A. More Implementation Details

A.1. Multi-PIE Dataset Preprocessing

We select 9 viewpoints (09, 08, 13, 14, 05, 05, 04, 19, and 20) and 11 flashes (03, 04, 05, 06, 07, 08, 09, 10, 11, 14, and 18) for reflectance parameter estimation. Please refer to [11] for the detailed configuration of the viewpoints and flashes. We develop a model-based method to reconstruct the camera parameters and the BFM09 [14] geometry coefficients for each identity. According to the Multi-PIE dataset [11], each selected viewpoint has one selected flash attached to it. Hence, we approximate the flash position as the camera position. We use the room-light images [11] for reconstruction. Specifically, we first adopt a CNN-based single-view face reconstruction method [5] to obtain the BFM09 coefficients, illumination coefficients, and head pose for each room-light image of a given identity. Then, we apply an offline optimization using the same loss function as [5] to improve the reconstruction accuracy. During the offline optimization, each room-light image shares the same BFM09 coefficients since they are the multi-view images of the given identity, and we initialize them as the average of the coefficients of all views predicted by the face reconstruction CNN. Similar to [5], we use the perspective camera model with a reasonable predefined focal length to represent the 3D-2D projection. After reconstruction, we can compute the camera parameters from the head pose $R$ and $t$ for each viewpoint:

$$R_{\text{cam}} = R^T, \quad t_{\text{cam}} = -R^T \cdot t$$

(1)

Here, $R_{\text{cam}}$ and $t_{\text{cam}}$ are the camera rotation and translation in the BFM09 canonical space, respectively. We repeat the steps above for all the identities in the Multi-PIE dataset.

Before reflectance parameter estimation, we obtain the OLAT image by removing the effect of the room light in the flash image. Specifically, we subtract the room-light image from the flash image in linear space with a reasonable mapping function:

$$I_{\text{OLAT}} = \frac{(I_{\text{flash}})^{1.2} - (I_{\text{roomlit}})^{1.2}}{2}$$

(2)

Here, $I_{\text{OLAT}}$ is the OLAT image in linear space, $I_{\text{flash}}$ and $I_{\text{roomlit}}$ are the flash and room-light image provided by the Multi-PIE dataset, respectively. We then estimate the reflectance parameters from $I_{\text{OLAT}}$ and build our morphable face reflectance model in linear space. To synthesis a face image in nonlinear space, we convert the shading $s$ to pixel color $c$ using the inverse mapping:

$$c = s^{1/2}$$

(3)

Demographics

Our initial morphable face reflectance model is built from a total of 128 manually selected individuals from the Multi-PIE dataset. We release the ID of the selected individuals in our code repository.

Feasibility of reflectance parameter estimation

The RGB diffuse color and 3 linear combination weights are the only unknowns in our reflectance representation. Theoretically, the ambiguity can be solved with 6 independent equations. We have 99 light-view direction pairs (the combination of 9 viewpoints and 11 light directions) in total, and if considering visibility, most of the vertices have 50+ light-view direction pairs. Different light-view direction pairs give independent equations. Thus, it’s feasible to estimate the BRDF parameters theoretically.

Practically, the light-view direction pairs which are not hitting the lobe of the BRDF would lead to a low activation value, and thus solving the reflectance parameters from these equations are highly ill-posed. In our setup, we find that the ill-posed scenario only happens on very few face vertices on the side face or with normal directions going down like nares. For most of the face vertices, our setup can provide enough well-conditioned equations with the corresponding light-view direction pairs hitting the lobe. Thus, it’s feasible to estimate the BRDF parameters practically.

We empirically find that performing image differencing in linear space leads to better reflectance parameter estimation than in non-linear space.
A.2. Model Finetuning

Recall that in model finetuning, the learnable parameters are the morphable model parameters, including the mean \( \tilde{R} \) and bases \( M_R \), and face reconstruction network parameters \( \theta \). We optimize them with the combination of a reconstruction loss \( L_{\text{rec}} \) and a regularization loss \( L_{\text{reg}} \):

\[
\arg \min_{\tilde{R}, \text{M}_R, \theta} L_{\text{rec}} + L_{\text{reg}}
\]

(4)

\( L_{\text{rec}} \) is the combination of a L1 term \( L_{11} \) and a perceptual term \( L_{\text{per}} \):

\[
L_{\text{rec}} = \omega_{11} \cdot L_{11} + \omega_{\text{per}} \cdot L_{\text{per}}, \quad \text{where}
\]

(5)

\[
L_{11} = M_{\text{skin}} \cdot ||I - \hat{I}||_1
\]

(6)

\[
L_{\text{per}} = 1 - \langle \phi_{\text{feat}}(\hat{I}), \phi_{\text{feat}}(I) \rangle
\]

(7)

Here, \( M_{\text{skin}} \) is the masked skin region, obtained by an off-the-shelf face parsing method [24]; \( \langle \cdot , \cdot \rangle \) is the inner product operation; \( \phi_{\text{feat}} \) is a pretrained FaceNet architecture [17] for feature extraction. Note that we directly compute the reconstruction loss \( L_{\text{rec}} \) in the linear space. Although \( \phi_{\text{feat}} \) is trained on images in the nonlinear space, we empirically find that it can still provide a reasonable supervision signal if the input image is in the linear space.

In our regularization loss \( L_{\text{reg}} \), we first adopt \( L_{\text{coef}} \) to constrain the predicted PCA coefficients \( \beta \) and \( \gamma \):

\[
L_{\text{coef}} = \sum_{i=1}^{N_R} (\beta_i / \sigma_{\beta_i})^2 + \sum_{i=1}^{N_L} (\gamma_i / \sigma_{\gamma_i})^2
\]

(8)

Here, \( \sigma_{\beta} \) and \( \sigma_{\gamma} \) are the standard deviations of the initial morphable face reflectance model and the lighting PCA model, respectively. Then, to constrain the updating of our morphable reflectance model, we design \( L_{\text{upd}} \) as:

\[
L_{\text{upd}} = ||\tilde{\text{R}} - \tilde{R}_0||_1 + ||\text{M}_R - \text{M}_{R_0}||_1
\]

(9)

Here, \( \tilde{R}_0 \) and \( \text{M}_{R_0} \) are the mean and bases of our initial morphable face reflectance model built from the Multi-PIE dataset. To resolve the color ambiguity between albedo and lighting, we involve \( L_{\text{light}} \) to encourage monochromatic environment lighting as [4]:

\[
L_{\text{light}} = ||l - l_{\text{mean}}||_2^2
\]

(10)

Here, \( l \) is the retrieved 8-th order SH coefficients; \( l_{\text{mean}} \) is the mean of \( l \) over the color channel dimension, representing the monochromatic counterpart of \( l \). Thus, our regularization loss \( L_{\text{reg}} \) can be written as:

\[
L_{\text{reg}} = \omega_{11} \cdot L_{\text{coef}} + \omega_{\text{per}} \cdot L_{\text{upd}} + \omega_{\text{light}} \cdot L_{\text{light}}
\]

(11)

In our experiments, we set \( \omega_{11}, \omega_{\text{per}}, \omega_{\text{coef}}, \omega_{\text{upd}}, \omega_{\text{light}} \) to 2, 0.1, 0.001, 10, 10, respectively.

Table 1. Quantitative face geometry reconstruction error on the validation set of the NoW challenge.

<table>
<thead>
<tr>
<th></th>
<th>Median (mm) ↓</th>
<th>mean (mm) ↓</th>
<th>std (mm) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFM09</td>
<td>1.44</td>
<td>2.06</td>
<td>2.51</td>
</tr>
<tr>
<td>Ours</td>
<td>1.51</td>
<td>2.15</td>
<td>2.61</td>
</tr>
</tbody>
</table>

B. More Results

B.1. Model Visualization

In Figure 2 and Figure 3, we visualize our model by showing random samples drawn from it before and after fine-tuning, respectively. The images are rendered in nonlinear space with a white frontal point light.

B.2. Face Reconstruction

More Reconstruction Results

We show more face reconstruction results on in-the-wild face images in Figure 1, including diverse ethics groups and challenging cases with facial occlusions and makeups. We multiply the linear combination weights (columns 3, 4, 5 in Figure 1) by 3 for better visualization.

Thanks to the model-finetuning process, our method is robust to handle diverse input images and predicts plausible reflectance attributes. However, it has the same limitation as previous in-the-wild face reconstruction methods [5, 19, 20]: i) the global skin tone can not be disentangled from the illumination due to the scale ambiguity between lighting and reflectance (row 5), and ii) shadow cast by external geometry (hat in row 9) bakes into the reflectance channels.

Evaluation on Geometry Reconstruction

Although our goal is not to better reconstruct face shape from images, we compare our method and BFM09 [14] on the validation set of the NoW challenge [16] to help the readers better understand our model. Note that both methods use the same BFM09 geometry model; we do not compare to AlbedoMM since AlbedoMM [18] is built on top of the BFM17 [10] geometry model.

In this experiment, we adopt a similar network architecture as [5] by simply modifying the number of neurons of the last fully-connect layer of \( E_B(\cdot) \) from \( N_R + N_L + 3 \) to \( N_S + N_E + N_P + N_R + N_L + 3 \) to predict the shape and expression coefficients and the head pose. We use the first 80 and 64 bases of the BFM09 shape and expression morphable model, respectively; thus, \( N_S = 80 \) and \( N_E = 64 \). For the head pose, we use the Euler angle to represent rotation and a 3D vector to represent translation; thus, \( N_P = 6 \). To train the network for geometry reconstruction, we involve a landmark loss term akin to previous works [5, 16, 21, 22]:

\[
L_{idm} = \sum_{n=1}^{68} ||\hat{q}_n - q_n||_2^2
\]

(12)
Figure 1. Face reconstruction results on diverse in-the-wild face images.
Figure 2. 60 random samples drawn from our initial morphable face reflectance model (before model finetuning). Rendered in nonlinear sRGB space with a white frontal point light.

Here, \(q_n\) are the 2D landmarks obtained from an off-the-shelf landmark detector \([1]\); \(\hat{q}_n\) are the 2D projection of the 3D landmarks defined on the reconstructed shape. In addition, we modify \(L_{\text{coef}}\) to add constraints on the shape and expression coefficients:

\[
L_{\text{coef}} = \sum_{i=1}^{N_S} \left( \frac{\alpha_i}{\sigma_{\alpha_i}} \right)^2 + \sum_{i=1}^{N_E} \left( \frac{\delta_i}{\sigma_{\delta_i}} \right)^2 + \sum_{i=1}^{N_R} \left( \frac{\beta_i}{\sigma_{\beta_i}} \right)^2 + \sum_{i=1}^{N_L} \left( \frac{\gamma_i}{\sigma_{\gamma_i}} \right)^2
\]

(13)

Here, \(\alpha \in \mathbb{R}^{N_S}\) and \(\delta \in \mathbb{R}^{N_E}\) are the predicted shape and expression coefficients, respectively; \(\sigma_{\alpha}\) and \(\sigma_{\delta}\) are the standard deviations of the shape and expression morphable model, respectively. Our full loss functions for geometry reconstruction can be written as:

\[
L = \omega_{l_{11}} \cdot L_{l_{11}} + \omega_{\text{per}} \cdot L_{\text{per}} + \omega_{\text{coef}} \cdot L_{\text{coef}} + \omega_{\text{light}} \cdot L_{\text{light}} + \omega_{\text{ldm}} \cdot L_{\text{ldm}}
\]

(14)

In the geometry reconstruction experiments, we set \(\omega_{l_{11}}, \omega_{\text{per}}, \omega_{\text{coef}}, \omega_{\text{light}}, \omega_{\text{ldm}}\) to 2, 0.2, 0.001, 10, 0.002, respectively. We train the geometry reconstruction network on the FFHQ \([12]\) dataset for 20 epochs.

As shown in Table 1, our method just obtains similar quantitative results compared to the BFM09 under the same CNN-based face geometry reconstruction pipeline. However, we believe that our model has the potential to achieve better geometry reconstruction results with the advance of lighting estimation and differentiable ray tracer.

B.3. Face Relighting and OLAT Rendering

See our project page for the video results.

C. Limitations and Discussions

Our method still has several limitations. We adopt the Lambertian BRDF to represent diffuse reflectance. Thus, we cannot model the subsurface scattering effect. Integrating a more complicated reflectance representation \([23]\) into our morphable face reflectance model to improve face rendering realism is an interesting direction.

Our model cannot well represent the specularities around the eyes. We try a straightforward way by adding more mirror-like specular terms in our reflectance representation but find it does not work. We attribute this to the following two reasons: i) the reconstructed geometry is inaccurate around eyes during inverse rendering, and ii) our BRDF reflectance representation cannot well model the complex properties of eyes (e.g. refraction).
During model finetuning, we use a differentiable rasterizer with an efficient local shading technique to render the reconstructed image, without considering global illumination effects like self-shadowing, considering that the illumination is soft, and the self-shadows are insignificant in most in-the-wild images. We believe that using a differentiable ray tracer [13] would slightly improve the current results as demonstrated in existing works [6–8]. Moreover, leveraging a multi-view in-the-wild face image dataset [2] or video dataset [3] could improve the face reconstruction results, as demonstrated by the previous works [9, 19]. We leave these as our future works.

In addition, there is an inevitably global scale between the reflectance parameters in our model and the ground truth since the low-cost data does not provide lighting information [15].

References

[1] 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. 4


