Towards Racially Unbiased Skin Tone Estimation via Scene Disambiguation
Supplementary Material

In the following we provide additional information on the benchmark and training datasets (Sec. 1), training details (Sec. 2), additional qualitative and quantitative results (Sec. 3), and details about the geometric and appearance model (Sec. 4).

1 Datasets

Benchmark. In Fig. 2 we show examples of full scene images from the evaluation dataset, along with ground-truth diffuse albedo. Note that the evaluations are performed in UV space (Fig. 1, left), not on the rendered albedo images.

The benchmark was rendered with the Unreal Engine 4.27 pathtracer (ray-tracing), using 155 head scans purchased from Triplegangers, and 50 HDRI panorama scenes from PolyHaven (omni-directional spherical images). The only light source of the scene is the HDRI panorama; additional global illumination, auto-exposure compensation and ground reflection were turned off. From each of these 50 environment maps we rendered 18 scene images, rotating the map 100° per rendering. Scene images with extremely low/high light intensity, which cause the face detector to fail, were removed.

The textures provided by Triplegangers contained tiny markers on the face for better alignment, which we manually removed. Each scene contains three head scans that share the same illumination. After classifying each scan according to its ITA value, we selected three heads for each scene as follows. We first uniformly sample a skin category from three possible groups, where each group contains two consecutive ITA skin types (I-II, III-IV, V-VI). We take this approach since type I contains very few samples (note that the population having type I skin is in general very low, as can be seen in the ITA graphs from [1]). We next randomly select a head scan from the designated skin category. This ensures an overall balance of skin tones in the benchmark.

To measure ITA we compute per-pixel ITA values over a mask that covers mainly the cheek region, as this is the most consistent area of the UV map in terms of skin color. Note that a similar protocol is usually employed in the field of dermatology [1], where measures are taken also on the cheek area. A visualization of this mask can be found in Fig 1, right.

The testing dataset contains 721 images and 2163 facial crops under different illumination conditions, with ground-truth UV maps. We also build a smaller validation set consisting of 234 images and 702 crops, which we use for ablation studies.

Synthetic training set. Fig. 4 shows examples of the synthetic dataset used for training the TRUST network. The synthetic training set was built by rendering 50K images using Unreal Engine 4.27 pathtracer (ray-tracing), with 1170 scans and 182 identities purchased from Renderpeople, as well as 273 HDRI panorama scenes from PolyHaven. Note that these scenes are different from the ones used in the benchmark dataset. As with the benchmark dataset, the only light source of the scene is the HDRI panorama; additional global illumination, auto-exposure compensation and ground reflection were turned off.

To generate each synthetic image we first randomly select the number of subjects in the scene (one to six), then uniformly sample an ethnicity for each subject (ethnicity labels were provided by Renderpeople), and finally randomly choose a scan according to the selected ethnicity. The scans are randomly positioned in the scene. Since our scene backgrounds are panoramas, we are able to render the images with any random section of each scene (we used a virtual camera with a focal length of 50mm, FOV 40 degrees).

We obtain pseudo ground-truth albedo by using the unlit version of the scan textures, as provided by Renderpeople. We obtain pseudo ground-truth for the environmental lighting as follows. First we insert a matte gray light probe in each of the 273 scenes, and render the light probe from six orthogonal views. We next optimize for the spherical harmonic coefficients that best explain the renders given the known shape and color, using an L1 photometric loss. We use the Adam optimizer [2] for 3000 iterations and a learning rate of 1e-4.

2 Training details

As explained in Sec. 5.2, we use a semi-supervised training strategy that combines a synthetic dataset with known ground-truth, with an in-the-wild dataset for generalization. This process is illustrated in Fig. 3. At train time the input batch contains $M \times N$ facial crops extracted from M scenes. Half of the scenes come from the synthetic dataset, while the other half comes from an in-the-wild dataset. In particular, we use for the latter the subset of the OpenImages dataset [3] that was labeled 'human face', 'woman' or 'man', and was not labeled as 'depiction' (i.e. drawings, etc).

For the synthetic half of the batch we apply an L1 loss between predicted and ground-truth albedo (L_{alb}), and an L1 loss between predicted and ground-truth spherical harmonics (L_{SH}).

Additionally, for both synthetic and real data we employ (1) the photometric loss L_{pho} and (2) the scene consistency loss L_{sc}. The scene consistency loss is applied over the N facial crops coming from the individual scene images (depicted with a same color in Fig. 3). Given the spherical harmonics parameters for each of the N crops, L_{sc} requires these to be close to each other. This is implemented by randomly permuting the SH of the crops from a same scene, and applying an L1 loss between original and permuted (see Fig.2 in main paper, right).

Note that we apply the photometric loss also on synthetic data because both the ground-truth light and the ground-truth albedo are only approximations, and

because the renderings were done with path tracing, which spherical harmonics can only approximate in a smooth way.

3 Additional results

This section shows additional qualitative and quantitative results.

First, Table 1 is an extension of Table 2 from the main paper that includes per-skin-type values. We include additional qualitative results that support this table in Figs. 5 and 6.

The effectiveness of the *scene consistency* loss L_{sc} is illustrated by ablating it as illustrated in Fig. 5. The results show that our full model is able to predict more faithful skin tone than the version without L_{sc}. This aligns with our quantitative evaluation results, where the bias score is improved by the use of L_{sc}.

In Fig. 6 we show the effectiveness of conditional albedo estimation. When the scene contains low light intensity, the model without light conditioning does not have enough information to disambiguate, hence predicting darker skin tones.

We also include an additional ablation in Table 2 that shows (1) the performance of our method when trained on synthetic data alone ("only w/syn"), but with all the proposed losses, and (2) a fully supervised version of our method ("full-sup"), where only L_{alb} is employed. Training the albedo encoder solely with the albedo supervision loss (Table 2, "full-sup") quickly leads to overfitting (indeed, after one epoch) that does not generalize to unseen data. Training the full model solely with synthetic data but with all the proposed losses ("only w/syn") leads also to poor generalization, mostly due to the lack of real-world data. These results support the need for combining a supervised approach with the proposed self-supervised losses, in order to achieve good generalization.

To evaluate the robustness of our method in terms of head rotations, we further conduct a controlled experiment on images of the same subject with yaw rotations ranging from -45 to +45 degrees. These rotations were applied to every sample in the benchmark test set. Fig. 7 shows that the estimated albedo per subject is relatively consistent with all head rotations, and that the skin tones are well reconstructed. Additional qualitative results are shown in Fig. 9, Fig. 10 and Fig. 11.

Limitations. We illustrate in Fig. 8 some of the limitations of our method:

1. Our model is trained with a limited number of scenes. Because of this, it will not perform well under very extreme lighting cases (e.g. a very dark scene), as shown in Fig. 8, first column.
2. When the scene background does not provide much information, the learned scene light prior cannot properly work, as shown in Fig. 8, second column.
3. Our assumptions fail when the local lighting on the face is different from the global scene lighting, e.g. a flash light or local shadows cast by a hat or nearby objects. This can be seen in Fig. 8, third and fourth columns.

Table 1: Ablation study (main paper) with additional per-skin-type results, conducted on the validation set of the FAIR benchmark. We show average ITA score per skin type in degrees (I: very light, VI: very dark), average ITA error over all skin types, bias score and total score.

Method	ITA per skin type ↓						ITA Avg ↓	Bias ↓	Score ↓
	I	II	III	IV	V	VI			
faceSH + self	37.30	28.44	19.60	10.97	10.17	38.51	24.16	11.45	35.62
faceSH + semi	14.16	8.77	9.49	12.96	14.14	28.69	14.70	6.61	21.31
fuseSH	14.60	7.78	6.95	10.19	12.49	39.31	15.22	11.08	26.31
fuseSH + sc	19.80	14.29	12.01	13.10	14.33	21.40	15.82	3.50	19.32
fuseSH + cond	14.53	10.92	8.34	10.47	12.47	28.21	**14.16**	6.56	20.72
Ours	14.34	12.19	10.75	13.64	15.01	19.16	14.18	**2.63**	**16.82**

Table 2: Additional ablation study, to investigate the effects of supervised training, conducted on the validation set of the FAIR benchmark. We show average ITA score per skin type in degrees (I: very light, VI: very dark), average ITA error over all skin types, bias score and total score.

Method	ITA per skin type ↓						Avg ITA ↓	Bias ↓	Score ↓
	I	II	III	IV	V	VI			
full-sup (only w/ syn, L_{alb})	45.81	37.09	27.37	22.91	20.56	29.43	30.53	9.44	39.96
Ours (only w/ syn)	12.60	11.95	12.02	13.34	20.67	48.95	19.92	14.60	34.53
Ours	14.34	12.19	10.75	13.64	15.01	19.16	**14.18**	**2.63**	**16.82**

4 Geometry and appearance model

This section describes in more detail the geometry and diffuse albedo model used in this work.

Geometry prior. We reconstruct geometry using the FLAME [4] statistical model, which parameterizes a head mesh using identity $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$, pose $\boldsymbol{\theta} \in \mathbb{R}^{3k+3}$ (with $k = 4$ joints for neck, jaw, and eyeballs), and expression $\boldsymbol{\psi} \in \mathbb{R}^{|\boldsymbol{\psi}|}$ latent vectors. The model is defined as

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = W(T_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}), \mathbf{J}(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathcal{W}), \tag{1}$$

with the blend skinning function $W(\mathbf{T}, \mathbf{J}, \boldsymbol{\theta}, \mathcal{W})$ that rotates the vertices in $\mathbf{T} \in \mathbb{R}^{3n}$ around joints $\mathbf{J} \in \mathbb{R}^{3k}$, linearly smoothed by blendweights $\mathcal{W} \in \mathbb{R}^{k \times n}$. The joint locations $\mathbf{J}$ are defined as a function of the identity $\boldsymbol{\beta}$, and

$$T_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbf{T} + B_S(\boldsymbol{\beta}; \mathcal{S}) + B_P(\boldsymbol{\theta}; \mathcal{P}) + B_E(\boldsymbol{\psi}; \mathcal{E}), \tag{2}$$

where $\mathbf{T}$ denotes the mean template in "zero pose", with shape blendshapes $B_S(\boldsymbol{\beta}; \mathcal{S}) : \mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{3n}$, pose correctives $B_P(\boldsymbol{\theta}; \mathcal{P}) : \mathbb{R}^{3k+3} \rightarrow \mathbb{R}^{3n}$, and expression blendshapes $B_E(\boldsymbol{\psi}; \mathcal{E}) : \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow \mathbb{R}^{3n}$, using the learned identity, pose, and expression bases (i.e. linear subspaces) $\mathcal{S}, \mathcal{P}$ and $\mathcal{E}$.

Diffuse albedo model. We train the diffuse albedo model using Principal Component Analysis. Given albedo parameters $\boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{\alpha}|}$, and the learned albedo bases $\mathbf{B_{alb}}$, we reconstruct a UV map by

$$A(\boldsymbol{\alpha}) = \bar{A} + \mathbf{B_{alb}}\boldsymbol{\alpha} \tag{3}$$

where $\bar{A}$ is the mean albedo from the training set, $A(\boldsymbol{\alpha})$ is a vectorized representation of the image, which is then transformed to a $d \times d$ UV map.

Fig. 1: Left: ground-truth UV map from the benchmark. Right: masked values used for ITA calculation.

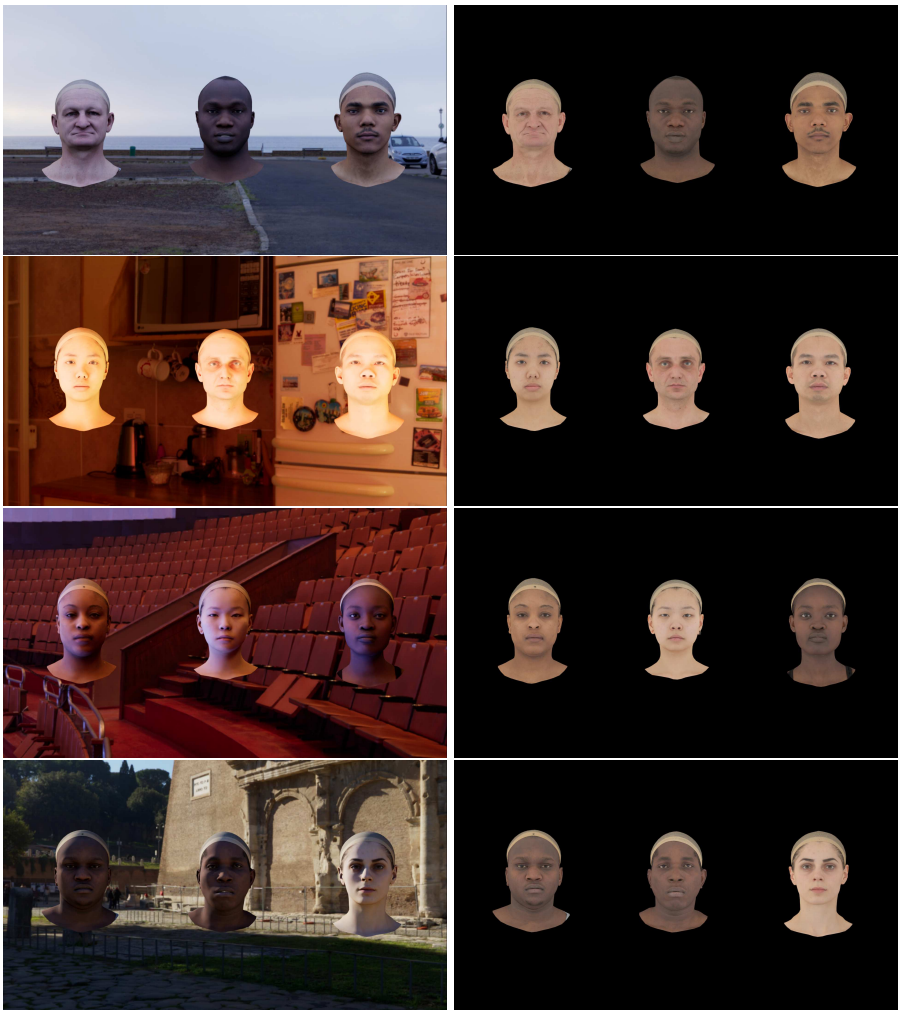

Fig. 2: Benchmark examples. Left: input image, right: albedo.

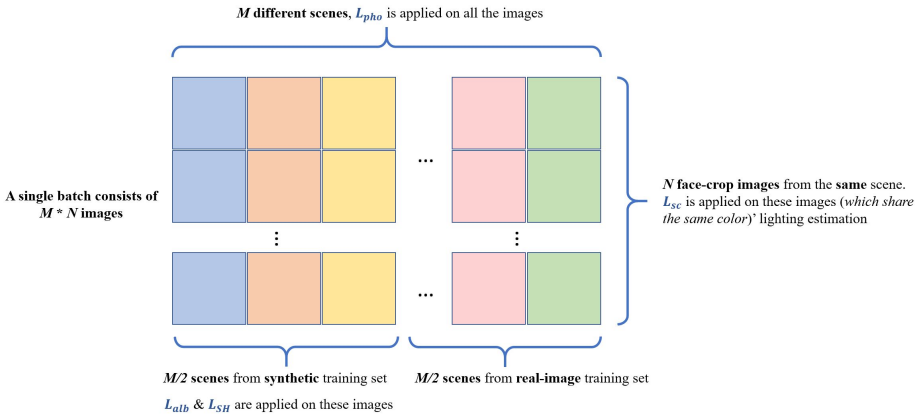

Fig. 3: Composition of a mini-batch during one train iteration. TRUST is trained with a combined dataset of synthetic and real images, and optimized with supervised and self-supervised loss terms.

Fig. 4: Synthetic training set examples. Left: input image, right: albedo.

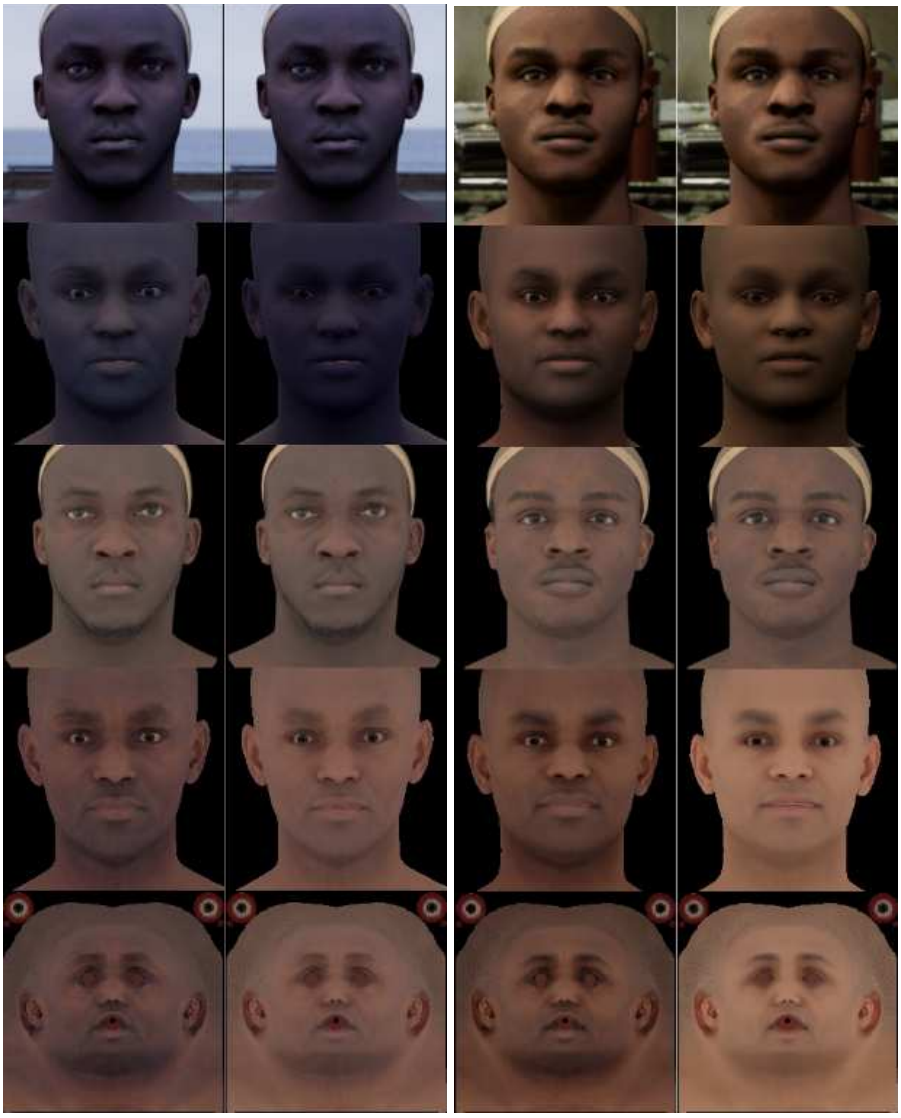

Fig. 5: Qualitative results illustrating the ablation of the scene consistency loss. We show two pairs of comparisons using the benchmark data (validation). Each pair shows our full model (left) and fuse-cond model (right), where no scene consistency was used. Images from top to bottom: Input cropped face image, predicted texture reconstruction (albedo + shading), ground-truth albedo (rendering), predicted albedo (rendering), predicted albedo (UV map).

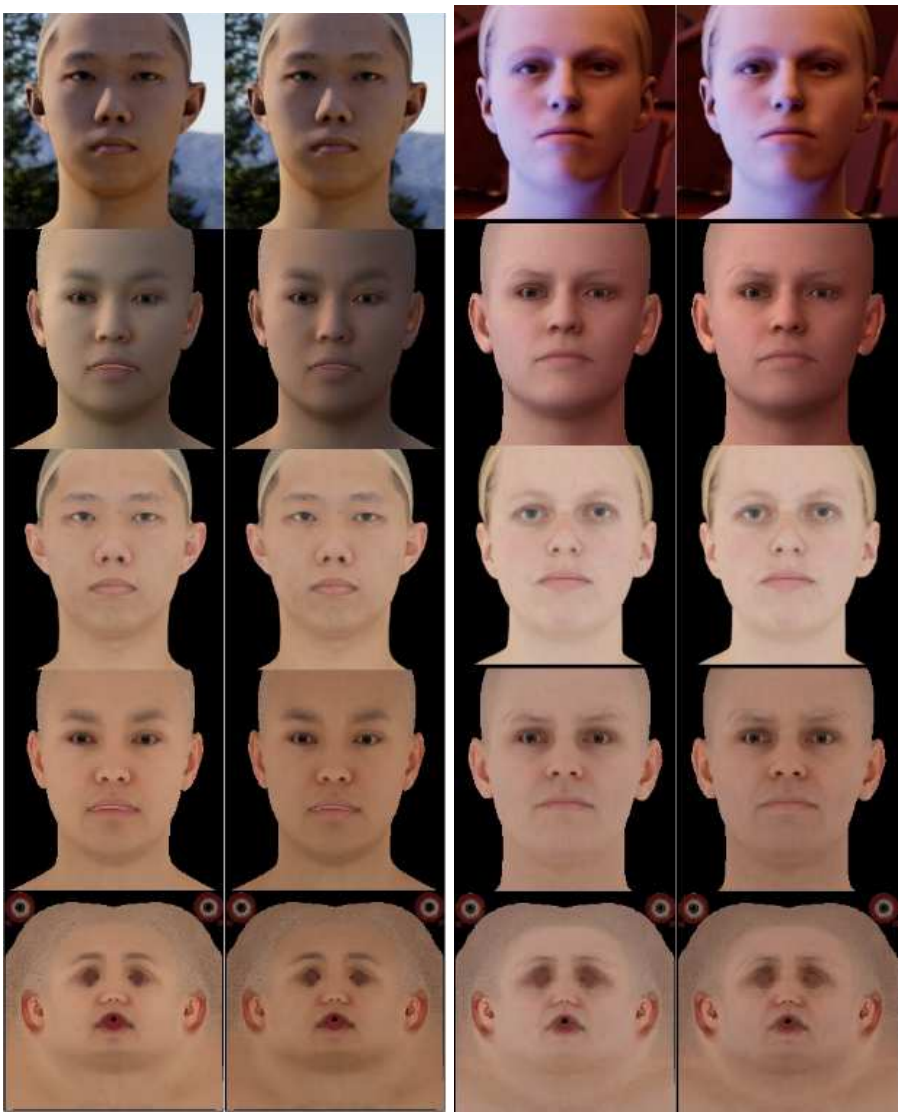

Fig. 6: Qualitative results illustrating the ablation of conditional albedo estimation. We show two pairs of comparisons using the benchmark data (validation). Each pair shows our full model (left) and fuse-sc model (right), where no conditioning was used. Images from top to bottom: Input cropped face image, predicted texture reconstruction (albedo + shading), ground-truth albedo (rendering), predicted albedo (rendering), predicted albedo (UV map).

Fig. 7: Qualitative results illustrating the robustness of TRUST on estimating consistent albedo under different yaw rotations of the same subject.

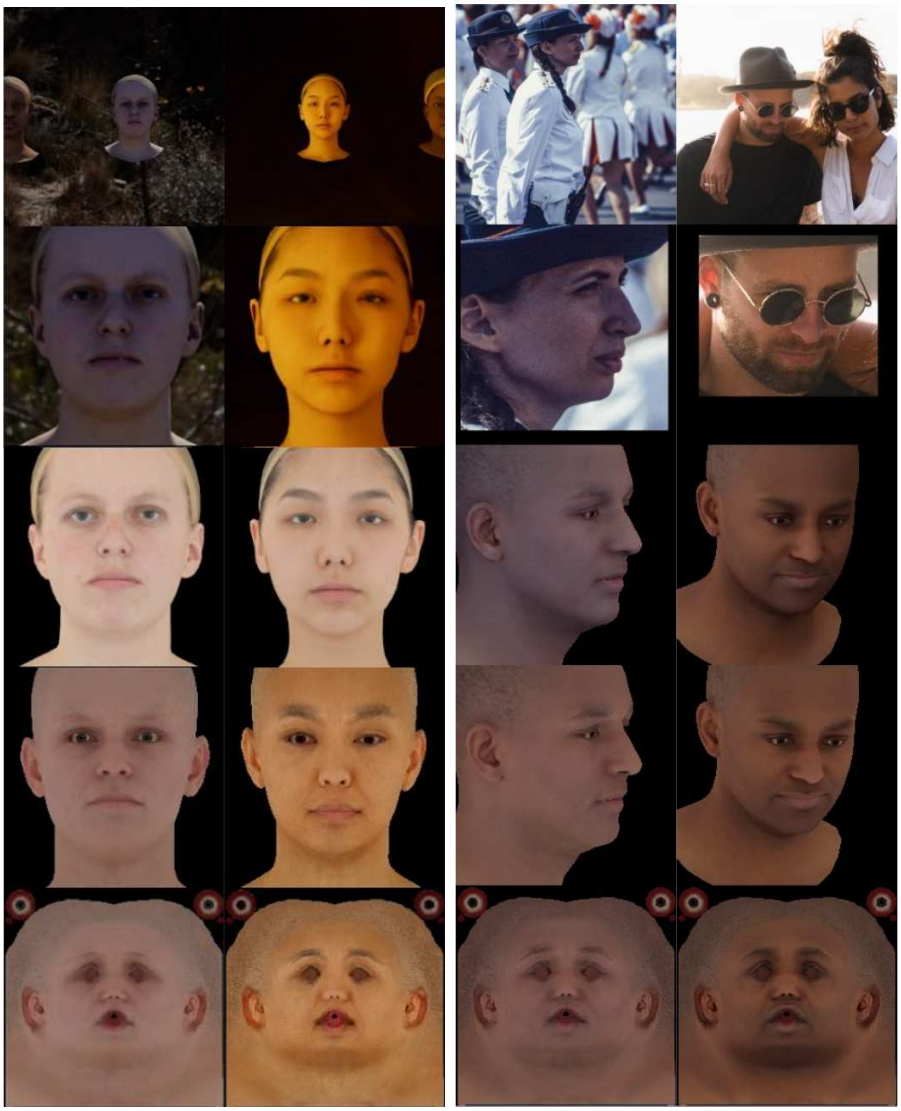

Fig. 8: Limitation examples. Left: samples from our proposed benchmark: (1) extreme low light, (2) example where the scene background does not provide lighting information. Right: real images: our assumption breaks under local shadows.

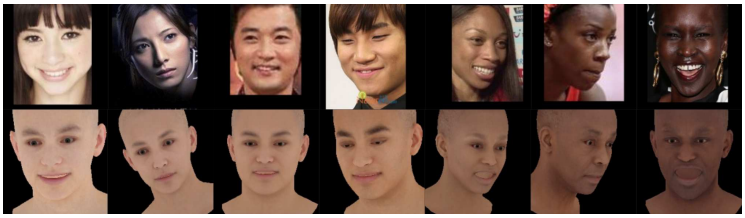

Fig. 9: Qualitative examples of face crops from BFW dataset. Top: input face crop, bottom: our predicted albedo (rendering).

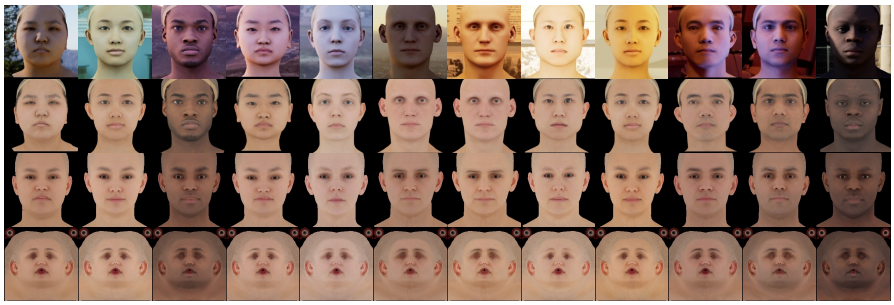

Fig. 10: Qualitative examples from the benchmark. From top to bottom: Input face crop, ground-truth albedo (rendering), our predicted albedo (rendering), our predicted albedo (UV map).

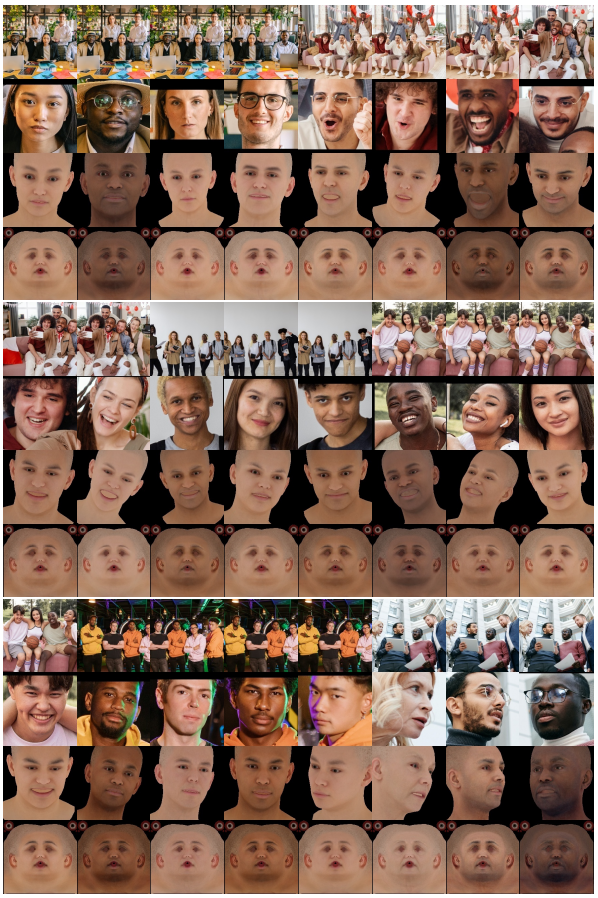

Fig. 11: Qualitative examples from real world images, downloaded from https://www.pexels.com/. From top to bottom: input scene image, input face crop, our predicted albedo (rendering), our predicted albedo (UV map).

Bibliography

[1] Del Bino, S., Bernerd, F.: Variations in skin colour and the biological consequences of ultraviolet radiation exposure. British Journal of Dermatology **169**, 33–40 (2013) 1

[2] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 2

[3] Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github. com/openimages **2**(3), 18 (2017) 2

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