

Towards Metrical Reconstruction of Human Faces

Wojciech Zieleńka Timo Bolkart Justus Thies

Max Planck Institute for Intelligent Systems, Tübingen, Germany

{wojciech.zielonka, timo.bolkart, justus.thies}@tuebingen.mpg.de

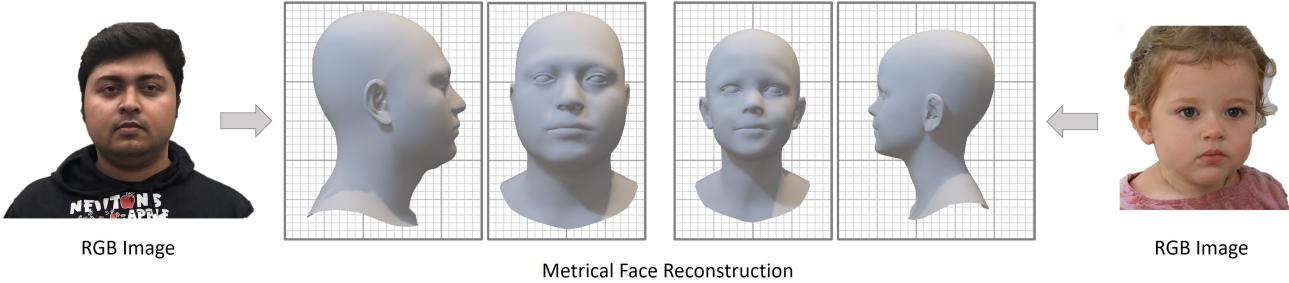


Figure 1. A single RGB image of a subject serves as input to MICA, which predicts a metrical reconstruction of the human face. This reconstruction can be used to measure distances for virtual try-on or alike, and we show how it can be used for analysis-by-synthesis-based face tracking. Images from NoW [64] and StyleGan2 [42].

Abstract

Face reconstruction and tracking is a building block of numerous applications in AR/VR, human-machine interaction, as well as medical applications. Most of these applications rely on a metrically correct prediction of the shape, especially, when the reconstructed subject is put into a metrical context (i.e., when there is a reference object of known size). A metrical reconstruction is also needed for any application that measures distances and dimensions of the subject (e.g., to virtually fit a glasses frame). State-of-the-art methods for face reconstruction from a single image are trained on large 2D image datasets in a self-supervised fashion. However, due to the nature of a perspective projection they are not able to reconstruct the actual face dimensions, and even predicting the average human face outperforms some of these methods in a metrical sense. To learn the actual shape of a face, we argue for a supervised training scheme. Since there exists no large-scale 3D dataset for this task, we annotated and unified small- and medium-scale databases. The resulting unified dataset is still a medium-scale dataset with more than 2k identities and training purely on it would lead to overfitting. To this end, we take advantage of a face recognition network pretrained on a large-scale 2D image dataset, which provides distinct features for different faces and is robust to expression, illumination, and camera changes. Using these features, we train our face shape estimator in a su-

pervised fashion, inheriting the robustness and generalization of the face recognition network. Our method, which we call MICA (Metrical fAce), outperforms the state-of-the-art reconstruction methods by a large margin, both on current non-metric benchmarks as well as on our metric benchmarks (15% and 24% lower average error on NoW, respectively). Project website: <https://zielon.github.io/mica/>.

1. Introduction

Learning to reconstruct 3D content from 2D imagery is an ill-posed inverse problem [4]. State-of-the-art RGB-based monocular facial reconstruction and tracking methods [20, 25] are based on self-supervised training, exploiting an underlying metrical face model which is constructed using a large-scale dataset of registered 3D scans (e.g., 33000 scans for the FLAME [51] model). However, when assuming a perspective camera, the scale of the face is ambiguous since a large face can be modeled by a small face that is close to the camera or a gigantic face that is far away from the camera. Formally, a point $\mathbf{x} \in \mathbb{R}^3$ of the face is projected to a point $\mathbf{p} \in \mathbb{R}^2$ on the image plane with the projective function $\pi(\cdot)$ and a rigid transformation composed of a rotation $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ and a translation $\mathbf{t} \in \mathbb{R}^3$:

$$\mathbf{p} = \pi(\mathbf{R} \cdot \mathbf{x} + \mathbf{t}) = \pi(s \cdot (\mathbf{R} \cdot \mathbf{x} + \mathbf{t})) = \pi(\mathbf{R} \cdot (s \cdot \mathbf{x}) + (s \cdot \mathbf{t})).$$

The perspective projection is invariant to the scaling factor $s \in \mathbb{R}$, and thus, if \mathbf{x} is scaled by s , the rigid transforma-

tion can be adapted such that the point still projects onto the same pixel position p by scaling the translation t by s . In consequence, face reconstruction methods might result in a good 2D alignment but can fail to reconstruct the metrical 3D surface and the meaningful metrical location in space. However, a metric 3D reconstruction is needed in any scenario where the face is put into a metric context. E.g., when the reconstructed human is inserted into a virtual reality (VR) application or when the reconstructed geometry is used for augmented reality (AR) applications (teleconferencing in AR/VR, virtual try-on, etc.). In these scenarios, the methods mentioned above fail since they do not reproduce the correct scale and shape of the human face. In the current literature [27, 64, 89], we also observe that methods use evaluation measurements not done in a metrical space. Specifically, to compare a reconstructed face to a reference scan, the estimation is aligned to the scan via Procrustes analysis, including an optimal scaling factor. This scaling factor favors the estimation methods that are not metrical, and the reported numbers in the publications are misleading for real-world applications (relative vs. absolute/metrical error). In contrast, we aim for a metrically correct reconstruction and evaluation that directly compares the predicted geometry to the reference data without any scaling applied in a post-processing step which is fundamentally different. As discussed above, the self-supervised methods in the literature do not aim and cannot reconstruct a metrically correct geometry. However, training these methods in a supervised fashion is not possible because of the lack of data (no large-scale 3D dataset is available). Training on a small- or medium-scale 3D dataset will lead to overfitting of the networks (see study in the appendix). To this end, we propose a hybrid method that can be trained on a medium-scale 3D dataset, reusing powerful descriptors from a pre-trained face recognition network (trained on a large-scale 2D dataset). Specifically, we propose the usage of existing 3D datasets like LYHM [18], FaceWarehouse [11], Stirling [28], etc., that contain RGB imagery and corresponding 3D reconstructions to learn a metrical reconstruction of the human head. To use these 3D datasets, significant work has been invested to unify the 3D data (i.e., to annotate and non-rigidly fit the FLAME model to the different datasets). This unification provides us with meshes that all share the FLAME topology. Similar to the AMASS [54] dataset collection, we plan to release this processed data for other researchers. Our method predicts the head geometry in a neutral expression, only given a single RGB image of a human subject in any pose or expression. To generalize to unseen in the wild images, we use a state-of-the-art face recognition network [19] that provides a feature descriptor for our geometry-estimating network. This recognition network is robust to head poses, different facial expressions, occlusions, illumination changes, and different focal lengths,

thus, being ideal for our task (see Figure 3). Based on this feature, we predict the geometry of the face with neutral expression within the face space spanned by FLAME [51], effectively disentangling shape and expression. As an application, we demonstrate that our metrical face reconstruction estimator can be integrated in a new analysis-by-synthesis face tracking framework which removes the requirement of an identity initialization phase [75]. Given the metrical face shape estimation, the face tracker is able to predict the face motion in a metrical space.

In summary, we have the following contributions:

- a dataset of 3D face reference data for about 2300 subjects, built by unifying existing small- and medium-scale datasets under common FLAME topology.
- a metrical face shape predictor – MICA– which is invariant to expression, pose and illumination, by exploiting generalized identity features from a face recognition network and supervised learning.
- a hybrid face tracker that is based on our (learned) metrical reconstruction of the face shape and an optimization-based facial expression tracking.
- a metrical evaluation protocol and benchmark, including a discussion on the current evaluation practise.

2. Related Work

Reconstructing human faces and heads from monocular RGB, RGB-D, or multiview data is a well-explored field at the intersection of computer vision and computer graphics. Zollhöfer et al. [91] provide an extensive review of reconstruction methods, focusing on optimization-based techniques that follow the principle of analysis-by-synthesis. Primarily, the approaches that are based on monocular inputs are based on a prior of face shape and appearance [6, 7, 29, 30, 44, 72–75, 83, 84]. The seminal work of Blanz et al. [8] introduced such a 3D morphable model (3DMM), which represents the shape and appearance of a human in a compressed, low-dimensional, PCA-based space (which can be interpreted as a decoder with a single linear layer). There is a large corpus of different morphable models [23], but the majority of reconstruction methods use either the Basel Face Model [8, 57] or the Flame head model [51]. Besides using these models for an analysis-by-synthesis approach, there is a series of learned regression-based methods. An overview of these methods is given by Morales et al. [55]. In the following, we will discuss the most relevant related work for monocular RGB-based reconstruction methods.

Optimization-based Reconstruction of Human Faces. Along with the introduction of a 3D morphable model

for faces, Blanz et al. [8] proposed an optimization-based reconstruction method that is based on the principle of analysis-by-synthesis. While they used a sparse sampling scheme to optimize the color reproduction, Thies et al. [74, 75] introduced a dense color term considering the entire face region that is represented by a morphable model using differentiable rendering. In addition, a GPU-based second-order optimizer allows for real-time reconstruction rates running the analysis-by-synthesis loop several hundred times per second. This method has been adapted for avatar digitization from a single image [40] including hair, is used to reconstruct high-fidelity facial reflectance and geometry from a single images [85], for reconstruction and animation of entire upper bodies [76], or avatars with dynamic textures [56]. Recently, these optimization-based methods are combined with learnable components such as surface offsets or view-dependent surface radiance fields [35]. In addition to a photometric reconstruction objective, additional terms based on dense correspondence [38] or normal [1, 35] estimations of neural network can be employed. Optimization-based methods are also used as a building block for neural rendering methods such as deep video portraits [44], deferred neural rendering [73], or neural voice puppetry [72]. Note that differentiable rendering is not only used in neural rendering frameworks but is also one of the key components for self-supervised learning of regression-based reconstruction methods covered in the following.

Regression-based Reconstruction of Human Faces. Learning-based face reconstruction methods can be categorized into supervised and self-supervised approaches. A series of methods are based on synthetic renderings of human faces to perform a supervised training of a regressor that predicts the parameters of a 3D morphable model [22, 45, 61, 62]. Genova et al. [34] propose a 3DMM parameter regression technique that is based on synthetic renderings (where ground truth parameters are available) and real images (where multi-view identity losses are applied). It uses FaceNet [65] to extract features for the 3DMM regression task. Tran et al. [77] and Chang et al. [13] (ExpNet) directly regress 3DMM parameters using a CNN trained on fitted 3DMM data. Tu et al. [80] propose a dual training pass for images with and without 3DMM fittings. Jackson et al. [41] propose a model-free approach that reconstructs a voxel-based representation of the human face and is trained on paired 2D image and 3D scan data. PRN [26] is trained on ‘in-the-wild’ images with fitted 3DMM reconstructions [90]. It is not restricted to a 3DMM model space and predicts a position map in the UV-space of a template mesh. Instead of working in UV-space, Wei et al. [82] propose to use graph convolutions to regress the coordinates of the vertices. MoFA [71] is a network trained to regress the 3DMM parameters in a self-supervised fashion. As a supervision signal, it uses

the dense photometric losses of Face2Face [75]. Within this framework, Tewari et al. proposed to refine the identity shape and appearance [70] as well as the expression basis [69] of a linear 3DMM. In a similar setup, one can also train a non-linear 3DMM [79] or personalized models [14]. RingNet [64] regresses 3DMM parameters and is trained on 2D images using losses on the reproduction of 2D landmarks and shape consistency (different images of the same subject) and shape inconsistency (images of different subjects) losses. DECA [25] extends RingNet with expression dependent offset predictions in UV space. It uses dense photometric losses to train the 3DMM parameter regression and the offset prediction network. This separation of a coarse 3DMM model and a detailed bump map has been introduced by Tran et al. [78]. Chen et al. [15] use a hybrid training composed of self-supervised and supervised training based on renderings to predict texture and displacement maps. Deng et al. [20] train a 3DMM parameter regressor based on multi-image consistency losses and ‘hybrid-level’ losses (photometric reconstruction loss with skin attention masks, and a perception-level loss based on FaceNet [65]). On the NoW challenge [64], DECA [25] and the method of Deng et al. [20] show on-par state-of-the-art results. Similar to DECA’s offset prediction, there are GAN-based methods that predict detailed color maps [32, 33] or skin properties [48, 49, 63, 85] (e.g., albedo, reflectance, normals) in UV-space of a 3DMM-based face reconstruction. In contrast to these methods, we are interested in reconstructing a metrical 3D representation of a human face and not fine-scale details. Self-supervised methods suffer from the depth-scale ambiguity (the face scale, translation away from the camera, and the perspective projection are ambiguous) and, thus, predict a wrongly scaled face, even though 3DMM models are by construction in a metrical space. We rely on a strong supervision signal to learn the metrical reconstruction of a face using high-quality 3D scan datasets which we unified. In combination with an identity encoder [19] trained on in-the-wild 2D data, including occlusions, different illumination, poses, and expressions, we achieve robust geometry estimations that significantly outperform state-of-the-art methods.

3. Metrical Face Shape Prediction

Based on a single input RGB image I , MICA aims to predict a metrical shape of a human face in a neutral expression. To this end, we leverage both ‘in-the-wild’ 2D data as well as metric 3D data to train a deep neural network, as shown in Figure 2. We employ a state-of-the-art face recognition network [19] which is trained on ‘in-the-wild’ data to achieve a robust prediction of an identity code, which is interpreted by a geometry decoder.

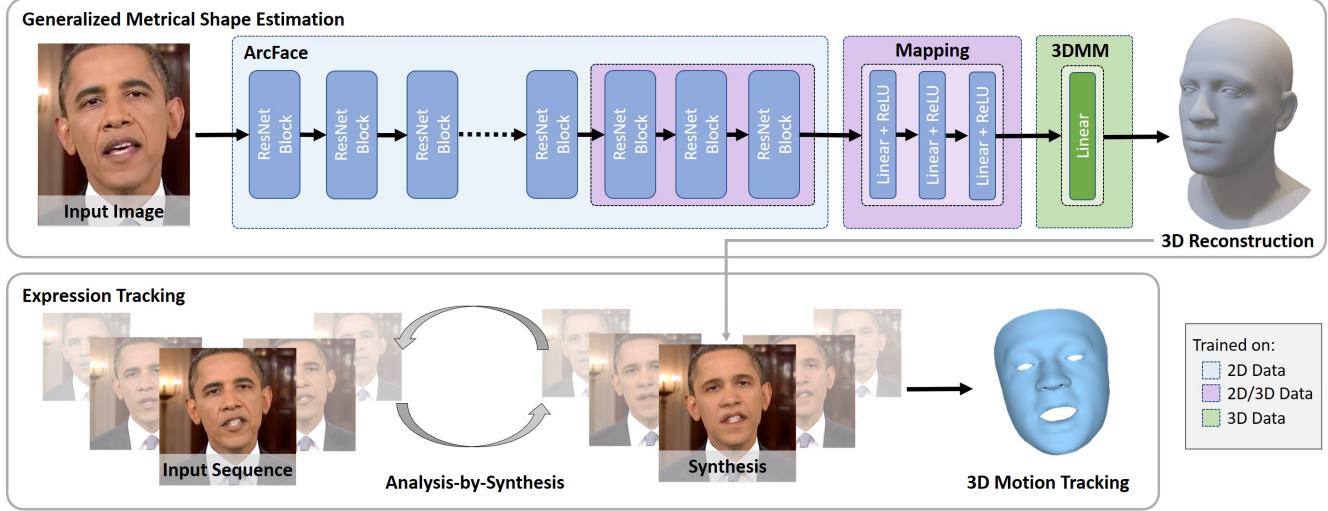


Figure 2. We propose a method for metrical human face shape estimation from a single image which exploits a supervised training scheme based on a mixture of different 2D, 2D/3D and 3D datasets. This estimation can be used for facial expression tracking using analysis-by-synthesis which optimizes for the camera intrinsics, as well as the per-frame illumination, facial expression and pose.

Identity Encoder. As an identity encoder, we leverage the ArcFace [19] architecture which is pretrained on Glint360K [2]. This ResNet100-based network is trained on 2D image data using an additive angular margin loss to obtain highly discriminative features for face recognition. It is invariant to illumination, expression, rotation, occlusion, and camera parameters which is ideal for a robust shape prediction. We extend the ArcFace architecture by a small mapping network \mathcal{M} that maps the ArcFace features to our latent space, which can then be interpreted by our geometry decoder:

$$\mathbf{z} = \mathcal{M}(\text{ArcFace}(I)),$$

where $\mathbf{z} \in \mathbb{R}^{300}$. Our mapping network \mathcal{M} consists of three fully-connected linear hidden layers with ReLU activation and the final linear output layer.

Geometry Decoder. There are essentially two types of geometry decoders used in the literature, model-free and model-based. Throughout the project of this paper, we conducted experiments on both types and found that both perform similarly on the evaluation benchmarks (see appendix C). Since a 3DMM model efficiently represents the face space, we focus on a model-based decoder. Specifically, we use FLAME [51] as a geometry decoder, which consists of a single linear layer:

$$\mathcal{G}_{3DMM}(\mathbf{z}) = \mathbf{B} \cdot \mathbf{z} + \mathbf{A},$$

where $\mathbf{A} \in \mathbb{R}^{3N}$ is the geometry of the average human face and $\mathbf{B} \in \mathbb{R}^{3N \times 300}$ contains the principal components of the 3DMM and $N = 5023$.

Supervised Learning. The networks described above are trained using paired 2D/3D data from existing, unified datasets \mathcal{D} (see Section 5). We fix large portions of the pre-trained ArcFace network during the training and refine the last 3 ResNet blocks. Note that ArcFace is trained on a much larger amount of identities, therefore, refining more hidden layers results in worse predictions due to overfitting. We found that using the last 3 ResNet blocks gives the best generalization (see Appendix A). The training loss is:

$$\mathcal{L} = \sum_{(I, \mathcal{G}) \in \mathcal{D}} |\kappa_{mask}(\mathcal{G}_{3DMM}(\mathcal{M}(\text{ArcFace}(I))) - \mathcal{G})|, \quad (1)$$

where $\mathcal{G} \sim \mathcal{D}$ is the ground truth mesh and κ_{mask} is a region dependent weight (the face region has weight 150.0, the back of the head 1.0, and eyes with ears 0.01). We use AdamW [53] for optimization with fixed learning rate $\eta = 1e-5$ and weight decay $\lambda = 2e-4$. We select the best performing model based on a validation set loss using the Florence dataset [3]. The model was trained for 160k steps on Nvidia Tesla V100.

4. Face Tracking

Based on our shape estimate, we demonstrate optimization-based face tracking on monocular RGB input sequences. To model the non-rigid deformations of the face, we use the linear expression basis vectors and the linear blendskinning of the FLAME [51] model, and use a linear albedo model [24] to reproduce the appearance of a subject in conjunction with a Lambertian material assumption and a light model based on spherical harmon-

ics. We adapt the analysis-by-synthesis scheme of Thies et al. [75]. Instead of using a multi-frame model-based bundling technique to estimate the identity of a subject, we use our one-shot shape identity predictor. We initialize the albedo and spherical harmonics based on the same first frame using the energy:

$$E(\phi) = w_{dense} E_{dense}(\phi) + w_{lmk} E_{lmk}(\phi) + w_{reg} E_{reg}(\phi), \quad (2)$$

where ϕ is the vector of unknown parameters we are optimizing for. The energy terms $E_{dense}(\phi)$ and $E_{reg}(\phi)$ measure the dense color reproduction of the face (ℓ_1 -norm) and the deviation from the neutral pose respectively. The sparse landmark term $E_{lmk}(\phi)$ measures the reproduction of 2D landmark positions (based on Google’s mediapipe [36, 43] and Face Alignment [9]). The weights w_{dense} , w_{lmk} and w_{reg} balance the influence of each sub-objectives on the final loss. In the first frame vector ϕ contains the 3DMM parameters for albedo, expression, and rigid pose, as well as the spherical harmonic coefficients (3 bands) that are used to represent the environmental illumination [59]. After this initialization, the albedo parameters are fixed and unchanged throughout the sequence tracking.

Optimization. Our prototype is implemented in PyTorch and uses Adam [46] to optimize the objective function Equation (2). While recent soft-rasterizers [52, 60] are popular, we rely on a sampling based scheme as introduced by Thies et al. [75] to implement the differentiable rendering for the photo-metric reproduction error $E_{dense}(\phi)$. Specifically, we use a classical rasterizer to render the surface of the current estimation. The rasterized surface points that survive the depth test are considered as the set of visible surface points \mathcal{V} for which we compute the energy term $E_{dense}(\phi) = \sum_{i \in \mathcal{V}} |I(\pi(\mathbf{R} \cdot p_i(\phi) + \mathbf{t})) - c_i(\phi)|$ where p_i and c_i being the i -th vertex and color of the reconstructed model, and I the RGB input image.

5. Dataset Unification

In the past, methods and their training scheme were limited by the availability of 3D scan datasets of human faces. While several small and medium-scale datasets are available, they are in different formats and do not share the same topology. To this end, we unified the available datasets such that they can be used as a supervision signal for face reconstruction from 2D images. Specifically, we register the FLAME [51] head model to the provided scan data. In an initial step, we fit the model to landmarks and optimize for the FLAME parameters based on an iterative closest point (ICP) scheme [5]. We further jointly optimize FLAME’s model parameters, and refine the fitting with a non-rigid deformation regularized by FLAME, similar to Li and Bolkart et al. [51]. In Table 1, we list the datasets that we unified

Dataset	#Subj.	#Min. Img.	Expr.
Stirling [28]	133	8	✓
D3DFACS [17]	10	videos	✓
Florence 2D/3D [3]	53	videos	✓
BU-3DFE [87]	100	83	✓
LYHM [18]	1211	2	✗
FaceWarehouse [11]	150	119	✓
FRGC [58]	525	7	✓
BP4D+ [88]	127	videos	✓

Table 1. Overview of our unified datasets. The used datasets vary in the capture modality and the capture protocol. Here, we list the number of subject, the minimum number of images per subjects, and whether the dataset includes facial expressions. In total our dataset contains 2309 subjects with FLAME topology.

for this project. We note that the datasets vary in the capturing modality and capturing script (with and without facial expressions, with and without hair caps, indoor and outdoor imagery, still images, and videos), which is suitable for generalization. The datasets are recorded in different regions of the world and are often biased towards ethnicity. Thus, combining other datasets results in a more diverse data pool. In the appendix, we show an ablation on the different datasets. *Upon agreement of the different dataset owners, we will share our unified dataset, i.e., for each subject one registered mesh with neutral expression in FLAME topology.* Note that in addition to the datasets listed in Table 1, we analyzed the FaceScape dataset [86]. While it provides a large set of 3D reconstructions ($\sim 17k$), which would be ideal for our training, the reconstructions are not done in a metrical space. Specifically, the data has been captured in an uncalibrated setup and faces are normalized by the eye distance, which has not been detailed in their paper (instead, they mention sub-millimeter reconstruction accuracy which is not valid). This is a fundamental flaw of this dataset, and also questions their constructed morphable face model as well as their reconstruction benchmark [89].

6. Results

Our experiments mainly focus on the metrical reconstruction of a human face from in the wild images. In addition, we show results for the sequential tracking of facial motions using our metrical reconstruction as initialization. In the Appendix A, we discuss further experiments and ablation studies. The following experiments are conducted with the original models of the respective publications including their reconstructions submitted to the given benchmarks. Note that these models are trained on their large-scale datasets, training them on our medium-scale 3D dataset would lead to overfitting.

NoW-Metric Challenge				Non-Metrical [64]			Metrical (mm)		
Method				Median	Mean	Std	Median	Mean	Std
Average Face (FLAME [51])	1.21	1.53	1.31	1.49	1.92	1.68			
3DMM-CNN [77]	1.84	2.33	2.05	3.91	4.84	4.02			
PRNet [26]	1.50	1.98	1.88	—	—	—			
Deng et al [20] (TensorFlow)	1.23	1.54	1.29	2.26	2.90	2.51			
• Deng et al [20] (PyTorch)	1.11	1.41	1.21	1.62	2.21	2.08			
• RingNet [64]	1.21	1.53	1.31	1.50	1.98	1.77			
• 3DDFA-V2 [37]	1.23	1.57	1.39	1.53	2.06	1.95			
• MGCNet [67]	1.31	1.87	2.63	1.70	2.47	3.02			
• UMDFA [47]	1.52	1.89	1.57	2.31	2.97	2.57			
• Dib et al. [21]	1.26	1.57	1.31	1.59	2.12	1.93			
• DECA [25]	1.09	1.38	1.18	1.35	1.80	1.64			
• FOCUS [50]	1.04	1.30	1.10	1.41	1.85	1.70			
• Ours	0.90	1.11	0.92	1.08	1.37	1.17			

Table 2. Quantitative evaluation of the face shape estimation on the *NoW Challenge* [64]. Note that we list two different evaluations: the non-metrical evaluation from the original NoW challenge and our new metrical evaluation (including a cumulative error plot on the left). The original NoW challenge cannot be considered metrical since Procrustes analysis is used to align the reconstructions to the corresponding reference meshes, including scaling. We list all methods from the original benchmark and additionally show the performance of the average human face of FLAME [51] as a reference (first row).

Stirling (NoW Protocol)	Non-Metrical						Metrical (mm)					
	Median		Mean		Std		Median		Mean		Std	
	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ
Average Face (FLAME [51])	1.23	1.22	1.56	1.55	1.38	1.35	1.44	1.40	1.84	1.79	1.64	1.57
RingNet [64]	1.17	1.15	1.49	1.46	1.31	1.27	1.37	1.33	1.77	1.72	1.60	1.54
3DDFA-V2 [37]	1.26	1.20	1.63	1.55	1.52	1.45	1.49	1.38	1.93	1.80	1.78	1.68
Deng et al. [20] (TensorFlow)	1.22	1.13	1.57	1.43	1.40	1.25	1.85	1.81	2.41	2.29	2.16	1.97
Deng et al. [20] (PyTorch)	1.12	0.99	1.44	1.27	1.31	1.15	1.47	1.31	1.93	1.71	1.77	1.57
DECA [25]	1.09	1.03	1.39	1.32	1.26	1.18	1.32	1.22	1.71	1.58	1.54	1.42
Ours w/o. Stirling	0.96	0.92	1.22	1.16	1.11	1.04	1.15	1.06	1.46	1.35	1.30	1.20

Table 3. Quantitative evaluation of the face shape estimation on the *Stirling Reconstruction Benchmark* [27] using the NoW protocol [64]. We list two different evaluations: the non-metric evaluation from the original benchmark and the metric evaluation. *Note that for this experiment, we exclude the Stirling dataset from our training set.*

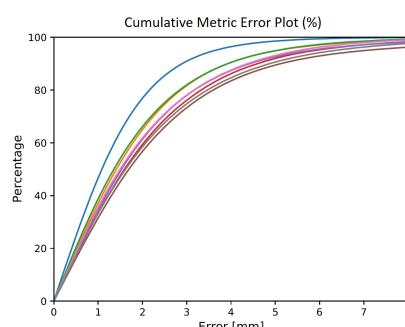
6.1. Face Shape Estimation

In recent publications, face shape estimation is evaluated on datasets where reference scans of the subjects are available. The NoW Challenge [64] and the benchmark of Feng et al. [27] which is based on Stirling meshes [28] are used in the state-of-the-art methods [20, 25, 64]. We conduct several studies on these benchmarks and propose different evaluation protocols.

Non-Metrical Benchmark. The established evaluation methods on these datasets are based on an optimal scaling step, i.e., to align the estimation to the reference scan, they optimize for a rigid alignment and an additional scaling factor which results in a non-metric/relative error. This scaling compensates for shape mispredictions, e.g., the mean error

evaluated on the NoW Challenge for the average FLAME mesh (Table 2) drops from 1.92mm to 1.53mm because of the applied scale optimization. This is an improvement of around 20% which has nothing to do with the reconstruction quality and, thus, creates a misleading benchmark score where methods appear better than they are. Nevertheless, we evaluate our method on these benchmarks and significantly outperform all state-of-the-art methods as can be seen in Tables 2 and 4 ('Non-Metrical' column).

Metrical Benchmark. Since for a variety of applications, actual metrical reconstructions are required, we argue for a new evaluation scheme that uses a purely rigid alignment, i.e., without scale optimization (see Figure 7). This new evaluation scheme enables a comparison of methods based on metrical quantities (see Tables 2 and 4) and, thus, is fun-



Stirling/ESRC Benchmark	Non-Metrical [27]						Metrical (mm)					
	Median		Mean		Std		Median		Mean		Std	
	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ
Average Face (FLAME [51])	1.58	1.62	2.06	2.08	1.82	1.83	1.70	1.62	2.19	2.09	1.96	1.85
RingNet [64]	1.56	1.60	2.01	2.05	1.75	1.76	1.67	1.64	2.16	2.09	1.90	1.81
3DDFA-V2 [37]	1.58	1.49	2.03	1.90	1.74	1.63	1.70	1.56	2.16	1.98	1.88	1.70
Deng et al. [20] (TensorFlow)	1.56	1.41	2.02	1.84	1.77	1.63	2.13	2.14	2.71	2.65	2.33	2.12
Deng et al. [20] (PyTorch)	1.51	1.29	1.95	1.64	1.71	1.39	1.78	1.54	2.28	1.97	1.97	1.68
DECA [25]	1.40	1.32	1.81	1.72	1.59	1.50	1.56	1.45	2.03	1.87	1.81	1.64
Ours w/o. Stirling	1.26	1.22	1.62	1.55	1.41	1.34	1.36	1.26	1.73	1.60	1.48	1.37

Table 4. Quantitative evaluation of the face shape estimation on the *Stirling Reconstruction Benchmark* [27]. We list two different evaluations: the non-metric evaluation from the original benchmark and the metric evaluation. This benchmark is based on an alignment protocol that only relies on reference landmarks and, thus, is very noisy and dependent on the landmark reference selection (in our evaluation, we use the landmark correspondences provided by the FLAME [51] model). We use the image file list from [64] to compute the scores (i.e., excluding images where a face is not detectable). Note that for this experiment, we exclude the Stirling dataset from our training set.

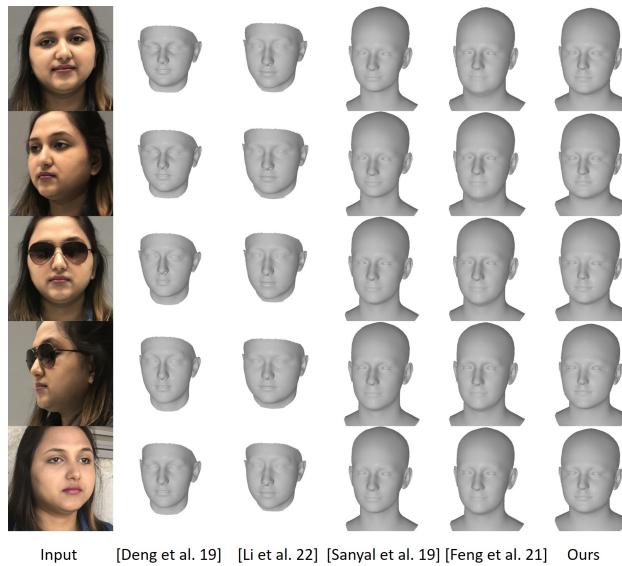


Figure 3. Qualitative results on NoW Challenge [64] to show the invariance of our method to changes in illumination, expression, occlusion, rotation, and perspective distortion in comparison to other methods.

damentally different from the previous evaluation schemes. In addition, the benchmark of Feng et al. [27] is based on the alignment using sparse facial (hand-selected) landmarks. Our experiments showed that this scheme is highly dependent on the selection of these markers and results in inconsistent evaluation results. In our listed results, we use the marker correspondences that come with the FLAME model [51]. To get a more reliable evaluation scheme, we evaluate the benchmark of Feng et al. using the dense iterative closest point (ICP) technique from the Now challenge, see Table 3. On all metrics, our proposed method significantly improves the reconstruction accuracy. Note that

some methods are even performing worse than the mean face [51].

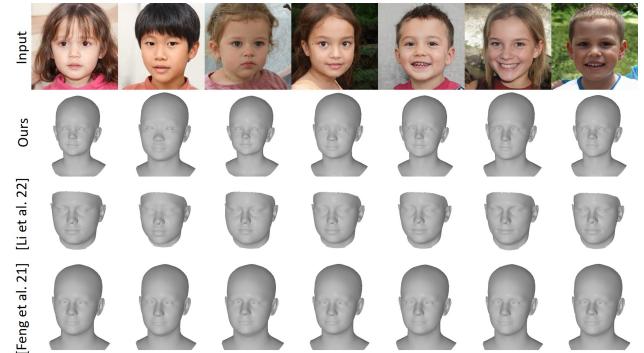


Figure 4. Current methods are not predicting metrical faces, which becomes visible when displaying them in a metrical space and not in their image spaces. To illustrate we render the prediction of the faces of toddlers in a common metrical space using the same projection. State-of-the-art approaches trained in a self-supervised fashion like DECA [25] or weakly-supervised like FOCUS [50] scale the face of an adult to fit the observation in the image space, thus, the prediction in 3D is non-metrical. In contrast, our reconstruction method is able to recover the physiognomy of the toddlers. Input images were generated by StyleGan2 [42].

Qualitative Results. In Figure 3, we show qualitative results to analyze the stability of the face shape prediction of a subject across different expressions, head rotation, occlusions, or perspective distortion. As can be seen our method is more persistent compared to others, especially, in comparison to Deng et al. [20] where shape predictions vary the most. Figure 4 depicts the challenging scenario of reconstructing toddlers from single images. Instead of predicting a small face for a child, the state of the art methods are pre-

dicting faces of adults. In contrast, MICA predicts the shape of a child.

In Figure 8 reconstructions for randomly sampled identities from the VoxCeleb2 [16] dataset are shown. Some of the baselines, especially, RingNet [64], exhibits strong bias towards the mean human face. In contrast, our method is able to not only predict better overall shape but also to reconstruct challenging regions like nose or chin, even though the training dataset contains a much smaller identity and ethnicity pool. Note that while the reconstructions of the baseline methods look good under the projection, they are not metric as shown in Tables 2 and 4.

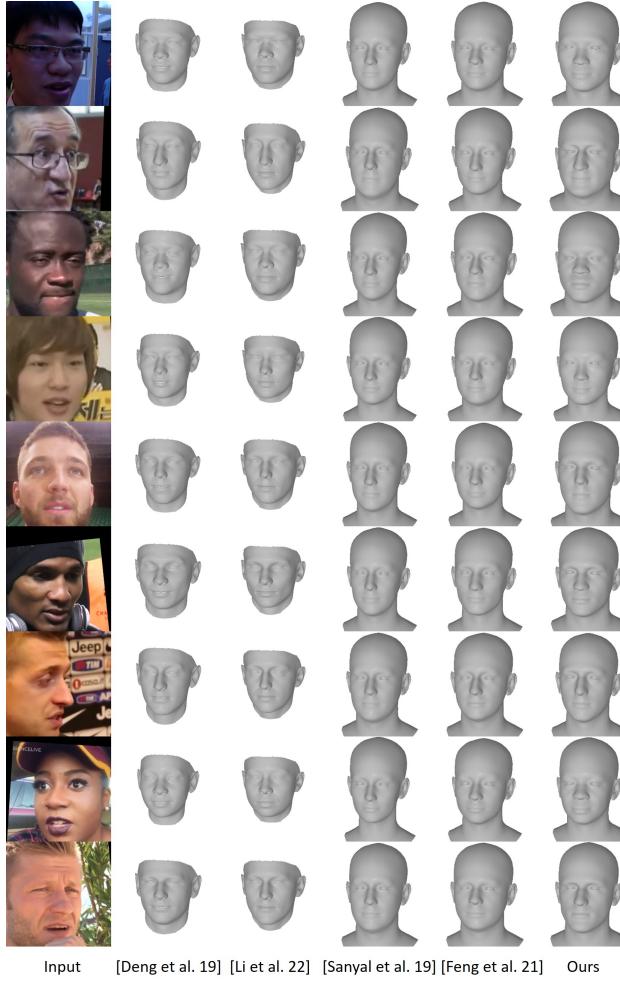


Figure 5. Qualitative comparison on randomly sampled images from the VoxCeleb2 [16] dataset. Our method is able to capture face shape with intricate details like nose and chin, while being metrical plausible (c.f., Tables 2 and 4).

6.2. Face Tracking

Our metrical face shape prediction can be used to initialize facial expression tracking. In contrast to methods

like [20, 25], our method uses a perspective camera model, which allows us to predict a depth. In Figure 10, we show a sample sequence from [81] with an according error plot (more results in the Appendix B).

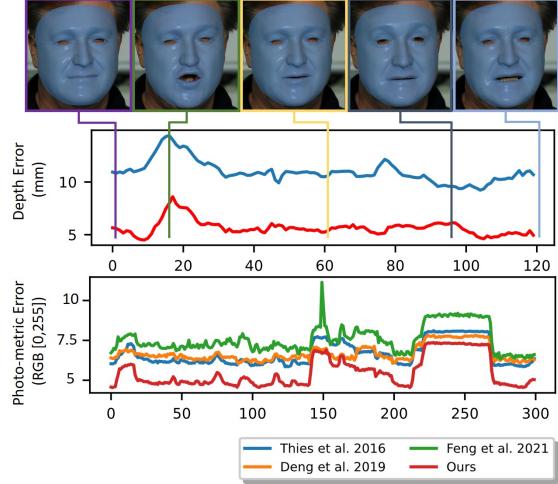


Figure 6. Evaluation of the tracking error on the 'Volker' sequence of [81]. The RMSE depth error is computed based on the reference depth maps which have been reconstructed using a stereo system. The photo-metric error is computed based on an RMSE error metric assuming RGB in $[0, 255]$ ³.

The error plots shown in Figure 10 contain the metrical reconstruction error of the depth (RMSE). It is based on the reference depth information of the sequence, which has been reconstructed from a passive stereo system. We also evaluate the dense photometric error (RMSE), which can be computed for [20, 25] too. In comparison to the method Face2Face [75] which also uses a perspective camera model (11.0mm mean RMSE depth error), our metrical face shape estimation improves the tracking quality significantly (5.7mm mean RMSE).

6.3. Limitations

Our method is not designed to predict shape and expressions in one forward pass, instead, we reconstruct the expression separately using an optimization-based tracking method. However, this optimization-based tracking leads to temporally coherent results, as can be seen in the suppl. video. In contrast to DECA [25] or Deng et al. [20], the focus of our method is the reconstruction of a metrical 3D model, reconstructing high-frequent detail on top of our prediction is an interesting future direction. Our method fails, when the used face detector [66] does not recognize a face in the input.

7. Discussion & Conclusion

In this paper, we analyzed various face reconstruction methods and their performance on metrical face shape re-

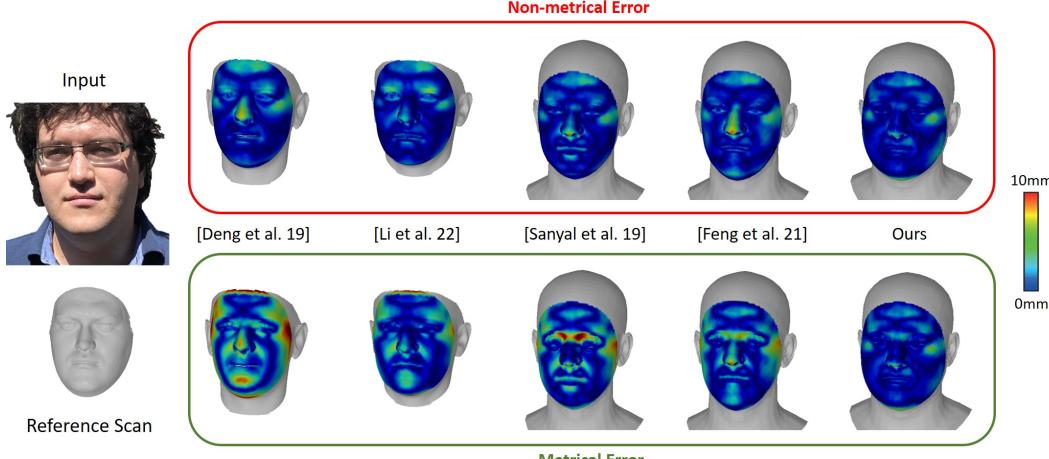


Figure 7. Established evaluation benchmarks like [27, 64] are based on a non-metrical error metric (top-row). We propose a new evaluation protocol which measures reconstruction errors in a metrical space (bottom row) (c.f. Table 2). Image from the NoW [64] validation set.

construction. A metrical reconstruction is key for any application that requires the measurement of distances and dimensions. It is essential for the composition of reconstructed humans and scenes where objects of known size are in, thus, it is especially important for virtual reality and augmented reality applications. However, we show that recent methods and evaluation schemes are not designed for this task. While the established benchmarks report numbers in millimeters, they are computed with an optimal scale to align the prediction and the reference. We strongly argue against this practice, since it is misleading and the errors are not absolute metrical measurements. To this end, we propose a simple, yet fundamental adjustment of the benchmarks to enable metrical evaluations. Specifically, we remove the optimal scaling, and only allow rigid alignment of the prediction with the reference shape.

As a stepping stone towards metrical reconstructions, we unified existing small- and medium-scale datasets of paired 2D/3D data. This allows us to establish 3D supervised losses in our novel shape prediction framework. While our data collection is still comparably small (around 2k identities), we designed MICA that uses features from a face recognition network pretrained on a large-scale 2D image dataset to generalize to in-the-wild image data. We validated our approach in several experiments and show state-of-the-art results on our newly introduced metrical benchmarks as well as on the established scale-invariant benchmarks. We hope that this work inspires researchers to concentrate on metrical face reconstruction, and that our discussion clarifies the misunderstandings that can occur with the established non-metrical evaluation benchmarks.

Acknowledgement. We thank Haiwen Feng for support with NoW and Stirling evaluations, and Chunlu Li for providing FOCUS results. The authors thank the International Max Planck Re-

search School for Intelligent Systems (IMPRS-IS) for supporting Wojciech Zienonka.

Disclosure. While TB is part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI.

References

- [1] Victoria Fernandez Abrevaya, Adnane Boukhayma, Philip H.S. Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4978–4988, 2020. 3
- [2] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Fu Ying. Partial fc: Training 10 million identities on a single machine. In *Arxiv 2010.05222*, 2020. 4
- [3] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2D/3D hybrid face dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU ’11, page 79–80, New York, NY, USA, 2011. Association for Computing Machinery. 4, 5
- [4] Anil Bas and William A. P. Smith. What does 2D geometric information really tell us about 3D face shape? *International Journal of Computer Vision (IJCV)*, 127(10):1455–1473, 2019. 1
- [5] 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. 5
- [6] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *EUROGRAPHICS (EG)*, volume 22, pages 641–650, 2003. 2
- [7] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. *Computer Graphics Forum*, 23(3):669–676, 2004. 2

- [8] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. [2](#), [3](#)
- [9] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. [5](#)
- [10] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. 2021. [15](#)
- [11] Chen Cao, Yanlin Weng, Shun Zhou, Yiyi Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *Transactions on Visualization and Computer Graphics*, 20:413–425, 01 2013. [2](#), [5](#)
- [12] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5799–5809, 2021. [15](#), [16](#)
- [13] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 122–129, 2018. [3](#)
- [14] Bindita Chaudhuri, Noranart Vesdapunt, Linda Shapiro, and Baoyuan Wang. Personalized face modeling for improved face reconstruction and motion retargeting, 2020. [3](#)
- [15] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9429–9439, 2019. [3](#)
- [16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In "INTER-SPEECH", 2018. [8](#), [14](#)
- [17] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facts valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 International Conference on Computer Vision*, pages 2296–2303, 2011. [5](#), [15](#)
- [18] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision (IJCV)*, 128(2):547–571, 2019. [2](#), [5](#), [15](#)
- [19] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision (ECCV)*, pages 741–757, 2020. [2](#), [3](#), [4](#), [14](#), [16](#)
- [20] 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 *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2019. [1](#), [3](#), [6](#), [7](#), [8](#), [16](#)
- [21] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *International Conference on Computer Vision (ICCV)*, pages 12819–12829, 2021. [6](#)
- [22] Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks, 2017. [3](#)
- [23] 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. *Transactions on Graphics (TOG)*, 39(5), 2020. [2](#)
- [24] Haven Feng and Timo Bolkart. Photometric FLAME fitting, 2020. [4](#)
- [25] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(8), 2021. [1](#), [3](#), [6](#), [7](#), [8](#), [14](#), [16](#)
- [26] 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 Conference on Computer Vision (ECCV)*, pages 534–551, 2018. [3](#), [6](#)
- [27] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter J. B. Hancock, Xiaojun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3D reconstruction from 2D face images in the wild. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 780–786, 2018. [2](#), [6](#), [7](#), [9](#)
- [28] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter J. B. Hancock, Xiaojun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. *CoRR*, abs/1803.05536, 2018. [2](#), [5](#), [6](#), [15](#), [16](#)
- [29] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thor-maehlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4217–4224, 2014. [2](#)
- [30] Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. VDub - modifying face video of actors for plausible visual alignment to a dubbed audio track. In *EUROGRAPHICS (EG)*, pages 193–204, 2015. [2](#)
- [31] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, volume 32, pages 158:1–158:10, November 2013. [15](#)
- [32] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [3](#)
- [33] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. [3](#)

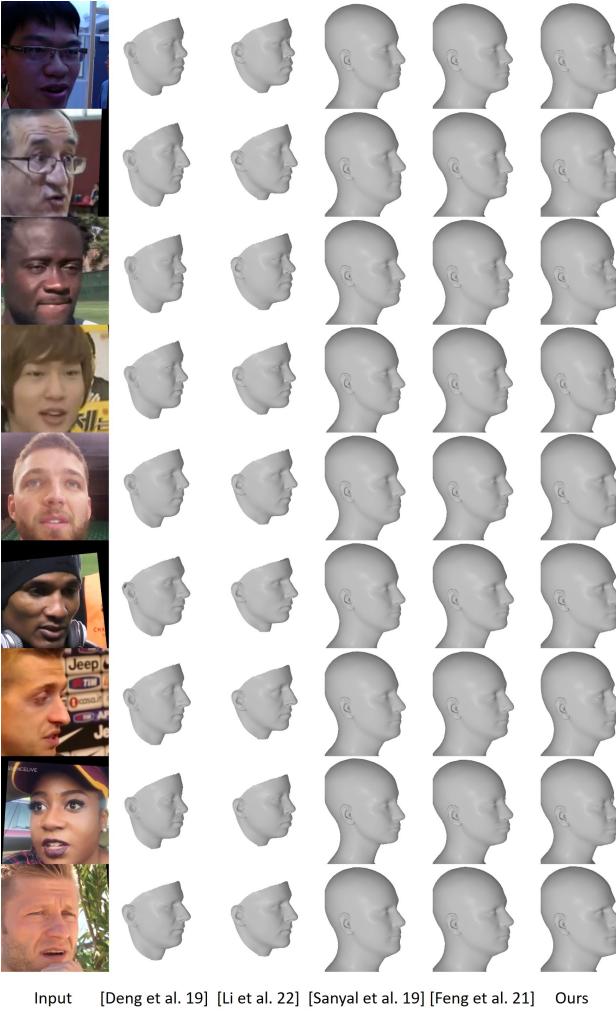
- [34] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression, 2018. 3
- [35] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos, 2021. 3
- [36] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time, 2020. 5
- [37] 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 (ECCV)*, 2020. 6, 7
- [38] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild, 2017. 3
- [39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 16
- [40] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.*, 36(6), nov 2017. 3
- [41] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression, 2017. 3
- [42] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 1, 7
- [43] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus, 2019. 5
- [44] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2, 3
- [45] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inverse-FaceNet: Deep monocular inverse face rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3
- [46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 5
- [47] Tatsuro Koizumi and William A. P. Smith. "look ma, no landmarks!" - unsupervised, model-based dense face alignment. In *European Conference on Computer Vision (ECCV)*, volume 12347, pages 690–706, 2020. 6
- [48] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction" in-the-wild". In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 760–769, 2020. 3
- [49] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P. Zafeiriou. AvatarMe++: Facial shape and BRDF inference with photorealistic rendering-aware GANs. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2021. 3
- [50] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. *CoRR*, abs/2106.09614, 2021. 6, 7
- [51] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 2, 4, 5, 6, 7, 15, 16
- [52] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *International Conference on Computer Vision (ICCV)*, Oct 2019. 5
- [53] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 4
- [54] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 2
- [55] Araceli Morales, Gemma Piella, and Federico M. Sukno. Survey on 3d face reconstruction from uncalibrated images, 2021. 2
- [56] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. Pagan: Real-time avatars using dynamic textures. *ACM Trans. Graph.*, 37(6), dec 2018. 3
- [57] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *International conference on advanced video and signal based surveillance*, pages 296–301, 2009. 2
- [58] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 947–954 vol. 1, 2005. 5
- [59] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '01*, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery. 5
- [60] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 5
- [61] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data, 2016. 3
- [62] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image, 2017. 3

- [63] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks, 2016. 3
- [64] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 6, 7, 8, 9, 14
- [65] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. 3, 14, 16
- [66] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended light-face: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 8
- [67] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision (ECCV)*, volume 12360, pages 53–70, 2020. 6
- [68] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 16
- [69] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 3
- [70] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3
- [71] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 3
- [72] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *European Conference on Computer Vision (ECCV)*, 2020. 2, 3
- [73] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 3
- [74] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *Transactions on Graphics (TOG)*, 34(6), 2015. 2, 3
- [75] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 2, 3, 5, 8
- [76] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Niessner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics*, 37(4):1–13, Aug 2018. 3
- [77] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1599–1608, 2017. 3, 6
- [78] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3
- [79] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019. 3
- [80] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. Joint 3D face reconstruction and dense face alignment from a single image with 2D-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359*, 2019. 3
- [81] Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012)*, volume 31, pages 187:1–187:11, November 2012. 8
- [82] Huawei Wei, Shuang Liang, and Yichen Wei. 3d dense face alignment via graph convolution networks, 2019. 3
- [83] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Re-altime performance-based facial animation. In *Transactions on Graphics (TOG)*, volume 30, 2011. 2
- [84] Thibaut Weise, Hao Li, Luc J. Van Gool, and Mark Pauly. Face/Off: live facial puppetry. In *SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 7–16, 2009. 2
- [85] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.*, 37(4), jul 2018. 3
- [86] 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)*, June 2020. 5, 15
- [87] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M.J. Rosato. A 3d facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition (FG2006)*, pages 211–216, 2006. 5, 15
- [88] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew

- Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Li-jun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, 2016. 5
- [89] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *arXiv preprint arXiv:2111.01082*, 2021. 2, 5, 15
- [90] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3D solution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. 3
- [91] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum (Eurographics State of the Art Reports)*, 37(2), 2018. 2

A. Additional Results

Our 3DMM-based shape estimation method presented in the main paper has two key components, (1) the encoder based on a face recognition network with a mapping network and (2) the 3DMM-based geometry decoder. In this section, we present several ablation studies w.r.t. those modules and the used training data. All experiments were done with the same optimizer and hyper-parameter configuration as the main method except where stated otherwise. The Stirling dataset was excluded from all the ablation experiments.

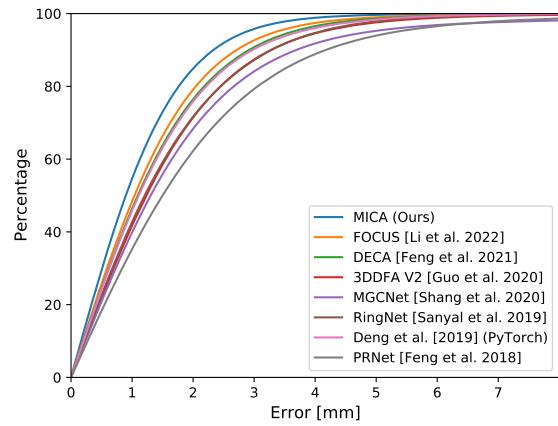


Input [Deng et al. 19] [Li et al. 22] [Sanyal et al. 19] [Feng et al. 21] Ours

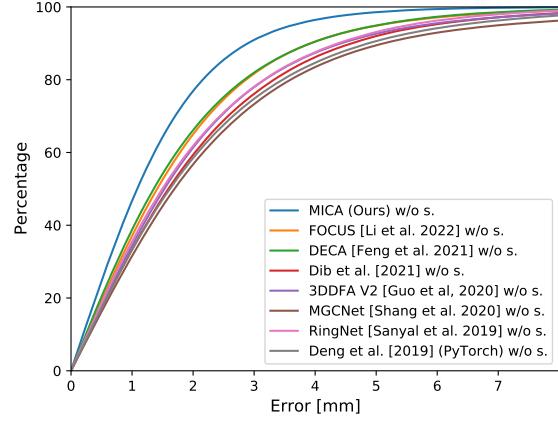
Figure 8. Qualitative comparison on randomly sampled images from the VoxCeleb2 [16] dataset for side views.

The difference between our and the state-of-the-art methods w.r.t. their reconstruction quality gets well visible in the cumulative error plots in Figure 9. Moreover, Figure 8 depicts side views of the reconstructions, which gives a better look at the shape quality.

Encoder Ablation Studies. Exploiting generalized facial features from a face recognition network is a key component of our method to predict geometry from in-the-wild 2D data. However, completely refining the latent space of the face recognition network is not possible with our medium-size dataset, thus, we can only retrain selected layers to maintain generalizability. In Table 5, we compare the performance of the two face recognition methods ArcFace [19] and FaceNet [65]. Overall, the pretrained ArcFace outperforms the pretrained FaceNet in terms of reconstruction quality in our shape estimation architecture. To further improve the results of ArcFace, we refine the last ResNet layer of ArcFace.



(a) Cumulative error plot for NoW Challenge (non-metrical)



(b) Cumulative error plot for NoW-Metric Challenge

Figure 9. Cumulative plots for the (a) NoW [64] and (b) NoW-Metric (w/o scale) challenges. We refer to the main paper for the detailed statistics.

Similarly, we conducted experiments on fine-tuning DECA [25] using our medium-scale dataset and our reconstruction loss based on an ℓ_1 error metric. We trained the

network on the same datasets like ArcFace for around 500 epochs. The fine-tuning of partial layers or entire pipeline leads to huge overfitting of the training data with significantly worse reconstructions on the test dataset (see Table 5). In contrast, the partial fine-tuning of ArcFace in our approach gives the lowest mean reconstruction error of 1.35mm. It shows that we can effectively use the generalized features from the ArcFace network for the task of metrical face reconstruction.

Decoder Ablation Studies. The decoder is defined by the 3DMM FLAME [51]. For our experiments in the main paper, we used 300 eigenvectors of the PCA basis. In Table 6, we present an ablation study on the number of used eigenvectors (i.e., the size of the latent geometry code z). As can be seen, exploiting the full linear space of FLAME leads to the best performance.

Decoder - #PC	Median		Mean (mm)		Std	
	LQ	HQ	LQ	HQ	LQ	HQ
50	1.19	1.12	1.50	1.41	1.33	1.23
100	1.15	1.10	1.45	1.39	1.28	1.22
200	1.15	1.06	1.47	1.36	1.31	1.20
300	1.15	1.06	1.46	1.35	1.30	1.20

Table 6. Evaluation of the influence of the number of principle components (PCs) used for the shape decoder. Based on the Stirling dataset [28] using NoW protocol (metrical).

Dataset Ablation Studies. As described in the main paper, we used several datasets to construct our training set. We perform a leave-one-out analysis in Table 7 on the Stirling dataset. The LYHM dataset contains 1211 subjects and, thus, has the most significant influence on the training.

In addition to the datasets listed in the main paper, we also processed FaceScape [86, 89]. While FaceScape is a large-scale dataset, it has been recorded within an uncalibrated setup, thus, being in a none metrical scale which is introducing a bias in our prediction.

B. Studies on the Facial Expression Tracking

In the supplemental video, we show several tracking results which demonstrate that our proposed technique is temporally stable. In Figure 10, we plot the photo-metric error of the three sequences from Garrido et al. [31] with different subjects and motion signatures. As can be seen, our method results in the lowest photo-metric error in terms of a masked RMSE metric on the colors.

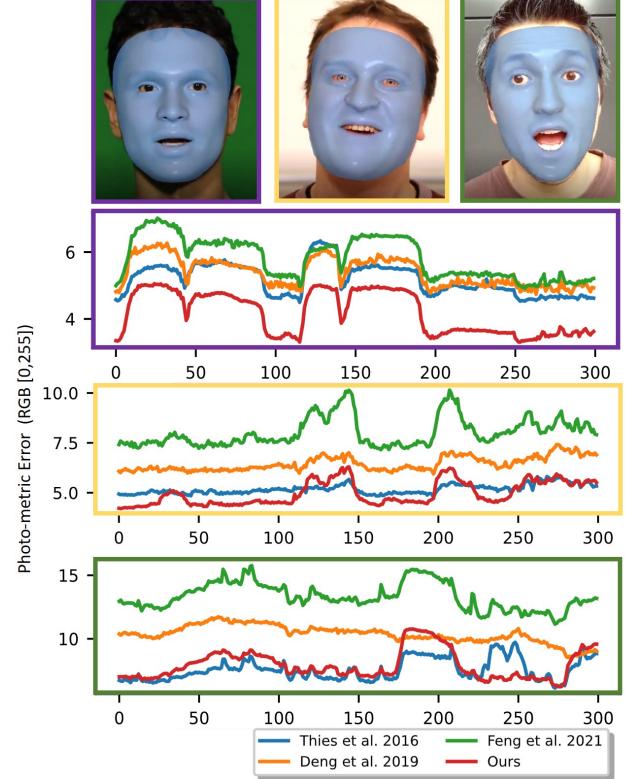


Figure 10. Photo-metric error on the three sequences from Garrido et al. [31] shown in the supplemental video. The photo-metric error is computed based on an RMSE error metric assuming RGB in $[0, 255]$ ³.

Dataset	Median		Mean (mm)		Std	
	LQ	HQ	LQ	HQ	LQ	HQ
w/o LYHM [18]	1.18	1.12	1.51	1.43	1.35	1.27
w/o FRGC [87]	1.15	1.06	1.47	1.36	1.33	1.23
w/o BP4D+ [87]	1.14	1.09	1.46	1.38	1.30	1.21
w/o BU3DFE [87]	1.14	1.08	1.45	1.37	1.29	1.21
w/o D3DFACS [17]	1.13	1.06	1.43	1.35	1.28	1.19
w/o Face Warehouse [17]	1.13	1.07	1.43	1.36	1.27	1.20
All	1.15	1.06	1.46	1.35	1.30	1.20

Table 7. To analyze the contribution of a single dataset, we perform a leave-one-out analysis. We report the reconstruction quality for images from Stirling dataset [28] (where we exclude Stirling from training). As can be seen, LYHM [18] has the highest influence on the reconstruction quality, leaving it out leads to an increase of the mean error for HQ images from 1.35mm to 1.43mm and for LQ images from 1.46mm to 1.51mm on the Stirling dataset [28].

C. Model-free Decoder

Inspired by pi-GAN [12] and Dynamic Surface Function Networks [10], we also evaluated a coordinated-based multi

Encoder	Median		Mean (mm)		Std	
	LQ	HQ	LQ	HQ	LQ	HQ
DECA [25] (frozen)	1.32	1.22	1.71	1.58	1.54	1.42
DECA [25] (fully trainable)	1.54	1.42	1.96	1.82	1.71	1.61
DECA [25] ($L_3 - L_4$ trainable)	1.55	1.43	1.97	1.83	1.71	1.62
DECA [25] (L_4 trainable)	1.55	1.49	1.97	1.83	1.71	1.61
Ours – FaceNet [65] (frozen)	1.37	1.29	1.75	1.65	1.56	1.47
Ours – ArcFace [19] (frozen)	1.25	1.18	1.60	1.52	1.43	1.37
Ours – ArcFace [19] (fully trainable)	1.18	1.11	1.52	1.42	1.38	1.27
Ours – ArcFace [19] ($L_2 - L_3 - L_4$ trainable)	1.22	1.12	1.56	1.43	1.39	1.27
Ours – ArcFace [19] ($L_3 - L_4$ trainable)	1.17	1.10	1.51	1.40	1.37	1.25
Ours – ArcFace [19] (L_4 trainable)	1.15	1.06	1.46	1.35	1.30	1.20

Table 5. Ablation study w.r.t. our face encoding network based on Stirling dataset [28] using our metrical evaluation scheme. As a comparison, we also show the results for DECA [25] fine-tuned on our dataset. The respective ResNet [39] networks were refined in different configurations; $\{L_3, L_4\}$ denotes the set of selected trainable layers from $\{L_1, \dots, L_4\}$. Each layer is composed of several ResNet blocks, specifically, ArcFace uses $\{3, 13, 30, 3\}$ and DECA $\{3, 4, 6, 3\}$ ResNet blocks for the respective layers.

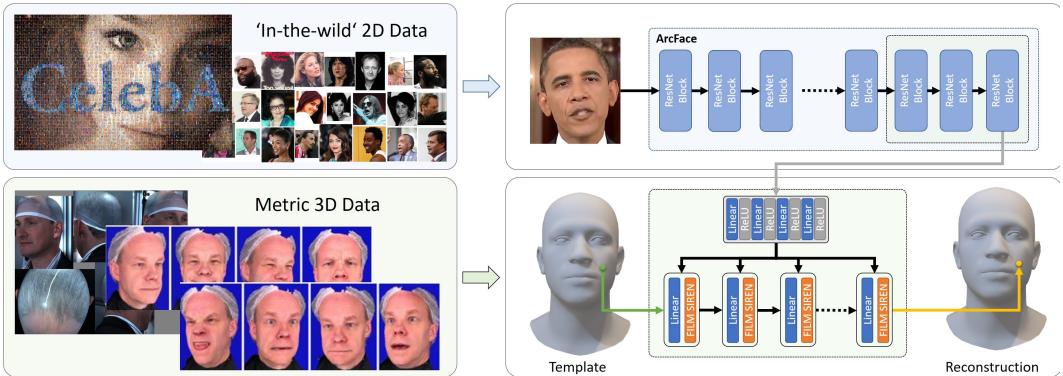


Figure 11. Overview of a model-free decoder. The model-free decoder is based on a Siren architecture [68] using FiLM conditionings [12]. In contrast to the FLAME-based decoder, this model-free decoder can be trained in conjunction to the encoder only based on the dataset with the paired 2D/3D data which is smaller than the dataset of 3D scans used for constructing the FLAME model.

Stirling (NoW Protocol)	Non-Metrical						Metrical (mm)					
	Median		Mean		Std		Median		Mean		Std	
	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ	HQ
Deng et al. [20] (PyTorch)	1.12	0.99	1.44	1.27	1.31	1.15	1.47	1.31	1.93	1.71	1.77	1.57
DECA [25]	1.09	1.03	1.39	1.32	1.26	1.18	1.32	1.22	1.71	1.58	1.54	1.42
Ours (SIREN)	1.01	0.94	1.28	1.19	1.15	1.06	1.20	1.09	1.53	1.39	1.35	1.23
Ours (FLAME)	0.96	0.92	1.22	1.16	1.11	1.04	1.15	1.06	1.46	1.35	1.30	1.20

Table 8. Quantitative evaluation of the face shape estimation using Stirling dataset [28] and NoW protocol (metrical).

layer perceptron (MLP) with sinusoidal activation functions (SIREN [68]) to represent the geometry of a face (see Figure 11). This architecture can be trained solely on the data of our unified dataset without requiring any 3DMM model. The network and its sinusoidal activation functions are controlled by a mapping network \mathcal{M}' to represent different

faces. The mapping network takes the identity code z as input and predicts the frequencies and phase shifts of the sinusoidal activation layers. The SIREN network \mathcal{S} is evaluated at the FLAME [51] template mesh vertices $A \in \mathbb{R}^{3N}$ and $N = 5023$ to leverage the correspondences of the 3D

training data.

$$\mathcal{G}_{SIREN}(\mathbf{z}) = \mathcal{S}(\mathbf{A} \mid \mathcal{M}'(\mathbf{z})).$$

Since this model does not rely on the PCA basis of the FLAME model, it can predict meshes outside the FLAME face space. In comparison to the 3DMM-based model presented in the main paper, this model-free approach performs on par on the different benchmarks (see Table 8). A benefit of the model-free decoder is that it can be trained solely on our dataset of paired 2D/3D data which is significantly smaller than the dataset of 3D scans used for the construction of the FLAME model (2.3k (our dataset) versus 4k subjects used for FLAME).

A drawback of this Siren-based approach is its runtime and complexity (3DMM only has a single linear layer for representing shape variations). The used SIREN network is a more compact representation using 8 hidden layers and 256 feature size with total 1976327 parameters, while the 3DMM has a linear layer with $(300 + 1) * 5023 * 3 = 4535769$ parameters.