Geometry-aware Single-image Full-body Human Relighting Supplementary Materials

Chaonan Ji[1], Tao Yu[1], Kaiwen Guo[2], Jingxin Liu[3], and Yebin Liu[1]

[1] Department of Automation, Tsinghua University, China
[2] Meta Reality Labs
[3] Guangdong OPPO Mobile Telecommunications Corp., Ltd

1 Network Architecture

1.1 Geometry Module

Given an input image I, our geometry module performs image-to-image translation of the input image to its corresponding normal map $\widehat{N}$ and ambient occlusion map $\widehat{AO}$ using a U-Net architecture similar to pix2pixHD [11]. We empirically found that training geometry networks separately produces more accurate results than joint training. The losses for geometry estimation can be expressed as follows:

$$L_N = \lambda_{geo}\left\|\widehat{N} - N\right\|_1 + Vgg(\widehat{N}, N)$$

$$L_{AO} = \lambda_{geo}\left\|\widehat{AO} - AO\right\|_1 + Vgg(\widehat{AO}, AO)$$

where L_N is the normal loss, L_{AO} is the ambient occlusion loss and Vgg is the Vgg loss. AO is the ground truth of ambient occlusion map and N is the ground truth of normal map. λ_{geo} is the weight factor and $\lambda_{geo} = 5$. Note that we add skip connections for the ambient occlusion map inference network to restore more details.

1.2 De-lighting Network

Fig. 1 shows the architecture of the de-lighting network. We remove the the downsampling operations of the first stage in HRNet [10] directly (which also avoid skip connections) to fuse multiscale features while maintaining high-resolution representations. Our de-lighting network contains 6 stages and the number modules of last five stages are [1,2,2,2,2]. The "NUM BLOCKS" is set to 2 for the last five stages and the last layer is followed by an extra 1×1 convolution layer to generate output. The convolution stride of the first stage is set to 1 and the upsample mode is set to bilinear. "Ours(5 stages)" only contains 5 stages and the rest of the network is the same as "Ours". "HRNet(w extra parameters)" is adpated from HRNet-W32 [10] and owns 6 stages (the number modules of last five stages are [1,4,3,2,1]). The "NUM BLOCKS" is set to 4.

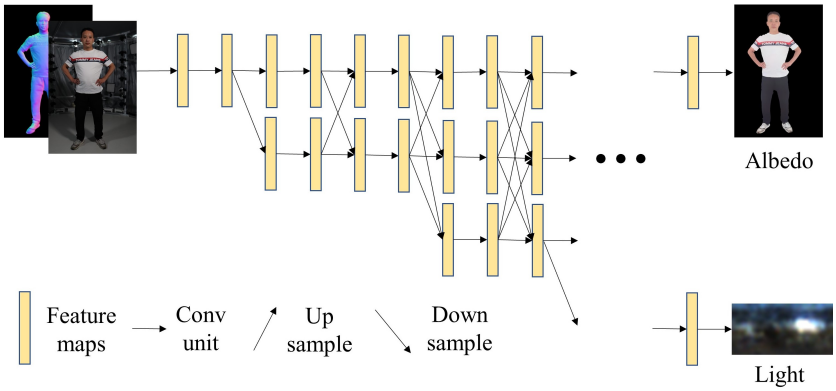

Fig. 1. The architecture of the de-lighting network. The de-lighting network takes the input image and inferred normal map as input, and outputs albedo at the layer with the highest resolution of the final stage and light at the layer with the lowest resolution of the final stage.

1.3 Full-body Refinement Network

Fig. 2 shows the architecture of the full-body shading refinement network. The network is modified from MIMO-Unet [3] and is deepen to 4 downsampling operations. It takes multi-scale input images as input and outputs multi-scale refined full-body shading maps. We concatenate the ray-traced shading map S^{body}_{coarse} and inferred ambient occlusion map $\widehat{AO}$ as the input to the highest-resolution convolution layer and the downsampled ambient occlusion maps are used as the inputs to low-resolution convolution layers. The FAM, SCM and AFF modules are the same with the MIMO-UNet [3].

1.4 Face Refinement Network

Fig. 3 shows the architecture of the face shading refinement network. The network takes the cropped face from the refined full-body shading map $\widehat{S^{body}_{fine}}$ and ray-traced face shading map S^{face}_{coarse} as input, and outputs the refined face shading map. The input is resized to 128×128. The network has 8 encoder-decoder layers and skip connections and each layer is run through 3×3 convolutions followed by LeakyReLU activations and BatchNorm (7×7 convolutions for the first and last layer, and the last layer is followed by an extra 3×3 convolution layer to generate output). The number of filters are 64,128,256,512 for the encoder,512 for the bottleneck, and 256,128,64,64 for the decoder respectively. Each convolution layer except the last layer is followed by a residual block and there are 3 residual blocks in the bottleneck.

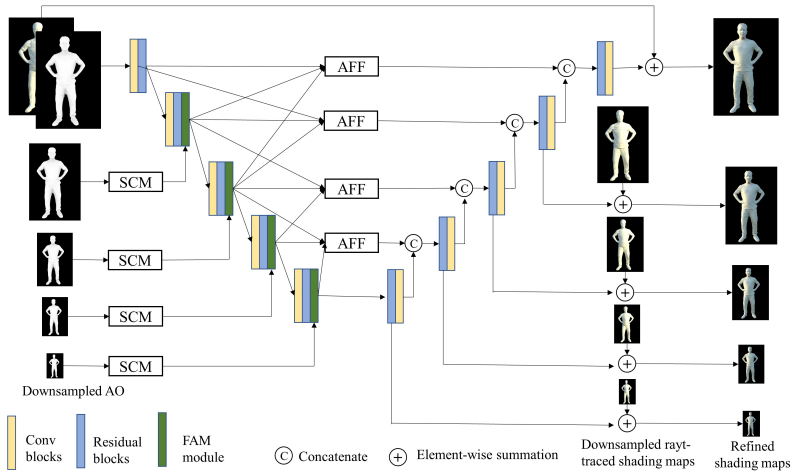

Fig. 2. The architecture of the full-body shading refinement network.

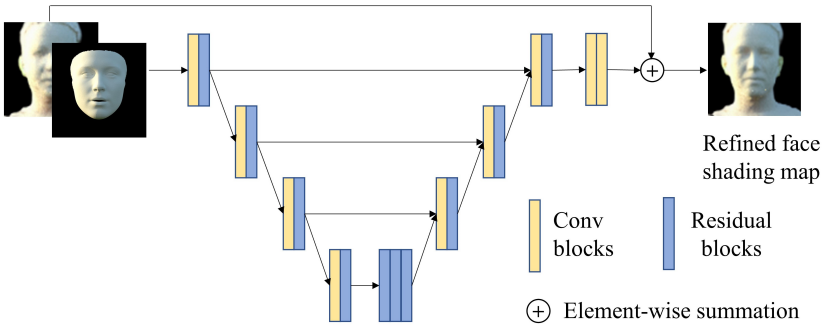

Fig. 3. The architecture of the face shading refinement network.

2 Implementation Details

2.1 Data

We use Cycle rendering engine in Blender [1] and a Principled BSDF shader to render dataset. For each model collected from Twindom [2], we set the pitch angle in [0,10,20,30], yaw angle in [-32,-24,-16,-8,0,8,16,24,32], and model scale in [0.8,1,1.1] and render the model under two different random lighting conditions for every setting. As shown in Fig. 4, each model can produce 216 sets of images where each set consists of the rendered image, albedo map, normal map, ambient occlusion map and corresponding mask. The human models are collected with well-behaved lighting conditions and the color of the texture is treated as ground truth albedo. To obtain the ambient occlusion map, we add extra output node in Cycles engine [1] and enable "AO pass" during rendering. In total, we generate

Fig. 4. Examples from our human image dataset. For each human model, from left to right are the rendered image, mask, albedo map, shading map, normal map and ambient occlusion map.

150k sets of images for training and 20k sets of images for testing. The resolution of the rendered images is 512×512. All training data are augmented using random blur, noise and color enhancement. The Twindom dataset we used is restricted to make public and we will release pretrained model based on the THUman 2.0 dataset [12] for research purposes.

2.2 Training and Testing Details

We train geometry networks, de-lighting nerwork and shading refinement networks separately in PyTorch. For all models that require training, we use the Adam optimizer and set learning rate to 1e-5. The learning rate for all discriminators is set to 1e-6. The size of inputs for geometry networks, de-lighting network and full-body shading refinement network is 512×512. For the face refinement network, all inputs are resized to 128×128.

All the training processes are conducted on a 4-RTX2080Tis-Server, which consists of: 1) 3D reconstruction using PIFuHD [9], 3DMM fitting [5] and ray-tracing take 2 days in total and 2) Training of Geometry Module, Albedo Module and Shading refinement module (body/face) takes 2 days, 4 days and 2 days respectively. The testing (on a single RTX2080Ti PC) efficiency is 17s per image which includes: 120 ms for estimating albedo, 14s for 3D reconstruction, 3s for ray-tracing and 50 ms for shading refinement.

For every 3D human model, there are two sets of rendered images under different lighting conditions for every pose. When testing, we use one set of images as input and predict the relit image under the lighting condition of another set. Therefore, we can compare the difference between the inferred relit result and the ground truth.

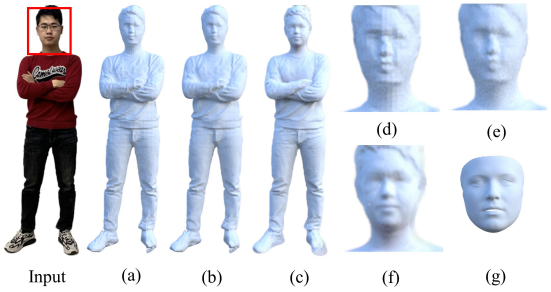

Fig. 5. Ray tracing results under the target illumination condition for the 3D model estimated from the input image. (a) Ray-traced shading map without smoothing (b) Ray-traced shading map with smoothing (c) Refined shading map (d) Cropped face from the ray-traced shading map without smoothing (e) Cropped face from the ray-traced shading map with smoothing (f) Cropped face from the refined shading map (g) ray-traced 3DMM face shading map

2.3 Ray-traced 3D Model Smoothing

The Fig. 5 shows the checkered artifacts mentioned in Section 5.1. The resolution of the estimated 3D model by PIFuHD [9] is low, and direct ray tracing may produce grid-like artifacts. We use Laplacian smoothing to smooth the surface of the estimated full-body model and improve the quality of the ray-traced shading map. The ray-traced 3DMM face shading map is shown in Fig. 5 (g). Compared with cropped faces from the ray-traced full-body shading maps, it owns clear facial geometry details and shadows.

2.4 Details of Comparisons with SOTA Methods [6, 4, 7]

All input images are resized to 512×512. Since all three methods use spherical harmonic lighting representation (but with different orders), we extract 25 SH coefficients for every HDR environmental lighting map, and use the first 9 SH coefficients for testing RH [6] and RHW [4] and all 25 SH coefficients for testing SFHR [7].

For the comparison with RH [6], we re-render our dataset using spherical harmonics lighting and pre-computed radiance transfer(PRT) to ensure that all the 3D models and lighting are the same as in our dataset. Experimentally we find that the model of RH [6] retrained on our larger dataset demonstrates superior performance. For a fair comparison, the qualitative results of RH [6] come from both the retrained model and the official pretrained model. We directly test the released models of RHW [4] and SFHR [7] on our testing dataset. Since these three methods require the brightness of target lighting to lie within [0.7,0.9], the relit results will fail under harsh illuminations and large areas of white pixels appear in the relit shading map. Therefore, we scale the pixel values of their relit shading to [0,1] by dividing the relit shading by its maximum value instead of clipping it to [0,1].

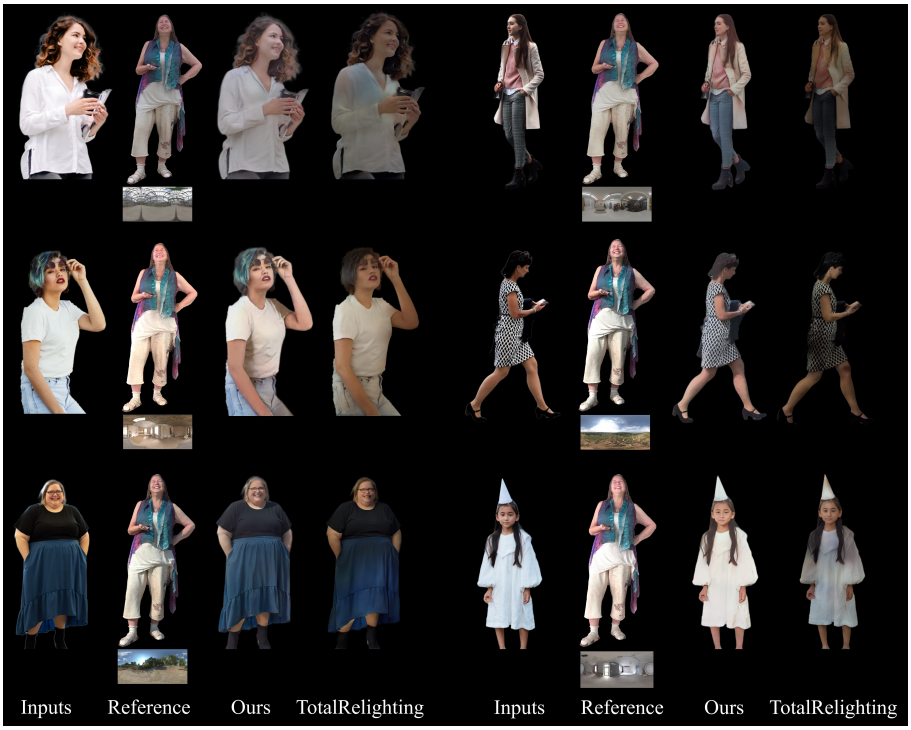

Fig. 6. Relit results on real-world images. "Reference" are the rendered images of a virtual 3D human model under the target lighting conditions and are used to indicate the position of shadows. The target HDR environment map is placed under the reference image.

3 Comparison with TotalRelighting [8]

We show results in comparison with TotalRelighting [8], which is trained on real OLAT dataset. As shown in Fig. 6, the TotalRelighting may produce patchy shadows and inaccurate albedo inference for clothing. By contrast, our method is able to generate photo-realistic relit results with high-frequency self-shadows (including hard cast shadows).

References

1. https://www.blender.org/
2. https://web.twindom.com/
3. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4641–4650 (2021)
4. Daichi Tajima, Yoshihiro Kanamori, Y.E.: Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. Computer Graphics Forum (Proc. of Pacific Graphics 2021) **40**(7), 205–216 (2021)
5. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. pp. 152–168. Springer (2020)
6. Kanamori, Y., Endo, Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. ACM Transactions on Graphics (TOG) **37**(6), 1–11 (2018)
7. Lagunas, M., Sun, X., Yang, J., Villegas, R., Zhang, J., Shu, Z., Masiá, B., Gutierrez, D.: Single-image full-body human relighting. CoRR **abs/2107.07259** (2021), https://arxiv.org/abs/2107.07259
8. Pandey, R., Escolano, S.O., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., Fanello, S.: Total relighting: learning to relight portraits for background replacement. ACM Transactions on Graphics (TOG) **40**(4), 1–21 (2021)
9. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 84–93 (2020)
10. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703 (2019)
11. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
12. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021)