15] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[16] Hang Dai, Nick Pears, William AP Smith, and Christian Duncan. A 3D morphable model of craniofacial shape and texture variation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2

[17] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 202018. 5

[18] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D Morphable Face Models: Past, Present, and Future. *ACM Transactions on Graphics*, 39(5):157:1–157:38, June 2020. 2, 8

[19] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. In *ACM Transactions on Graphics (TOG)*, 2021. 5

[20] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 2, 5, 7

[21] Barbara Flueckiger. Computer-generated characters in Avatar and Benjamin Button. *Digitalitat und Kino. Translation from German by B. Letzler*, 1, 2011. 2

[22] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z. Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, 2021. 5, 7

[23] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. Phace: Physics-based face modeling and animation. *ACM Trans. on Graphics (TOG)*, 36(4), 2017. 2

[24] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3d face reconstruction with geometry details from a single image. In *IEEE Transactions on Image Processing*, 2018. 5

[25] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 2

[26] Robert Lange and Peter Seitz. Solid-state time-of-flight range camera. *IEEE Journal of quantum electronics*, 37(3), 2001. 2

[27] John P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H. Pighin, and Zhigang Deng. Practice and Theory of Blendshape Facial Models. *Eurographics (State of the Art Reports)*, 1(8):2, 2014. 2

[28] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6), 2017. 2, 5

[29] Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1860–1873, 2017. 2

[30] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 704–720, 2018. 2, 5

[31] Joseph Roth, Yiying Tong, and Xiaoming Liu. Unconstrained 3D face reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[32] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5, 7, 8

[33] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017. 2, 5

[34] Greg Singer. The two towers: Face to face with gollum. *Animation World Network*, 1, 2003. 2

[35] J. Rafael Tena, Fernando De la Torre, and Iain Matthews. Interactive region-based linear 3D face models. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011. 2

[36] Sham Tickoo. *Autodesk Maya 2011: A Comprehensive guide*. Pearson Education India, 2017. 2

[37] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 2

[38] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 1991. 2

[39] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM Trans. on Graphics (TOG)*, 24(3), 2005. 2

[40] Pengrui Wang, Yi Tian, Wujun Che, and Bo Xu. Efficient and accurate face shape reconstruction by fusion of multiple landmark databases. In *IEEE International Conference on Image Processing (ICIP)*, 2019. 5

[41] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Trans. on Graphics (TOG)*, 24(3), 2005. 2

[42] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019. 7

[43] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[44] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 5, 7

[45] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3D portrait from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[46] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3D-aware face component transfer. In *ACM SIGGRAPH 2011*. 2011. 1

[47] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 601–610, 2020. 1, 2, 4

[48] Jing Yuan, Xingce Wang, and Zhongke Wu. Example-guided 3d human face reconstruction from sparse landmarks. In *International Conference on Cyberworlds (CW)*, 2021. 5

[49] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2315–2324, 2019. 5, 7

[50] Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. on Graphics*, 23(3), 2004. 2

[51] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8), 1999. 2

[52] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2), 2012. 2

[53] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *IEEE conference on computer vision and pattern recognition*, 2016. 5

[54] Michael Zollhöfer, Michael Martinek, Günther Greiner, Marc Stamminger, and Jochen Süßmuth. Automatic reconstruction of personalized avatars from 3D face scans. *Computer Animation and Virtual Worlds*, 22(2-3), 2011. Wiley. 1

[55] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. In *In Computer Graphics Forum (Vol. 37, No. 2, pp. 523-550)*, May 2018. 2