Supplementary material: FitDiff: Robust monocular 3D facial shape and reflectance estimation using Diffusion Models

Stathis Galanakis[1,2] Alexandros Lattas[1] Stylianos Moschoglou[1] Stefanos Zafeiriou[1]

[1]Imperial College London
[2]HUAWEI Noah's Ark Lab

1. Future Work

In this work, we present the first-of-its-kind diffusion model conditioned on expressive facial embeddings, which is essentially a step towards a facial foundation model. Such models require vast datasets of labeled data [19], in our case paired facial images with geometry, reflectance and identity embeddings. These are immensely challenging to acquire in numbers, and hence we have to rely on a synthetic dataset and inherit its method's limitations [13]. Nevertheless, our method could be trivially extended to larger datasets of even scanned datasets (e.g. [24]), given their availability. Moreover, the "fitting" nature of our method, is limited by the ambiguity between scene illumination and skin tone, especially in single-image inference. To that end, the recent method of TRUST [10], could be incorporated into our diffusion model, as an additional conditioning mechanism, given however the availability of training data.

2. Implementation Details

A comprehensive overview of this training approach is presented in Fig. 1. We provide the essential information required to reproduce our method. The code-base for the brached multi-modal AutoEncoder is built on the public repository of the VQGAN AutoEncoder [9]. We made the following changes: A) The first downsampling layer of the encoder $\mathcal{E}$ and the last upsampling layer of the decoder $\mathcal{D}$ are branched, by making 3 copies of the respective layers. B) As proposed in FitMe [13], we use a branched discriminator, in the essence of having 2 copies of the main discriminator, except the last convolutional layer. The branch, dedicated for diffuse ($\mathbf{A}_D$) and specular ($\mathbf{A}_S$) albedos, gets a 6-channel input whereas the normals ($\mathbf{N}$) branch gets a 3-channel input.

On the other hand, the main training phase is built on the public repository of Latent Diffusion Models [19]. We modified the provided UNet code by turning it into a 1-D UNet network and replaced the attention-based conditional mechanism with SPADE layers [17]. The hyper-parameters for the Conditional UNet using SPADE layers are presented in Tab. 2 while it gets trained for 800 epochs.

f	$\lvert\mathcal{Z}\rvert$	Embed. dim
8	16384	1
z channels	**Channels**	**Channels mult.**
4	128	1,2,2,4
Res. Blocks	**Attention Res.**	**Batch Size**
2	32	16

Table 1. Hyper-parameters used during training the branched multi-modal AutoEncoder.

As mentioned in the main paper, the overall training loss under which FitDiff is trained, is the following:

$$\mathcal{L} = \mathcal{L}_{noise} + \mathcal{L}_{id} + \mathcal{L}_{per} + \mathcal{L}_{verts}$$

where $\mathcal{L}_{noise}$ is the noise prediction loss as defined in Section 3.3, $\mathcal{L}_{id}$ is the identity distance, $\mathcal{L}_{per}$ the identity perceptual loss and $\mathcal{L}_{verts}$ the shape loss.

Identity distance To supervise the identity similarity between the ground truth and the predicted facial avatars, we follow the methodology presented in [12, 13]. We employ a face recognition network [7] with n layers : $\mathcal{C}^n(\mathbf{I}) : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{512}$. The identity distance is estimated by computing the identity similarity between the feature emebeddings of the input image $\mathbf{I}$, and the estimated initial image $\bar{\mathbf{I}}_0$:

$$\mathcal{L}_{id} = 1 - \frac{\mathcal{C}^n(\bar{\mathbf{I}}_0)\cdot\mathcal{C}^n(\mathbf{I})}{\parallel \mathcal{C}^n(\bar{\mathbf{I}}_0)\parallel_2 \cdot \parallel \mathcal{C}^n(\mathbf{I})\parallel_2} \tag{1}$$

Identity perceptual loss To enforce perceptual consistency between the generated and ground truth avatars, we also penalize the discrepancy between the intermediate activation layers of the face recognition network $\mathcal{C}$. The identity

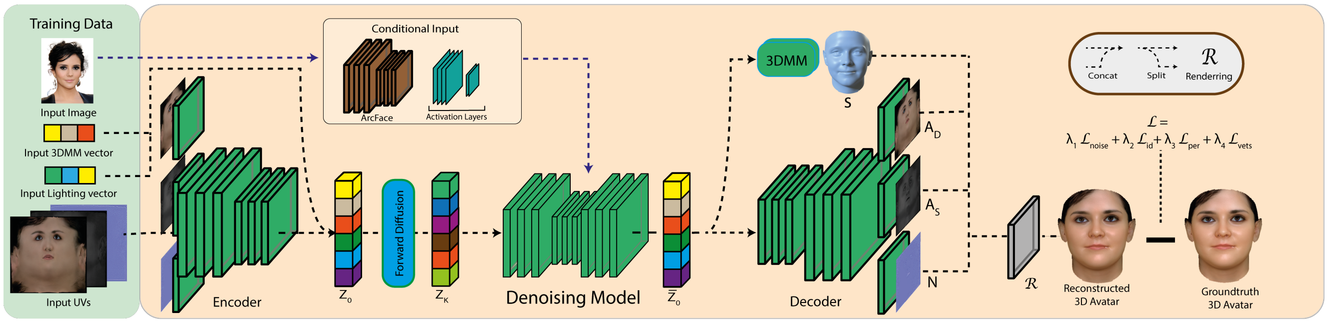


Figure 1. Overview of the main phase of our training scheme: At each training iteration, the facial reflectance maps are first projected into the latent space and subsequently concatenated to the latent vector $\mathbf{z}_0$, to which noise is introduced. After estimating the initial latent vector $\bar{\mathbf{z}}_0$ and rendering ($\mathcal{R}$) the estimated initial avatar, perceptual and face recognition losses are applied.

Diffusion steps 1000	**Noise Schedule** linear	**Input Channels** 1
Channels 192	**Cond. Dim** 1048	**SPADE dim.** 128
Channels mult 1,2,4,8	**Depth** 2	**Heads** 4
Heads Channels 32	**Batch size** 16	**LR** 3.2e-05

Table 2. Hyper-parameters of the main training phase.

perceptual loss is computed as:

$$\mathcal{L}_{per} = \sum_{j}^{n} \frac{\parallel \mathcal{C}^j(\bar{\mathbf{I}}_0) - \mathcal{C}^j(\mathbf{I}) \parallel_2}{H_{\mathcal{C}_j} \cdot W_{\mathcal{C}_j} \cdot C_{\mathcal{C}_j}} \tag{2}$$

where $H_{\mathcal{C}_j}$, $W_{\mathcal{C}_j}$, and $C_{\mathcal{C}_j}$ denote the height, width, and number of channels of the j-th activation map, respectively.

Shape loss The difference between the estimated facial shape $\bar{\mathbf{v}}_0$ and the ground truth facial shape $\mathbf{v}$ is calculated using the L1-norm:

$$\mathcal{L}_{verts} = \parallel \bar{\mathbf{v}}_0 - \mathbf{v} \parallel_1 \tag{3}$$

3. Guidance Algorithm

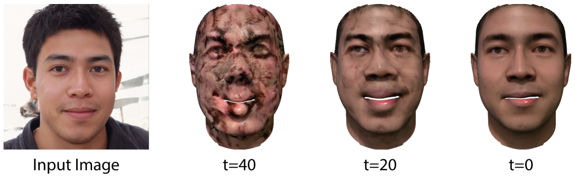


Figure 2. An example of the sampling process for $t = \{40, 20, 0\}$,

FitDiff is a diffusion-based architecture conditioned on an identity embedding vector. It accurately generates facial identities by incorporating an effective identity guidance method during the sampling phase. An example of this process is illustrated in Fig. 2. The proposed guidance method uses the guidance loss which is formulated as:

$$\mathcal{G} = \mathcal{G}_{id}^{cos} + \lambda_1 \mathcal{G}_{id}^{per} + \lambda_2 \mathcal{G}_{mse} + \lambda_3 \mathcal{G}_{lan} + \lambda_4 \mathcal{G}_{vgg} \tag{4}$$

The values of the used lambdas are $\lambda_1 = 50$, $\lambda_2 = 10$, $\lambda_3 = 200$, $\lambda_4 = 1$ where we use a gradient scale $s = 75$. We run our sampling method for $T = 50$ steps. When run on an NVIDIA Tesla V100-PCIE-32GB GPU, the diffusion sampling process takes about 54 seconds, timed comparable with other fitting methods like FitMe [13] and Relightify [16] which take about 50 seconds and about 1min respectively.

3.1. Sampling Guidance pseudo-code

Following Algorithm 1, we feed the input image $\mathbf{I}$ into $\mathcal{C}$, to extract the latent identity embedding vector $\mathbf{V}_{trgt}$ and the intermediate activation maps. On top of that, we conduct an alignment step wherein the scene parameters of $\mathbf{I}$ are extracted by using a face detection network [21] and a facial landmark detection network $\mathcal{M}$ [2]. For each reverse diffusion step $t \in \{T, \cdots, 1\}$, we firstly predict the injected noise and the corresponding noised variable $\mathbf{z}_t$. Then, according to the formula in line 5 of Algorithm 1, the initial expected latent vector $\bar{\mathbf{z}}_0$ is estimated, followed by the decoding step. The estimated initial facial texture $\bar{\mathbf{T}}_0$ and the estimated initial facial shape $\bar{\mathbf{S}}_0$ are computed using the multi-branch facial texture decoder $\mathcal{D}$ and the PCA model $\mathcal{F}_{shp}$, respectively. After being rendered, the expected facial image $\bar{\mathbf{I}}_0$ is generated. We compare the identity embedding vectors between the target image $\mathbf{I}$ and the expected facial image $\bar{\mathbf{I}}_0$ by using the identity cosine distance and identity perceptual loss as defined in [13]. Finally, we obtain accurate illumination and facial expression parameters by

Algorithm 1 Diffusion sampling using Identity Guidance

```
Input: A facial "in-the-wild" image $\mathbf{I}$, a gradient scale $s$, and networks $\mathcal{C}$ [7], $\mathcal{M}$ [2],$\mathcal{F}_{shp}$ [1], $\mathcal{V}$ [25], and the multi-modal decoder $\mathcal{D}$ .
Output: $\mathbf{z}_0 = \{\mathbf{z}_{tex}|\mathbf{z}_{shp}|\mathbf{z}_{ill}\}$.
$\mathbf{z}_T = \{\mathbf{z}_{tex_T}|\mathbf{z}_{shp_T}|\mathbf{z}_{ill_T}\} \backsim \mathcal{N}(\mathbf{0},\mathbf{1})$
$\mathbf{V}_{trgt} = \mathcal{C}(\mathbf{I})$
for all t from T to 1 do
    $\mu, \mathbf{\Sigma} \leftarrow \epsilon_\theta(\mathbf{z}_t, t, \mathbf{V}_{trgt})$
    $\bar{\mathbf{z}}_0 = \frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{z}_t, t, \mathbf{V}_{trgt})}{\sqrt{\bar{\alpha}_t}}$
    $\bar{\mathbf{I}}_0 \xleftarrow{\text{render}} \bar{\mathbf{T}}_0, \bar{\mathbf{S}}_0 \xleftarrow[\mathcal{D},\mathcal{F}_{shp}]{\text{decode}} \bar{\mathbf{z}}_0$
    $\mathcal{G}_{id}^{cos} \leftarrow (1 - cos(\mathbf{V}_{trgt}, \mathcal{C}(\bar{\mathbf{I}}_0))$
    $\mathcal{G}_{id}^{per} \leftarrow \sum_i \frac{\mathcal{C}^i(\bar{\mathbf{I}}_0) \cdot \mathcal{C}^i(\mathbf{I})}{H_{\mathcal{C}^i} \cdot W_{\mathcal{C}^i} \cdot C_{\mathcal{C}^i}}$
    $\mathcal{G}_{mse} \leftarrow \|\bar{\mathbf{I}}_0 - \mathbf{I}\|_2$
    $\mathcal{G}_{lan} \leftarrow \|\mathcal{M}(\bar{\mathbf{I}}_0) - \mathcal{M}(\mathbf{I})\|_2$
    $\mathcal{G}_{vgg} \leftarrow \|\mathcal{V}(\bar{\mathbf{I}}_0) - \mathcal{V}(\mathbf{I})\|_2$
    $\mathcal{G} = \mathcal{G}_{id}^{cos} + \lambda_1 \cdot \mathcal{G}_{id}^{per} + \lambda_2 \cdot \mathcal{G}_{mse} + \lambda_3 \cdot \mathcal{G}_{lan} + \lambda_4 \cdot \mathcal{G}_{vgg}$
    $\mathbf{z}_{t-1} \backsim \mathcal{N}(\mu - s\mathbf{\Sigma}\nabla_{\mathbf{z}_t}\mathcal{G}, \mathbf{\Sigma})$
end for
return $\mathbf{z}_0$
```

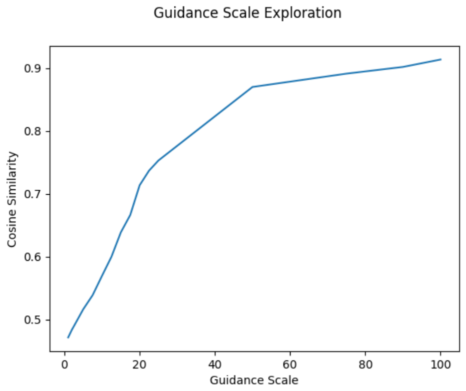


Figure 3. Guidance scale exploration: We randomly pick 10 facial images across the web. We measure the identity similarity between the ground truth image and the generated avatars for different guidance scales.

penalizing the disparity between the per-pixel color intensity and the 3D facial landmarks using $\mathcal{G}_{mse} = \| \bar{\mathbf{I}}_0 - \mathbf{I} \|_2$, $\mathcal{G}_{vgg} = \|\mathcal{V}(\bar{\mathbf{I}}_0) - \mathcal{V}(\mathbf{I})\|_2$, and $\mathcal{G}_{lan} = \| \mathcal{M}(\mathbf{I}_0) - \mathcal{M}(\mathbf{I}) \|_2$.

4. Controlling the generated identity

Choosing the guidance scale is an important factor for the trade-off between the intra-class diversity of the generated samples and the accuracy of the reconstruction. We conduct an experiment by choosing 10 "in-the-wild" images across the web and sample while using different guidance scales for a range of $s = [0, 100]$. We showcase the results in Fig. 3.

5. Partial Texture Completion

Inspired by the in-painting approach presented in Relightify [16], FitDiff finds another application in the domain of partial reflectance map completion, illustrated in Fig. 4. In certain scenarios, the input reflectance map may be provided partially completed. Due to the absence of ground truth facial reflectance maps and with the intention of demonstrating our model's ability to complete partial texture maps, we examine the following scenario: Given the input images illustrated in Fig. 4a, we firstly reconstruct the corresponding facial identity (Fig. 4b and 4e). The resulting diffuse albedo images are treated as pseudo-ground truth and a part of it is randomly masked (Fig. 4c). Obtaining completed diffuse albedo maps involves sampling while using only the input identity embedding vector. The resulting diffuse albedos are showcased in Fig. 4d whilst the corresponding renderings are shown in Fig 4f. By comparing those figures, it is evident that FitDiff clearly retrieves the masked parts, effectively completing the partially visible reflectance maps.

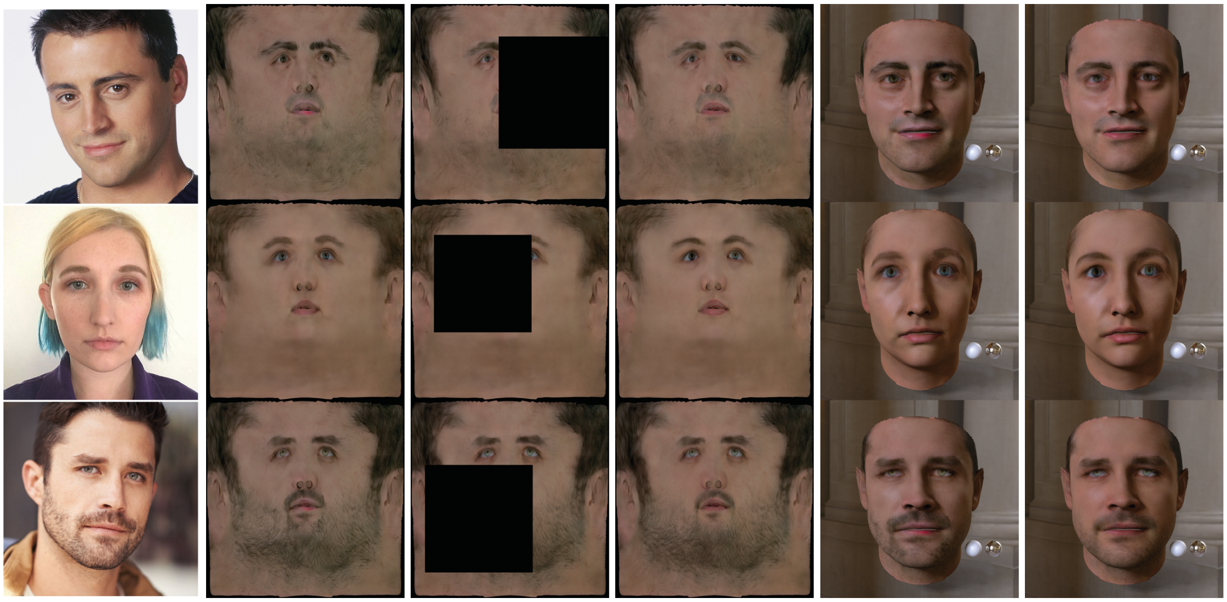
6. Disentanglement Control

Another application of the proposed guidance method in Relightify [16] is used for examining the disentanglement abilities of our proposed approach. More specifically, we consider the following scenarios a) given the facial texture maps as input, FitDiff generates unconditional facial shapes b) given the input facial shapes as input, our method generates facial reflectance maps. We present some results in Fig. 5.

7. Additional Results

7.1. Shape Reconstruction - REALY benchmark

We evaluate our method's shape reconstruction with state-of-the-art methods using REALY [3], a widely used public benchmark. It contains 100 high-quality face shapes from different ethnic and age backgrounds, based on the LYHM dataset [5]. Contrary to previous face geometry reconstruction challenges [20], the REALY benchmark computes geometric errors separately for each region of the human face while using the l_2 distance between the ground truth and the predicted meshes. The results of this benchmark are showcased in Tab. 3 and 4. Our method ranks 7th on the average reconstruction error and gets surpassed only by models that either focus solely on generating facial shape (HiFace [4]). or produce a single albedo map with baked-in illumination (HRN [14], Deep3D [8], AlbGAN [18], MGC-Net [22]), which restricts the resulting avatars from being relightable.



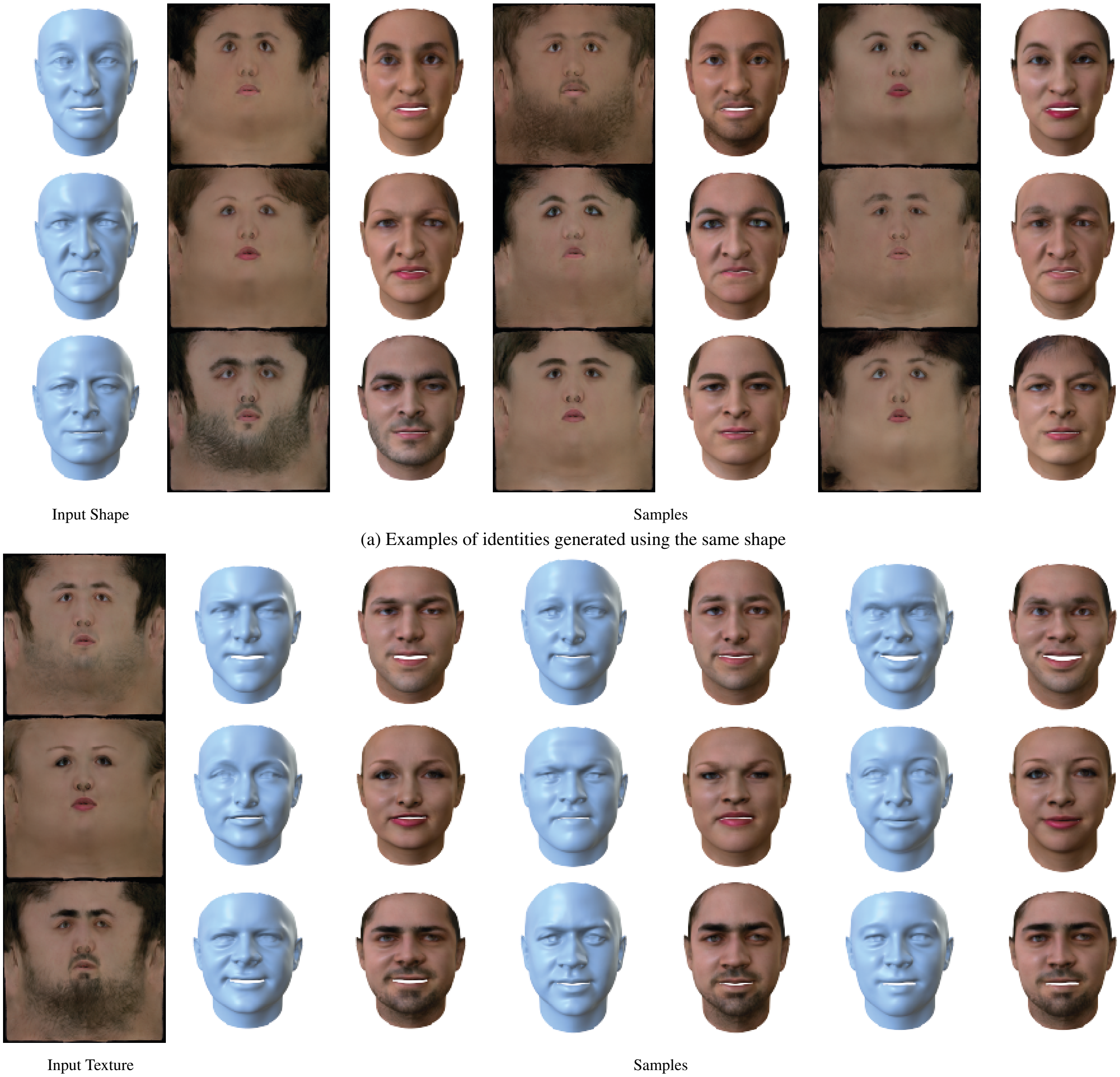
(a) Input Facial Images
(b) Acquired Diffuse Albedo ($\mathbf{A}_D$)
(c) Partial Diffuse Albedo ($\mathbf{A}_D$)
(d) Final Diffuse Albedo ($\mathbf{A}_D$)
(e) Rendering using acquired $\mathbf{A}_D$
(f) Rendering using completed $\mathbf{A}_D$

Figure 4. Our method can be used for facial texture completion.

	@Nose			@Mouth		
Method	avg	med	std	avg	med	std
HiFace-f [4]	**1.036**	**0.992**	**0.280**	1.450	1.388	0.413
HiFace-c [4]	1.054	1.021	0.317	1.461	1.381	0.430
HRN [14]	1.722	1.685	0.330	**1.357**	**1.226**	0.523
Deep3D [8]	1.719	1.683	0.354	1.368	1.301	0.439
AlbGAN [18]	1.656	1.636	0.374	2.087	1.927	0.839
MGCNet [22]	1.771	1.741	0.380	1.417	1.355	0.409
GANFit [12]	1.928	1.881	0.490	1.812	1.769	0.544
FitMe [13]	1.833	1.796	0.434	1.752	1.629	0.539
PSL [15]	1.708	1.688	0.349	1.708	1.777	0.563
DECA-c [11]	1.697	1.654	0.355	2.516	2.465	0.839
CEST [23]	2.779	2.717	0.835	1.448	1.438	**0.406**
EMOCA-c [6]	1.868	1.821	0.387	2.679	2.419	1.112
MICA [26]	1.585	1.542	0.325	3.478	3.439	1.204
DECA-f [11]	2.138	2.137	0.461	2.802	2.699	0.868
EMOCA-f [6]	2.532	2.563	0.539	2.929	2.676	1.106
FitDiff(Ours)	1.821	1.770	0.438	1.751	1.611	0.523

Table 3. Results in the REALY benchmark [3]

Our model's performance can be explained by the fact that our approach is trained using synthetic data obtained from a fitting methodology [13] and generates both albedo and normals UV maps. The utilization of synthetic data imposes inherent limitations on our method's ability to accurately retrieve facial shapes. This limitation stems from the constraints imposed by FitMe on shape retrieval performance. We eventually beat FitMe's performance, as well as similar methods (e.g. GANFit), showing that our results are bounded by the training data, and could improved given a real captured dataset. Additionally, as highlighted by the authors of FitMe, the concurrent reconstruction of both facial shape and texture normals introduces further constraints on the approach's shape reconstruction performance. Specifically, a portion of the shape information is occasionally encoded in the normals domain rather than in



(a) Examples of identities generated using the same shape

(b) Examples of identities generated having the same texture

Figure 5. FitDiff can efficiently disentangle the facial shape and reflectance maps.

the actual facial shape domain.

7.2. Comparison between a single model with separate models

In this section we analyze the effectiveness of a unified single-model architecture compared to a multi-model approach. Using a randomly selected subset of 50 identities from the REALY benchmark [3], we fit our model following two distinct strategies: a) independently sampling for facial shape and reflectance maps, and b) employing our proposed methodology. We then evaluate their identity similarity scores and the facial shape reconstruction performance by comparing the generated facial avatars against the ground truth using the evaluation pipeline provided by the REALY benchmark [3] . The results are presented in Tab. 5, and demonstrate that the unified single-model approach achieves superior performance in both metrics.

	@Forehead			@Cheek			All
Method	avg	med	std	avg	med	std	avg
HiFace-f [4]	**1.324**	**1.296**	**0.334**	1.291	1.240	0.362	**1.275**
HiFace-c [4]	1.331	1.307	0.347	1.342	1.304	0.384	1.297
HRN [14]	1.995	1.990	0.476	**1.072**	1.063	0.333	1.537
Deep3D [8]	2.015	2.007	0.449	1.528	1.442	0.501	1.657
AlbGAN [18]	2.102	2.035	0.512	1.141	1.103	**0.303**	1.746
MGCNet [22]	2.268	2.215	0.503	1.639	1.494	0.650	1.774
GANFit [12]	2.402	3.339	0.545	1.329	1.234	0.504	1.868
FitMe [13]	2.494	2.385	0.605	1.414	1.315	0.526	1.873
PSL [15]	2.350	2.343	0.551	1.593	1.482	0.540	1.882
DECA-c [11]	2.394	2.256	0.576	1.479	1.400	0.535	2.010
CEST [23]	2.384	2.302	0.578	1.456	1.321	0.485	2.017
EMOCA-c [6]	2.426	2.383	0.641	1.438	1.294	0.501	2.103
MICA [26]	2.374	2.251	0.683	1.099	**1.003**	0.324	2.134
DECA-f [11]	2.457	2.341	0.559	1.443	1.353	0.498	2.210
EMOCA-f [6]	2.595	2.505	0.631	1.495	1.360	0.469	2.388
FitDiff(Ours)	2.472	2.322	0.581	1.404	1.287	0.525	1.862

Table 4. Results in the REALY benchmark [3]

Method	Facial Shape ↓	ID similarity ↓
Separate Models	1.805	0.873
FitDiff(Ours)	**1.764**	**0.911**

Table 5. Comparison between the single model approach (FitDiff) and using 2 separate models for facial reconstruction

References

[1] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. 3

[2] 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. 2, 3

[3] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3, 4, 5, 6

[4] Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, and Jiang Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9087–9098, October 2023. 3, 4, 6

[5] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 2019. 3

[6] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 4, 6

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 3

[8] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 3, 4, 6

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021. 1

[10] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision*, 2022. 1

[11] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. Graph.*, 40(4), jul 2021. 4, 6

[12] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 4, 6

[13] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos

Zafeiriou. FitMe: Deep photorealistic 3D morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 1, 2, 4, 6

[14] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images, 2023. 3, 4, 6

[15] C. Otto, P. Chandran, G. Zoss, M. Gross, P. Gotardo, and D. Bradley. A perceptual shape loss for monocular 3d face reconstruction. *Computer Graphics Forum*, 42(7):e14945, 2023. 4, 6

[16] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. *arXiv preprint arXiv:2305.06077*, 2023. 2, 3

[17] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[18] Aashish Rai, Hiresh Gupta, Ayush Pandey, Francisco Vicente Carrasco, Shingo Jason Takagi, Amaury Aubel, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Towards realistic generative 3d face models. *arXiv preprint arXiv:2304.12483*, 2023. 3, 4, 6

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1

[20] 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)*, pages 7763–7772, June 2019. 3

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[22] 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. *arXiv preprint arXiv:2007.12494*, 2020. 3, 4, 6

[23] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3d face reconstruction via conditional estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13269–13278, 2021. 4, 6

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 601–610, 2020. 1

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[26] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pages 250–269. Springer, 2022. 4, 6