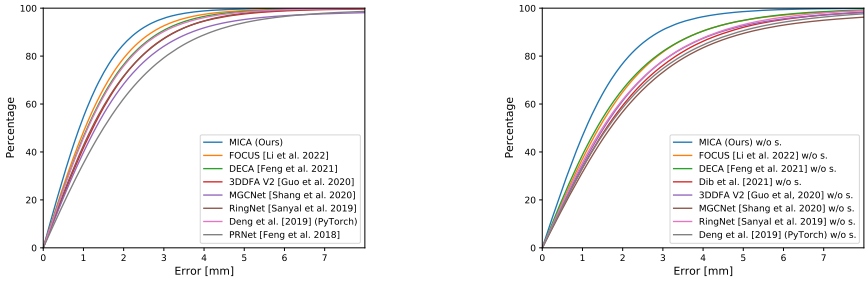
Towards Metrical Reconstruction of Human Faces –Supplemental Document–

Wojciech Zielonka, Timo Bolkart, and Justus Thies

Max Planck Institute for Intelligent Systems, Tübingen



(a) NoW Challenge (non-metrical) (b) NoW-Metric Challenge

Fig. 1: Cumulative plots for the (a) NoW [13] and (b) NoW-Metric (w/o scale) challenges. We refer to the main paper for the detailed statistics.

Abstract. In this supplemental document, we demonstrate the robustness of our proposed method in additional qualitative and quantitative experiments. The cumulative error plots from the NoW challenge presented in the main paper are also included in this document. Moreover, we present a justification of our architecture selection which is tailored for our unified dataset. Further, we discuss an alternative model-free estimation approach that does not rely on a 3DMM decoder and can be learned solely on our unified data.

1 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. 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 1. Moreover, Figure 2 depicts side views of the reconstructions, which gives a better look at the shape quality. 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.

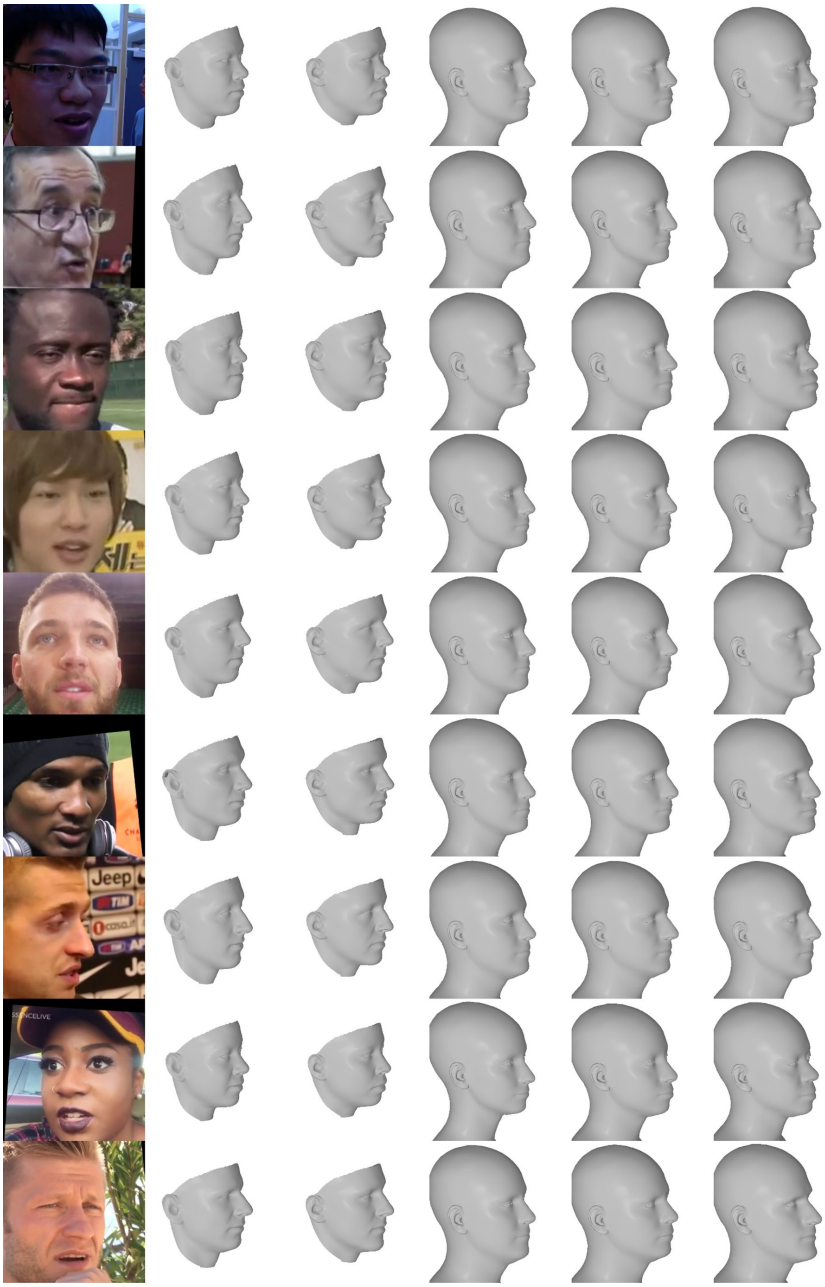


Fig. 2: Qualitative comparison on randomly sampled images from the VoxCeleb2 [3] dataset for side views.

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 1, we compare the performance of the two face recognition methods ArcFace [6] and FaceNet [14]. 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. Similarly, we conducted experiments on fine-tuning DECA [8] 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 1). 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.

Table 1: Ablation study w.r.t. our face encoding network based on Stirling dataset [9] using our metrical evaluation scheme. As a comparison, we also show the results for DECA [8] fine-tuned on our dataset. The respective ResNet [11] networks were refined in different configurations; $\{L3, L4\}$ denotes the set of selected trainable layers from $\{L1, ..., L4\}$. 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.

Encoder	Median		Mean (mm)		Std	
	LQ	HQ	LQ	HQ	LQ	HQ
DECA [8] (frozen)	1.32	1.22	1.71	1.58	1.54	1.42
DECA [8] (fully trainable)	1.54	1.42	1.96	1.82	1.71	1.61
DECA [8] ($L3 - L4$ trainable)	1.55	1.43	1.97	1.83	1.71	1.62
DECA [8] ($L4$ trainable)	1.55	1.49	1.97	1.83	1.71	1.61
Ours – FaceNet [14] (frozen)	1.37	1.29	1.75	1.65	1.56	1.47
Ours – ArcFace [6] (frozen)	1.25	1.18	1.60	1.52	1.43	1.37
Ours – ArcFace [6] (fully trainable)	1.18	1.11	1.52	1.42	1.38	1.27
Ours – ArcFace [6] ($L2 - L3 - L4$ trainable)	1.22	1.12	1.56	1.43	1.39	1.27
Ours – ArcFace [6] ($L3 - L4$ trainable)	1.17	1.10	1.51	1.40	1.37	1.25
Ours – ArcFace [6] ($L4$ trainable)	1.15	1.06	1.46	1.35	1.30	1.20

Decoder Ablation Studies. The decoder is defined by the 3DMM FLAME [12]. For our experiments in the main paper, we used 300 eigenvectors of the PCA basis. In Table 2, we present an ablation study on the number of used eigenvectors (i.e., the size of the latent geometry code $\boldsymbol{z}$). As can be seen, exploiting the full linear space of FLAME leads to the best performance.

Table 2: Evaluation of the influence of the number of principle components (PCs) used for the shape decoder (Stirling dataset [9] with NoW protocol (metrical)).

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

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 3 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 [18, 20]. 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.

Table 3: 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 [9] (where we exclude Stirling from training). As can be seen, LYHM [5] 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 [9].

Dataset	Median		Mean (mm)		Std	
	LQ	HQ	LQ	HQ	LQ	HQ
w/o LYHM [5]	1.18	1.12	1.51	1.43	1.35	1.27
w/o FRGC [19]	1.15	1.06	1.47	1.36	1.33	1.23
w/o BP4D+ [19]	1.14	1.09	1.46	1.38	1.30	1.21
w/o BU3DFE [19]	1.14	1.08	1.45	1.37	1.29	1.21
w/o D3DFACS [4]	1.13	1.06	1.43	1.35	1.28	1.19
w/o Face Warehouse [4]	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

2 Studies on the Facial Expression Tracking

Our metrical face shape prediction can be used to initialize facial expression tracking. In contrast to methods like [8], our method uses a perspective camera model, which allows us to predict a depth. In Figure 4, we show a sample sequence from [17] with an according depth and photo-metric error plot.

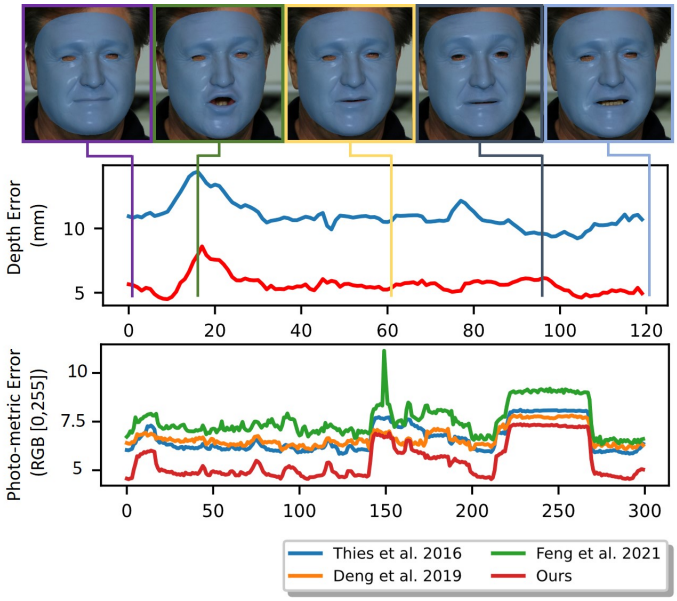


Fig. 3: Evaluation of the tracking error on the 'Volker' sequence of [17]. The RMSE depth error is computed based on the reference depth maps which have been reconstructed using a stereo system. The photo-metric error is computed based on an RMSE error metric assuming RGB in $[0, 255]^3$.

As can be seen, our method results in the lowest photo-metric error in terms of a masked RMSE metric on the colors. The error plots shown in Figure 3 contain the metric 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 [7, 8] too. In comparison to the method Face2Face [16] 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). In the supplemental video, we show several tracking results which demonstrate that our proposed technique is temporally stable.

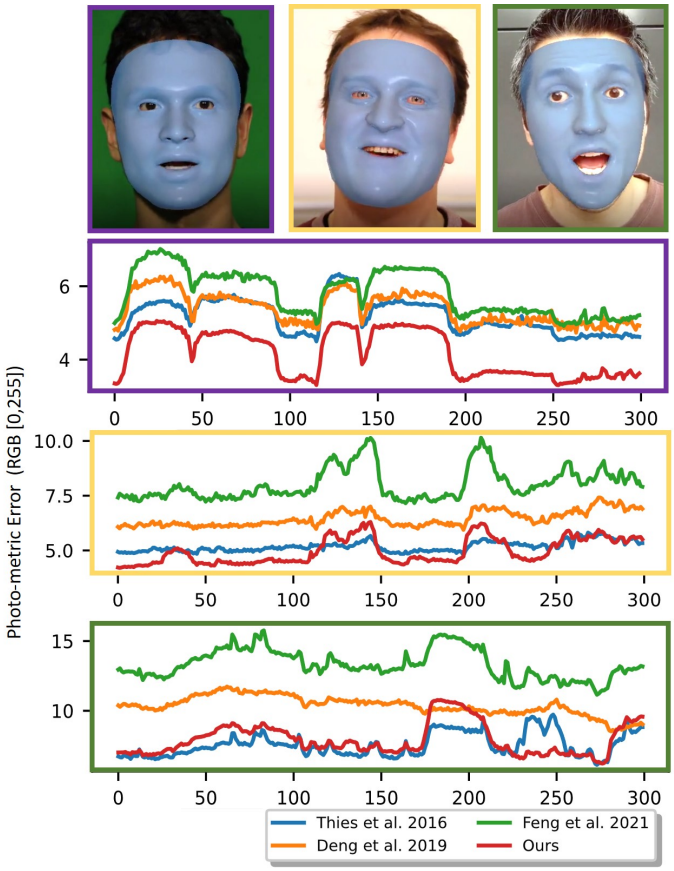


Fig. 4: Photo-metric error on the three sequences from Garrido et al. [10] shown in the supplemental video.The photo-metric error is computed based on an RMSE error metric assuming RGB in $[0, 255]^3$.

3 Model-free Decoder

Inspired by pi-GAN [2] and Dynamic Surface Function Networks [1], we also evaluated a coordinated-based multi layer perceptron (MLP) with sinusoidal activation functions (SIREN [15]) to represent the geometry of a face (see Figure 5). 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 [12] template mesh vertices $\boldsymbol{A} \in \mathbb{R}^{3N}$ and $N = 5023$ to leverage the correspondences of the 3D training data.

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

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 4). 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).

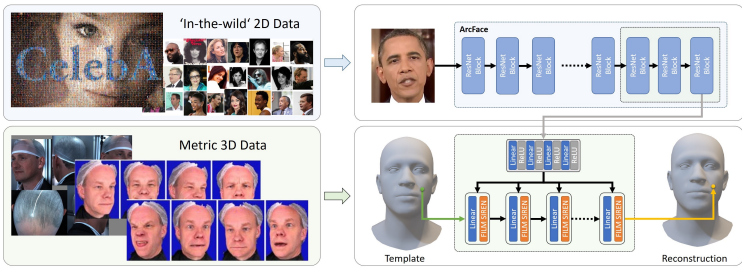


Fig. 5: Overview of a model-free decoder. The model-free decoder is based on a Siren architecture [15] using FiLM conditionings [2]. In contrast to the FLAME-based decoder, this model-free decoder can be trained in conjunction to the encoder only based on the dataset with the paired 2D/3D data which is smaller than the dataset of 3D scans used for constructing the FLAME model.

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.

Table 4: Quantitative evaluation of the face shape estimation using Stirling dataset [9] and NoW protocol (metrical).

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. [7] (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 [8]	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**

Bibliography

[1] Burov, A., Nießner, M., Thies, J.: Dynamic surface function networks for clothed human bodies (2021) 6

[2] Chan, E., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5799–5809 (2021) 6, 7

[3] Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep speaker recognition. In: "INTERSPEECH" (2018) 2

[4] Cosker, D., Krumhuber, E., Hilton, A.: A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In: 2011 International Conference on Computer Vision. pp. 2296–2303 (2011). https://doi.org/10.1109/ICCV.2011.6126510 4

[5] Dai, H., Pears, N., Smith, W., Duncan, C.: Statistical modeling of craniofacial shape and texture. International Journal of Computer Vision (IJCV) **128**(2), 547–571 (2019). https://doi.org/10.1007/s11263-019-01260-7 4

[6] Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: European Conference on Computer Vision (ECCV). pp. 741–757 (2020) 3

[7] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W) (2019) 5, 7

[8] Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. Transactions on Graphics, (Proc. SIGGRAPH) **40**(8) (2021) 3, 5, 7

[9] Feng, Z., Huber, P., Kittler, J., Hancock, P.J.B., Wu, X., Zhao, Q., Koppen, P., Rätsch, M.: Evaluation of dense 3d reconstruction from 2d face images in the wild. CoRR **abs/1803.05536** (2018), http://arxiv.org/abs/1803.05536 3, 4, 7

[10] Garrido, P., Valgaerts, L., Wu, C., Theobalt, C.: Reconstructing detailed dynamic face geometry from monocular video. In: ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013). vol. 32, pp. 158:1–158:10 (November 2013). https://doi.org/10.1145/2508363.2508380, http://doi.acm.org/10.1145/2508363.2508380 6

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385 3

[12] Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813 4, 6

[13] Sanyal, S., Bolkart, T., Feng, H., Black, M.: 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

[14] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2015). https://doi.org/10.1109/cvpr.2015.7298682, http://dx.doi.org/10.1109/CVPR.2015.7298682 3

[15] Sitzmann, V., Martel, J.N., Bergman, A.W., Lindell, D.B., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 6, 7

[16] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: Real-time face capture and reenactment of RGB videos. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2387–2395 (2016) 5

[17] Valgaerts, L., Wu, C., Bruhn, A., Seidel, H.P., Theobalt, C.: Lightweight binocular facial performance capture under uncontrolled lighting. In: ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012). vol. 31, pp. 187:1–187:11 (November 2012). https://doi.org/10.1145/2366145.2366206, http://doi.acm.org/10.1145/2366145.2366206 5

[18] Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4

[19] Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3d facial expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06). pp. 211–216 (2006). https://doi.org/10.1109/FGR.2006.6 4

[20] Zhu, H., Yang, H., Guo, L., Zhang, Y., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. arXiv preprint arXiv:2111.01082 (2021) 4