Multi-level Attention Aggregation for Aesthetic Face Relighting

1. Synthetic dataset for training

As described in the main paper, we created a training dataset using 8 synthetic 3D human models. We varied the pose, location and orientation of these 3D human models w.r.t the camera to improve the dataset diversity. Fig 1 highlights a few of the variations observed in the training dataset. We trained the model on 21,000 images created using 7 3D human models and validated it on 3,000 images created from one other 3D human model (rightmost images on the last row). We used 3 female and 4 male 3D models for training. We can observe the diversity in position, pose and illumination of the foreground subjects across different images. Further, we can also observe the diversity of the 7 synthetic models in terms of gender, ethnicity, face pose, facial hair, facial structure, hair colour, etc. We created the training dataset by combining these variations in the training images with our novel dataset composition strategy (described in Section 3 of the revised paper). During training the model learns the relationship between the face location & orientation, light source position and the expected relit image. This ensures that our relighting model is able to generalize to real images from vastly different data distribution despite being trained on a synthetic dataset created from a limited number of 3D human models.

2. Relighting results

To fit the paper within the limits, we had to reduce the size and resolution of the images in Fig 6 of the revised paper. We show results on test images at higher resolution in Fig 2. We can observe that both our stage 1 and stage 2 models are able to estimate accurate and photo-realistic shadows. However, the stage 2 model renders sharper shadows and corrects small differences in skin colour[1] (tone) to further improve the photo-realism of the estimated relit image.

Further, we show the generalization of our model to different types of input images (brightness variations), facial structures, facial hair and complex lighting conditions. Our model is able to estimate accurate shadows and render aesthetic/photo-realistic images. Our model provides flexibility to smoothly control the light intensity on the face. Thus, it can be tweaked based on user preference.

A small limitation of our method is that the estimated relit image smoothes out the face (visible in the nose region in Fig 2) on some test images. This is because the training dataset consists of images generated from synthetic models which have very smooth skin textures. These issues do not affect the photo-realism of the estimated relit image, and thus we plan to address it in future work by possible adding a few synthetic 3D models with less smooth skin texture in the training dataset.

3. Ablation study results

Fig 3 shows the qualitative results on a few images for the ablation study shown in Table 2 of the revised paper. We found that the qualitative result back up the quantitative metrics. We observed that without attention layers the relit image lacks sharpness in shadows. The shadows are sharper with attention layers at higher levels (HLA) as compared to attention layers at lower levels (LLA). Without global loss (GL) the shadows are slightly more sharper than without local loss (LL). However, without local loss leads to better colour (skin tone) rendering of the estimated relit image. Each design choice contributes something unique, however we can observe that the best results are observed with our full stage 1 model.

[1] Visible differences in the colour/skin tone are mainly due to the light falling on the surface (face / shirt).

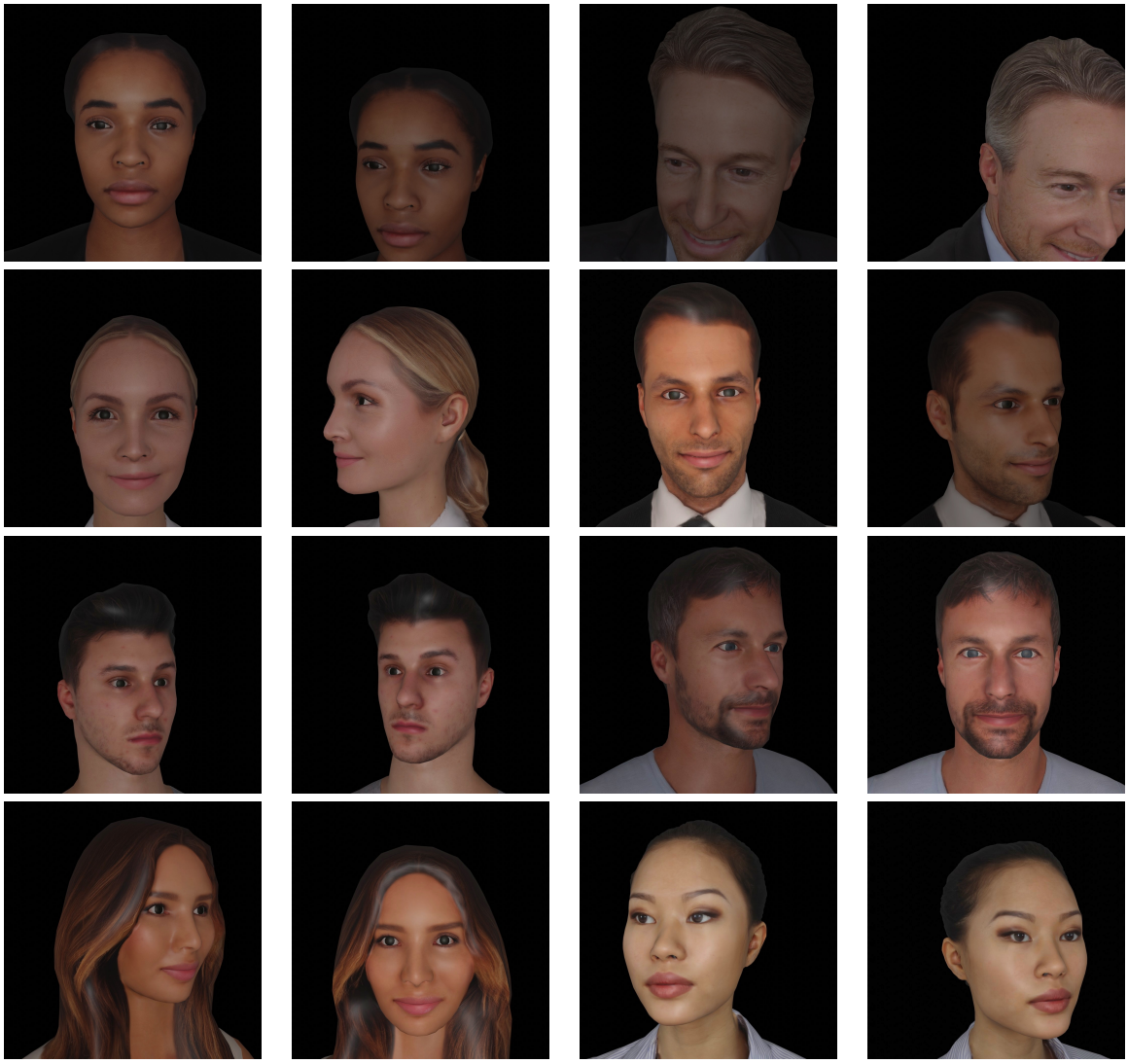
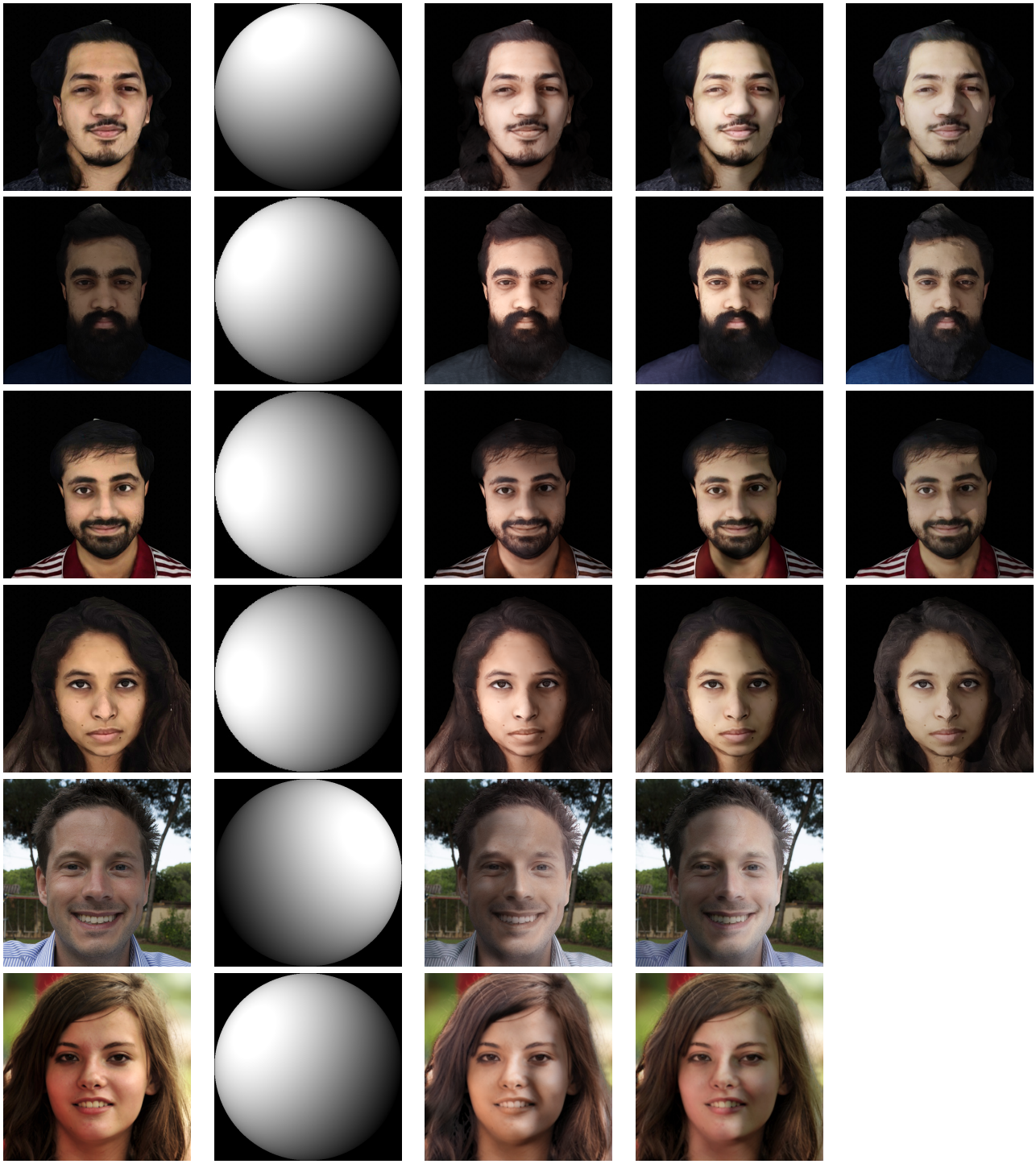
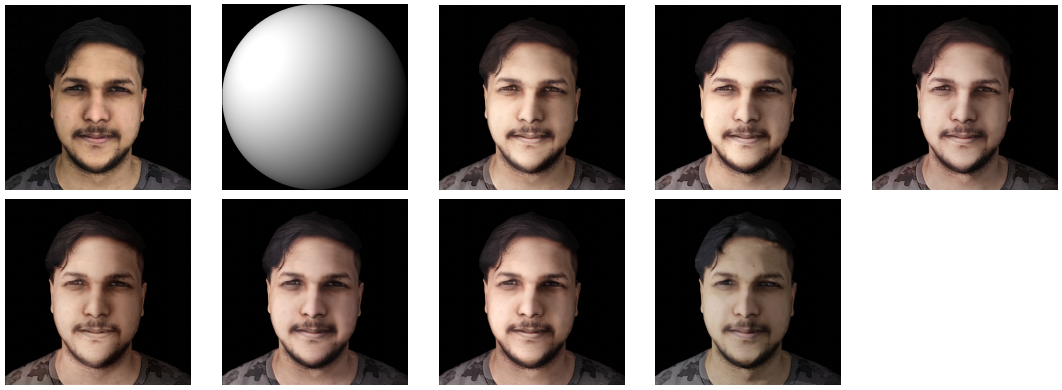


Figure 1. Sample input images used for training our relighting model. The two rightmost images on the last row are from the the validation dataset, while rest of the images are from the training dataset. Images are best viewed in colour.

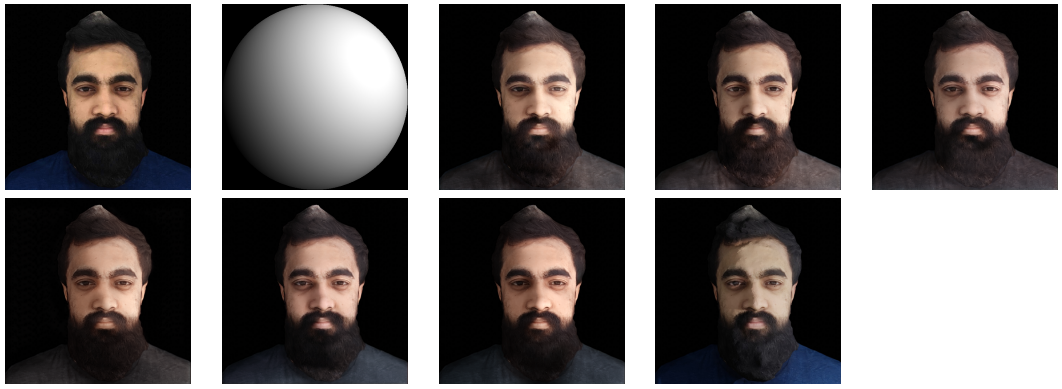


(a) Input image (b) Target lighting (c) Ours (stage1) (d) Ours (stage2) (e) Ground truth

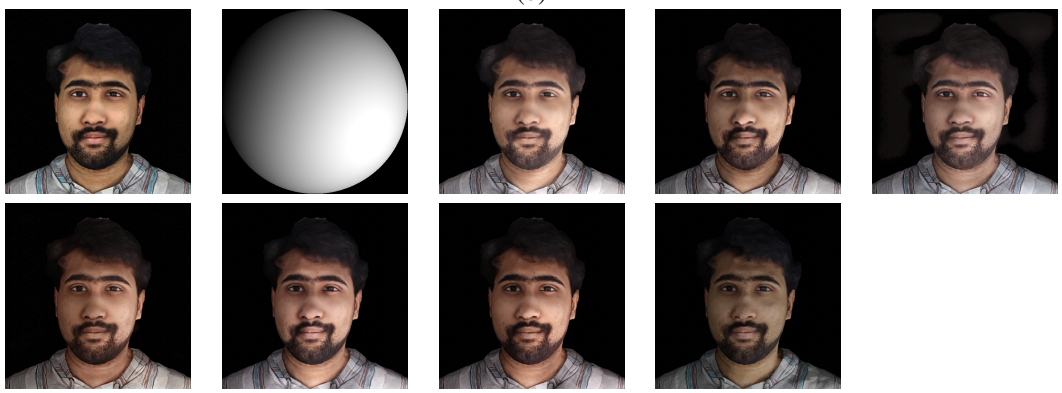
Figure 2. Qualitative results from stage 1 and stage 2 of our relighting models. Rows 1-4 are images from our real human test dataset, while Rows 5-6 are images from Celeb-FFHQ dataset for which we do not have the ground truth relit image. Images are best viewed in colour.



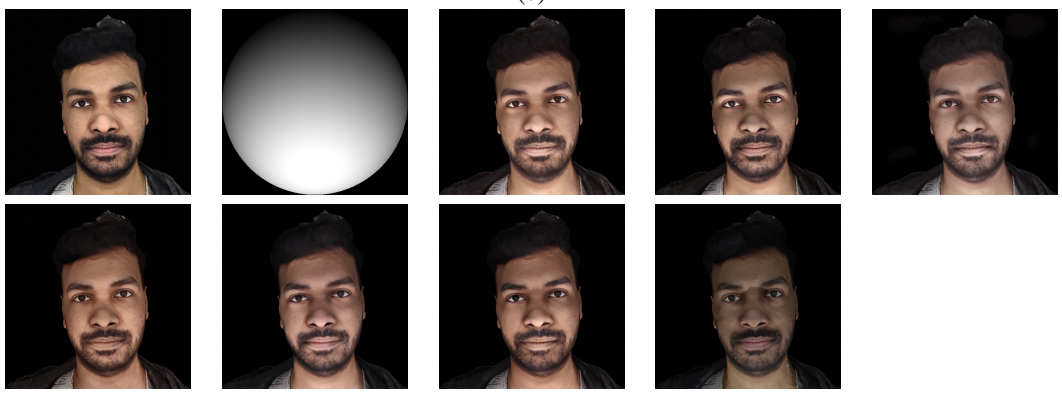
(a)



(b)



(c)



(d)

Figure 3. Qualitative results of the ablation study shown in Table 2 of the revised paper. Left to right for each row: Input image, light source position, without attention, with HLA, with LLA, without GL, without LL, full model (stage 1), ground truth. Images are best viewed in colour.