RelightAnyone: A Generalized Relightable 3D Gaussian Head Model

Supplementary Material

6. Implementation Details

Network Details. Our 2D convolutional decoders, *i.e.*, $\mathcal{D}_g$, $\mathcal{D}_c$, $\mathcal{D}_{ci}$ and $\mathcal{D}_{cv}$, have a nearly identical architecture, differing only in their specific input/output layers and skip connections. The input vector is first linearly mapped and reshaped into an initial feature map $\mathbf{z}' \in \mathbb{R}^{256\times8\times8}$ (channels $\times$ height $\times$ width). Then, at each layer, it is progressively upsampled by a factor of two until it reaches the final 1024×1024 resolution. All intermediate layers are followed by LeakyReLU activations. $\mathcal{E}$ is a mirrored version of $\mathcal{D}_{ci}$ and $\mathcal{D}_{cv}$. We apply specific activation functions to the final output: a softplus function for the Gaussian scales $\mathbf{s}_k$, a sigmoid function for opacity o_k and specular visibility v_k, and an exponential function for the roughness σ_k. Gaussian colors are clamped to be non-negative before splatting.

Training Details. We set the loss balancing weights as follows: $\lambda_{\text{l1}} = 10$, $\lambda_{\text{ssim}} = 0.2$, $\lambda_{\text{geo}} = 0.4$, $\lambda_s = 0.01$, $\lambda_{c\_} = 0.01$, $\lambda_{\text{mono}} = 0.01$, $\lambda_{\text{id}} = 0.01$, and $\lambda_{\text{lr}} = 1$. Several weights are linearly annealed: λ_t is initialized at 1 and decreased to 0.001 by iteration 20000; λ_n is initialized at 1 and decreased to 0 by iteration 5000; λ_ρ is initialized as 10 and decreased to 0.01 by iteration 10000. We use the Adam optimizer with a learning rate of $1e^{-3}$ to for Stage 1, and $5e^{-4}$ for Stage 2 and model fitting. A batch size of 16 is used for both stages. Both Stage 1 and Stage 2 models are trained for one day on 4 Quadro RTX 6000/8000 GPUs. The model fitting process, including both the inversion and finetuning steps, typically converges within 3000 iterations, taking approximately 30 minutes on a single GPU.

7. Additional Experiments

Novel View Comparison. Fig. 10 compares our novel view synthesis with 3D GAN-based methods: NeRFFace-Lighting (NFL) [25], Lite2Relight [59] and 3DPR [60], which are trained only on 2D portrait collections, often produce distorted or "stretched" results for side poses. Moreover, they cannot be applied trivially to multi-view inputs of the same subject, as they typically encode each view into a different latent vector. In contrast, our method learns an explicit volumetric representation directly from multi-view data, resulting in better view-consistency. We note that when fitted to a single image, our method degrades only slightly in side poses, particularly in reconstructing the ears and the facial silhouette. Our results also better preserve the identity (see Fig. 8 for a real photo of this subject).

Figure 10. Novel view comparison with 3D GAN-based relighting methods: Since the baseline methods are trained only on 2D portrait collections, they struggle with view-consistency at the side poses. Furthermore, since the baselines can only take a single image input, we also show our method fitted to only a single image which slightly degrades the quality in side poses.

Figure 11. Effect of finetuning. While optimizing only the identity code and lighting produces a rough likeness, finetuning our model recovers high-frequency, person-specific details that better match the ground truth input.

Effect of Finetuning. Fig. 11 demonstrates the effect of finetuning on a single in-with-wild input image. Without finetuning, *i.e.*, optimizing only the identity code and the scene lighting (see "w/o finetuning" column), the rendered image captures only a rough likeness of the subject with low-frequency appearance. By finetuning the Stage 1 model, we can capture more person-specific details, resulting in a rendered image more closely matches the ground truth.

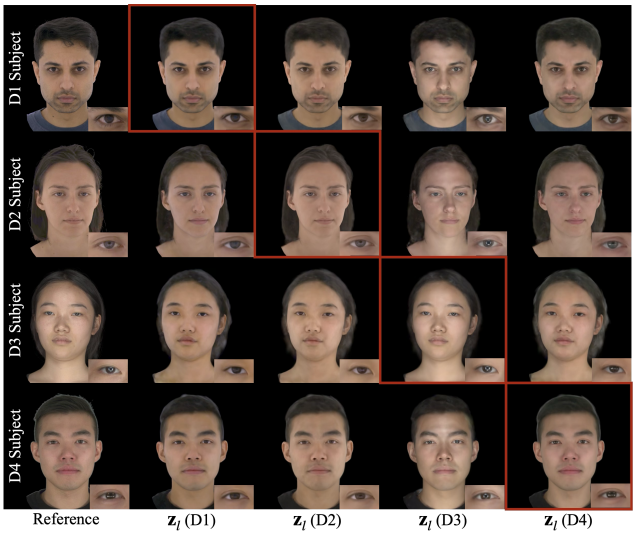

Figure 12. Self-supervised lighting alignment. Our model learns a distinct lighting code $\mathbf{z}_l$ for each dataset. Each row shows a subject rendered with the lighting codes from different datasets (D1-D4). Red boxes indicate the subject's original dataset. Note how our model captures dataset-specific lighting variations, especially in the specular reflections on the nose and in the eyes. Image best viewed zoomed-in.

Lighting Alignment. Our model enables self-supervised lighting alignment by introducing a dataset-specific lighting code $\mathbf{z}_l$. As shown in Fig. 12, each row corresponds to a subject from a different dataset (D1, D2, D3 and D4, from top to bottom). The first column shows the ground truth images for reference, and the subsequent columns show fully-lit renders generated using the lighting codes from each of the four datasets. Red boxes indicate the original lighting condition for each subject. Although all datasets are generally evenly-lit, our $\mathbf{z}_l$ code successfully learns their subtle, distinct lighting distributions. For example, subjects in dataset D3 are lit from four frontal flashes, resulting in stronger specular highlights in the central part of the face. Our model correctly captures this specific effect when applying the D3 lighting code. Similarly, while dataset D4 is also front-lit, it has a light background that reflects light from behind, making it closer to D1 and D2 (which are lit from 360 degrees). Even so, our model is able to capture the subtle differences in specular reflections on the nose and in the eyes.

Effect of $\mathcal{L}_\rho$ and $\mathcal{L}_{\text{mono}}$. Fig. 13 demonstrates the effect of the regularization terms $\mathcal{L}_\rho$ and $\mathcal{L}_{\text{mono}}$, which we introduced to enforce a meaningful decomposition of the diffuse albedo and the diffuse shading. Without these regularization terms, the final render may look plausible, but the underlying albedo and diffuse shading components exhibit severe color artifacts. This occurs because the model gets stuck in a local minima, which it cannot escape as training proceeds. Our loss terms are designed to prevent this, guiding the optimization toward a correct decomposition.

Figure 13. Effect of $\mathcal{L}_\rho$ and $\mathcal{L}_{\text{mono}}$ on intrinsic decomposition. Without our regularization (top row), the model produces a plausible render but fails to properly disentangle albedo and shading. Our full model (bottom row) achieves a clean and physically meaningful intrinsic decomposition.

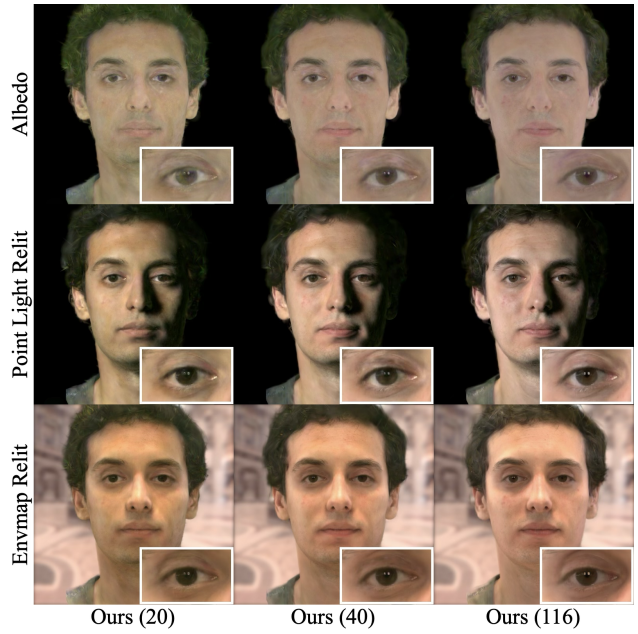

Figure 14. Effect of OLAT dataset size. The figure ablates the number of OLAT subjects used to train our Stage 2 network (20, 40, and 116 subjects, from left to right). While all models render the correct lighting distribution, training with more subjects produces a cleaner albedo and finer details which better preserves the subject's identity. Refer to Fig. 4 for a real photo of this subject.

Training with less OLAT Data. To evaluate the impact of the OLAT dataset size, we train our Stage 2 relighting

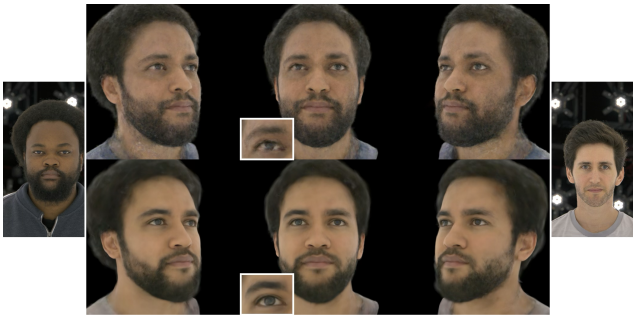

Figure 15. Novel view renders of an interpolated identity. Top row: Our Stage 1 model trained only on dataset D1 exhibits high-frequency artifacts. Bottom row: Our model trained on all four datasets yields a cleaner appearance. The references images for the two source identities used for interpolation are shown in the leftmost and rightmost columns.

network on subsets of 20, 40 and full 116 subjects in dataset D1. Fig. 14 shows their fitting results on a single in-the-wild image. We note that all models successfully capture the correct lighting distribution, including similar shadows and highlights, which demonstrates the strong generalizability of our model, even when trained with minimal OLAT data. However, adding more OLAT data improves the results: the model trained on more subjects learns a cleaner albedo and better preserves identity (*e.g.*, correct skin tone) and finer details.

Training with less Full-On Data. Combining multiple existing flat-lit datasets improves the quality of the identity latent space and the learned multi-view prior. We demonstrate this in Fig. 15 by visualizing novel view renders of an interpolated identity. We compare a model trained only on dataset D1 (top row of Fig. 15) to our full model trained on all four datasets (bottom row of Fig. 15). We can see that training with the combined datasets produces a "cleaner" and more plausible new identity. In contrast, the ablated model (trained on D1 only) exhibits significant high-frequency artifacts, indicating a less robust latent space.

Failure Cases. Finally, we show some failure cases of our method. The first type of failure is associated with accessories, such as the headscarf and glasses shown in Fig. 16. Because the OLAT dataset (*i.e.*, D1) does not contain these accessories, our model cannot infer their relightable parameters. As a result, the patterns on the relit headscarf appear blurred, and the glasses lack specular reflections. We note that this is also a limitation of RGCA, as its appearance model is designed for the human head and does not work well on the diverse materials found in accessories.

Second, our model struggles with the reconstruction and relighting of some hairstyles. We show an example in Fig. 17,

Figure 16. Failure case: accessories. Our model fails to infer relightable parameters for items like headscarf and glasses.

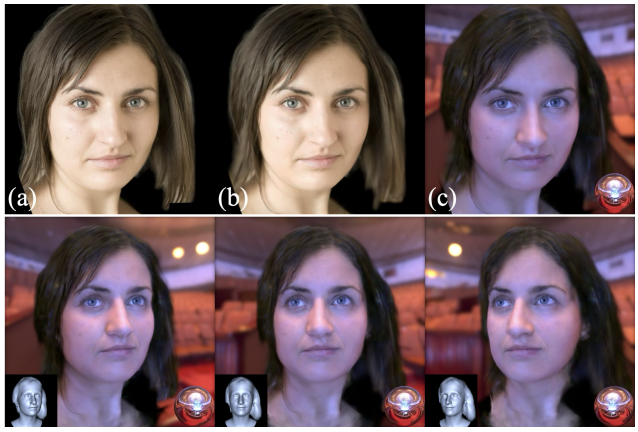

Figure 17. Failure case: open long hairstyle. Given an in-the-wild image of a subject with open long hair (a), the model fitting (b) and the original-view relit (c) appear plausible, but the hair appears as a texture-less cloud with color artifacts when rendered from novel views (bottom row). The corresponding tracked meshes are shown in the corner. Image best viewed zoomed-in.

where our model can be fitted closely to a subject and relight them plausibly from the original camera view, but the hair appears as a texture-less cloud. This is especially visible when rendering novel views under new environment lighting, where some Gaussians also exhibit distracting color artifacts. There are several causes: first, although the VHAP [54] face tracker deforms the FLAME template to cover the hair, the results are sometimes poor for subjects with long hair (see inset). Second, these inaccurate tracking results lead to bad UV correspondences, making it difficult to learn a universal relightable prior for various hairstyles. Third, FLAME UV parameterization compresses the hair region into a small area on the UV map, allocating an insufficient number of Gaussians to represent the intricate structures.

8. Ethics

All individuals portrayed in this paper provided informed consent for the use and publication of their images for research purposes.