Supplementary Materials for
Relightable Neural Human Assets from Multi-view Gradient Illuminations

Taotao Zhou[1,4*] Kai He[1*] Di Wu[1,3*] Teng Xu[1,4] Qixuan Zhang[1,3]
Kuixiang Shao[1,4] Wenzheng Chen[2] Lan Xu[1†] Jingyi Yu[1†]
[1]ShanghaiTech University [2]University of Toronto [3]Deemos Technology
[4]LumiAni Technology
{zhoutt, hekai, wudi, xuteng, zhangqx1, shaokx, xulan1, yujingyi}@shanghaitech.edu.cn
wenzheng@cs.toronto.edu

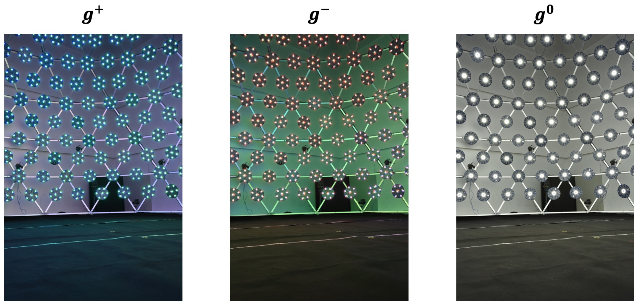


Figure 1. We show 3 illuminations in our capture settings. From left to right are positive gradient light, negative gradient light, and white light.

A. System Calibration and Capturing Settings

In this section, we provide more details about system calibration and capturing configurations. We first describe how to perform photometric calibration in Sec. A.1. Then we explain gradient illuminations in Sec. A.2. Lastly, we analyze the albedo acquisition in Sec. A.3.

A.1. Light Color Calibration

We conduct both geometric and photometric calibration for our system. Below we briefly talk about how we perform photometric calibration to reproduce user-specific lighting conditions within our system. Specifically, the LED beads in our system support six illumination colors (RGBWAC), denoted as $L_i, i \in \{1, 2, 3, 4, 5, 6\}$. Given an arbitrary 3-channel illuminations $P_j, j \in \{1, 2, 3\}$, our goal is to simulate P with the linear combinations of L.

We assume the light is linear and fulfills the superposition principle. Thus, each channel in P can be represented by the linear combination of L, where we calibrate coefficients a s.t. $P_j = a_{i,j} L_i$. Once the coefficients a are solved, we can represent any 3-channel illumination with the LED beads. Here, the LED beads adopt six illumination colors to obtain a more comprehensive color spectrum and therefore have a better capacity to approximate the given lighting conditions. Similar to [7], we utilize a precisely calibrated color chart and solve a non-negative linear equation to calibrate a. For more details, please refer to [7].

A.2. Gradient Illuminations

Following [3, 4, 8], we use the popular gradient illuminations to estimate the surface normal and corresponding surface reflectance properties. Assuming $L^0 \in \mathbb{R}^3$ is the equalized maximum lighting intensity, the positive gradient illuminations are defined as:

$$L = (\frac{1}{2} + \frac{1}{2}\Theta)L^0 \quad , \tag{1}$$

where $\Theta \in [-1, 1]^3$ is the lighting direction and $L \in \mathbb{R}^3$ is the corresponding RGB lighting intensity at Θ. We illuminate the scene with 3 different lights: positive gradient illumination, negative gradient illumination, and white light. The negative gradient illumination is defined in a similar way but with negative direction: $L = (\frac{1}{2} - \frac{1}{2}\Theta)L^0$, and the white light is defined as $L = L^0$, which means we turn all lights on. See Fig. 1 for all 3 illuminations.

We then record the 3 pixel values captured under 3 illuminations as $g^+ \in \mathbb{R}^3$, $g^- \in \mathbb{R}^3$, and $g^0 \in \mathbb{R}^3$, respectively. For more efficient data acquisition, we adopt colored gradient illuminations, encoding the gradients into RGB channels, similar to the approaches by Ma *et al.* [8] and Guo *et al.* [4].

[*] Equal contribution.
[†] Corresponding author.

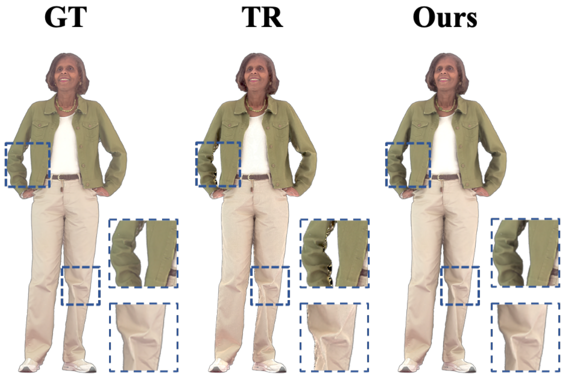


Figure 2. We qualitatively compare the estimated albedo between the method of Guo *et al*. [4](denoted as TR), and our method on a synthetic example. TR produces strong artifacts in occluded regions, mainly due to the inaccurate mesh normal estimation. Instead, we approximate the albedo map from the image captured under white light, which generates smoother and more accurate results.

A.3. Albedo Acquisition

In photometric stereo (PS), albedo can be estimated simultaneously with normal [5, 8] from gradient illuminations. However, in practice, we find the jointly recovered albedo has severe artifacts, especially for the occluded regions. Therefore, we choose to acquire albedo under white light. Below we show the comparison between jointly estimated albedo and our white-light albedo.

Jointly estimating albedo under gradient illuminations. Similar to our settings, Guo *et al*. [4] propose a method to estimate surface albedo $\mathbf{a}$ from two gradient illumination measurements g^+ and g^-:

$$\mathbf{a} = \frac{g^+ + g^- - (r_0, r_0, r_0)}{(1-o)(1-r_0)}, \tag{2}$$

where $r_0 = 0.04$ approximates the dielectric Fresnel term at normal incidence, $o \in [0, 1]$ is the ambient occlusion term. Equation (2) assumes that the sum of two gradient illuminations $g^+ + g^-$ contains the albedo at each pixel. It further adds r_0 and o to account for the Fresnel effect and the ambient occlusion term. o is defined as:

$$d = \frac{g^+ - g^-}{g^+ + g^-} \tag{3}$$

$$\beta = \frac{3}{2}(|d| - \frac{1}{3}) \tag{4}$$

$$\alpha = min(1, cos^{-1}(\mathbf{n}, \mathbf{n}^m)) \tag{5}$$

$$o = \beta^\alpha \tag{6}$$

Method	PSNR↑	SSIM↑	RMSE↓
Ours	**30.761**	**0.975**	**0.029**
TR	29.840	0.951	0.032

Table 1. We quantitatively evaluate the estimated albedo between the method of Guo *et al*. [4](denoted as TR), and our method on a synthetic example. Our method has fewer artifacts in occluded regions, resulting in better albedo estimation.

Here, $\mathbf{n}$ is the computed photometric surface normal (paper Eq. (1)), and $\mathbf{n}^m$ is the mesh normal, where the mesh is reconstructed by MVS and depth cameras. Intuitively, Guo *et al*. [4] assume that surfaces with a larger angle difference between the photometric surface normal $\mathbf{n}$ and mesh normal $\mathbf{n}^m$ will have a larger ambient occlusion term. It works well in their settings, potentially due to their depth cameras enabling accurate mesh reconstruction.

Albedo acquisition with white light. However, we find that directly applying the method of Guo *et al*. [4] to our system results in poor albedo estimation, as shown in Fig. 2 and Tbl. 1, mainly due to the varying capture settings. The method of Guo *et al*. depends on accurate geometry extraction to compute the correct occlusion term. While they employ depth cameras to fulfill the requirement, in our settings, we adopt multi-view normal maps to train an SDF field to extract the geometry. Due to the over-smoothed network prediction, the calculated occlusion term is rather noisy, especially for the regions containing sharp edges. The inaccurate occlusion term will also produce problematic albedo estimation and hamper the final rendering results.

Instead, the image captured under white illumination is naturally a better approximation in our settings. White light helps produce minimum shadows on human bodies, largely reducing the influence of self-occlusions. Moreover, normal and albedo are entangled in gradient illuminations, while images captured in white light can preserve most albedo information. We qualitatively and quantitatively compare the two methods on a synthetic example, where we render a synthetic 3D human model with 32 cameras and 3 illuminations, the same as our capture settings. We demonstrate in Fig. 2 and Tab. 1 that our method produces more accurate albedo estimation.

The approximation of albedo using white-light images presents bake-in shading and visibility issues due to inherent limitations. To overcome this problem, we propose the use of multi-view depth-guided G-buffer reprojection. This approach enhances the robustness of our results by considering six views instead of relying on a single view which might yield incorrect results. In Fig. 3, we compare the re-rendered images and real capture under gradient illuminations, demonstrating our method's ability to accurately render complex texture and geometry details.

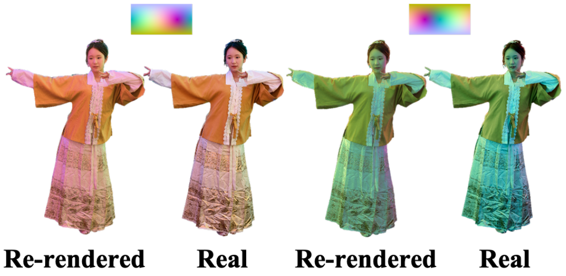


Figure 3. We compare the re-rendered images and real capture images under gradient illuminations. The complex texture and geometry details can be well rendered by approximating albedo with white-light images.

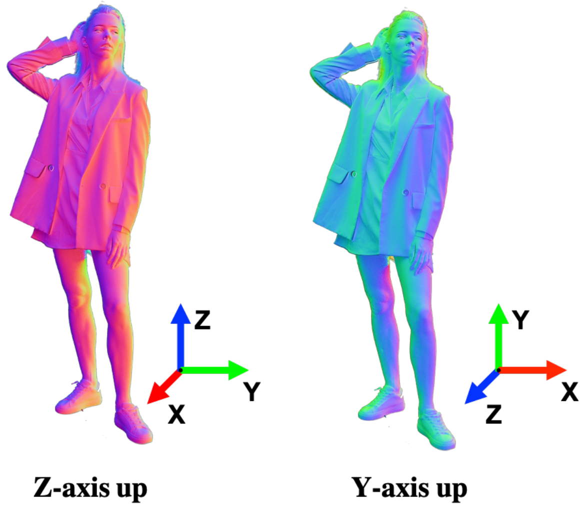


Figure 4. Illustration on the coordinate systems for normal maps. The left image demonstrates normal maps in a right-handed coordinate system with the Z-axis oriented upwards, while the right image displays a coordinate system with the Y-axis directed upwards.

Coordinate system of normal. The normal maps are calculated within UltraStage's world coordinate system, which is consistent with the camera and lighting systems. This right-handed coordinate system features the Z-axis oriented upwards. We convert XYZ values to RGB values within the range of 0-255. Typically, normal maps are represented in camera space, where the Z-axis points towards the camera and the Y-axis is directed upwards. However, to accommodate the multi-view capture setting, we opt to compute normal maps in world space. An illustration of normal maps can be found in Fig. 4, showcasing configurations with the Z-axis and Y-axis oriented upwards, respectively.

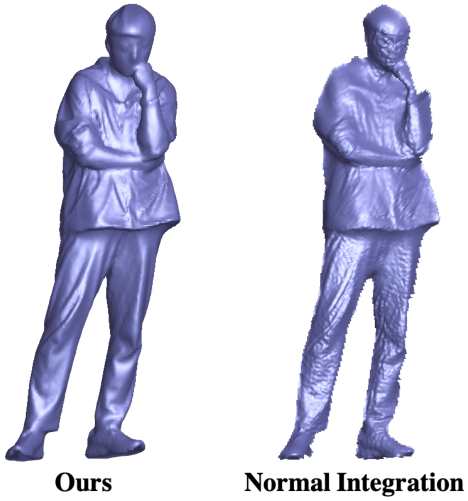


Figure 5. Qualitative comparison on geometry modeling between our approach and normal integration. Traditional normal integration generates a smoother and noisier result than the SDF field.

B. Dataset Overview

As mentioned in the main paper, UltraStage consists of over 2,000 static human poses. For each pose, we capture it by 32 surrounding cameras under three lighting conditions: color gradient illumination, inverse color gradient illumination, and white light. The three lighting patterns are switched at 5fps, so the subject is required to stay still for around 0.6 second. For each pose, we provide 96 images at resolution 6016×4016. We show more examples in Fig. 10, Fig. 11, Fig. 12, and Fig. 13.

C. High-quality Neural Geometry Modeling

In order to augment the mesh surface details, we have also explored the conventional normal integration technique. However, we qualitatively compared this method with our approach in Fig. 5. Despite our efforts, no considerable enhancement in the geometry detail was observed. In fact, the normal integration produced a result characterized by both increased smoothness and noise in comparison to the SDF field. Consequently, we have decided against incorporating normal integration as a means to refine the geometric detail.

D. Depth-guided G-buffer Reprojection

Instead of applying volume rendering to generate normal and albedo maps under novel views, we perform depth-guided G-buffer reprojection to synthesize normal and albedo maps from its nearby PS views. We have shown in paper Fig. 6 and Tbl. 3 that it achieves more photorealistic albedo and accurate normal maps. Below we explain

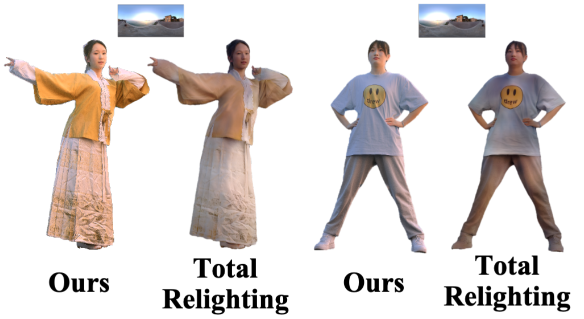


Figure 6. We qualitatively compare our per-scene neural asset relighting approach with Total Relighting. Our method produces more realistic specular effects, as evidenced by the silk dress regions.

how to train the blending networks.

Specifically, we train our blending networks on the Twindom dataset. We pick up 2040 3D models from the Twindom dataset and render 90 training views for each 3D model. Given a random novel view with known camera pose and depth map (In training, we use GT depth map, while in testing, it can be generated by volume rendering, see paper Sec. 4.2), we pick up its 6 closest views, taking them as sources views to synthesize the novel view image.

With the generated depth map in the novel view, for each pixel we compute its world coordinate and re-project it on the 6 selected views and collect 6 RGB colors. While occlusions might happen in reprojection, we further compute the depth difference between the novel view and the reprojected views. We concatenate all the information, including 6 RGB images and 6 depth difference maps, and feed them into a UNET to learn 6 blending weights for each pixel. We then apply weighted average to synthesize the novel view image. While training on a synthetic dataset, we find it generalizes well on the real captures. We refer to [13] for more details.

E. Material Optimization

Following [12], we apply spherical Gaussians (SGs) to approximate the rendering equation (paper Eq. (2)) to accelerate the rendering process. An spherical Gaussian (SG) $\mathcal{G}$ is formulated as:

$$\mathcal{G}(\nu;\xi,\lambda,\mu)=\mu e^{\lambda(\nu\cdot\xi-1)} \quad , \tag{7}$$

where ν is the query direction, ξ is the SG lobe axis, λ is the lobe sharpness and μ is the lobe amplitude. We then represent each term in the rendering equation with SGs. We first approximate the incoming light with 128 SG lobes and the cosine foreshortening term with one SG, similar to [2, 11, 12]. As for the BRDF, we adopt a simplified version of Disney BRDF model [1],

$$f_r(\omega_o,\omega_i;\mathbf{x})=\frac{\mathbf{a}}{\pi}+\frac{D(\omega_h)F(\omega_o,\omega_h)G(\omega_i,\omega_o)}{4|\omega_i\cdot\mathbf{n}||\omega_o\cdot\mathbf{n}|} \quad , \tag{8}$$

where $\mathbf{a}$ is the diffuse albedo, $\omega_h=(\omega_o+\omega_i)/\|\omega_o+\omega_i\|_2$ is the half-vector, $D(\omega_h)$ is the normal distribution function (NDF), $F(\omega_o,\omega_h)$ is the Fresnel term and $G(\omega_i,\omega_h)$ accounts for shadowing effects. Suppose roughness $R\in\mathcal{R}_+$, the NDF $D(\omega_h)$ is defined as:

$$D(\omega_h)=\mathcal{G}(\omega_h;\mathbf{n},\frac{2}{R^4},\frac{1}{\pi R^4}) \quad . \tag{9}$$

The Fresnel term $F(\omega_o,\omega_h)$ is computed as:

$$F(\omega_o,\omega_h)=F_0+(1-F_0)\cdot 2^{(-5.55473\omega_o+6.8316)(\omega_o\cdot\omega_h)} \quad , \tag{10}$$

where F_0 is the specular reflectance. The shadowing terms $G(\omega_i,\omega_o)$ is computed as:

$$G(\omega_i,\omega_o)=\frac{\omega_o\cdot\mathbf{n}}{\omega_o\cdot\mathbf{n}(1-k)+k}\cdot\frac{\omega_i\cdot\mathbf{n}}{\omega_i\cdot\mathbf{n}(1-k)+k} \quad , \tag{11}$$

where $k=\frac{(R+1)^2}{8}$.

Human-centric dataset material analysis. Materials like aluminum and iron will have a larger F_0 while dielectric objects like ceramic will have a smaller specular reflectance. Since UltraStage is a human-centric dataset, the captured images are mainly composed of human skin and daily clothes. The most common materials in our dataset are dielectric, such as cotton shirts, jeans, wool scarves, and human skin. As a result, following [12], we also set F_0 to 0.02, which is suitable for most dielectric materials. We find it works well for our human-centric capture content.

Optimization choice. Zhang *et al.* [12] also models visibility in direct illumination and indirect illumination. Differently, we set the visibility term as 1 and ignore the indirect illuminations. As for the visibility term, our albedo maps already contain occlusion information. We find modeling visibility didn't make too much difference, as shown in Fig. 7. As for the indirect illuminations, the indirect light networks require extra time to train. Due to the time limit, we didn't add this part. We do agree that human skin contains complex indirect illumination effects like subsurface scattering, and applying indirect light networks will potentially improve the performance. Our neural pipeline only serves as a starting point and we believe adding support for advanced rendering effects will be a promising direction.

Per-scene neural asset relighting. We evaluated the effectiveness of UltraStage on two relighting scenarios: per-scene neural asset relighting and learning-based single-image relighting. For the former, we compare our approach with the state-of-the-art method Total Relighting [9]

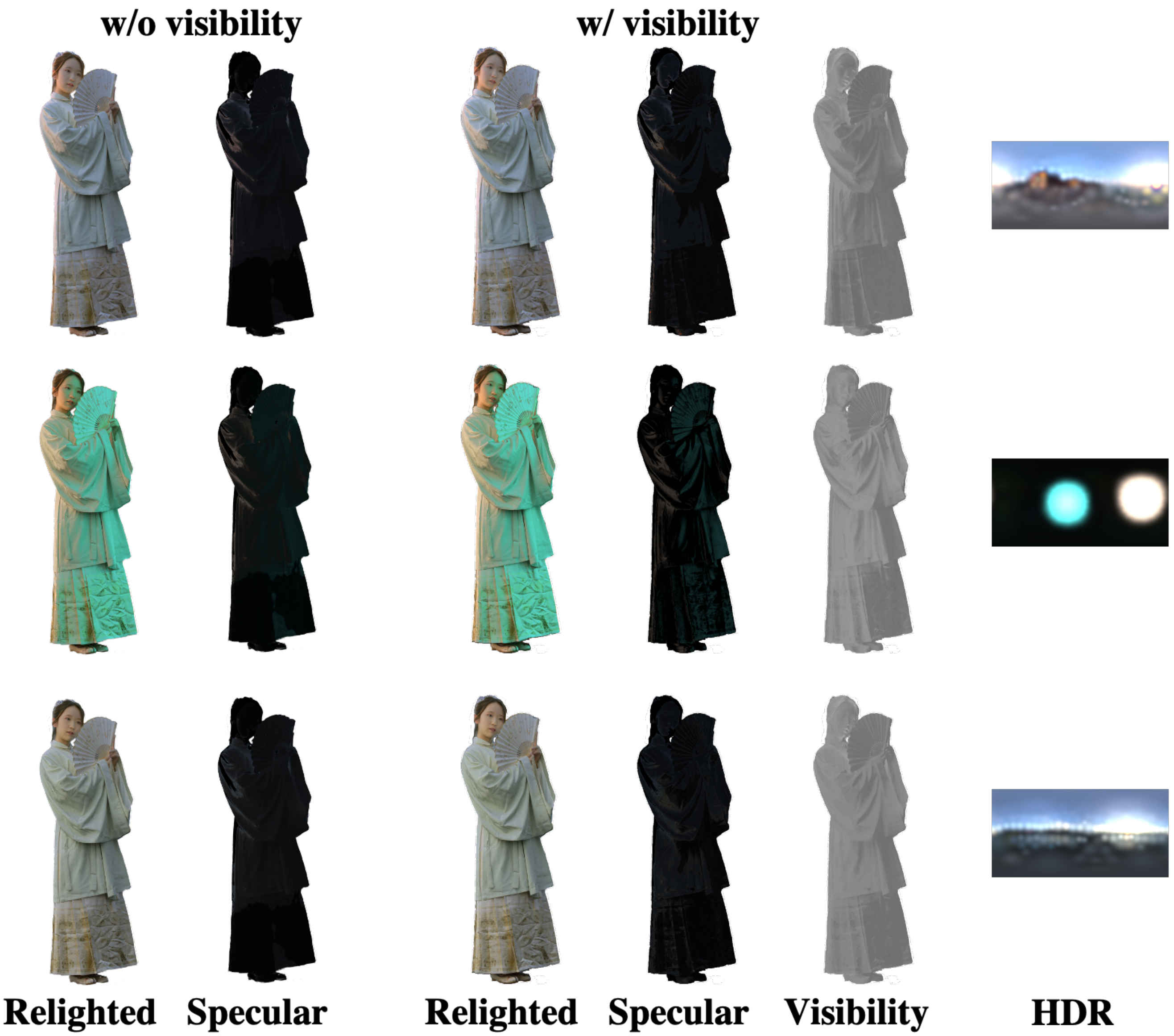


Figure 7. We show whether to model visibility or not doesn't make too much difference. The learned visibility maps are close to a constant.

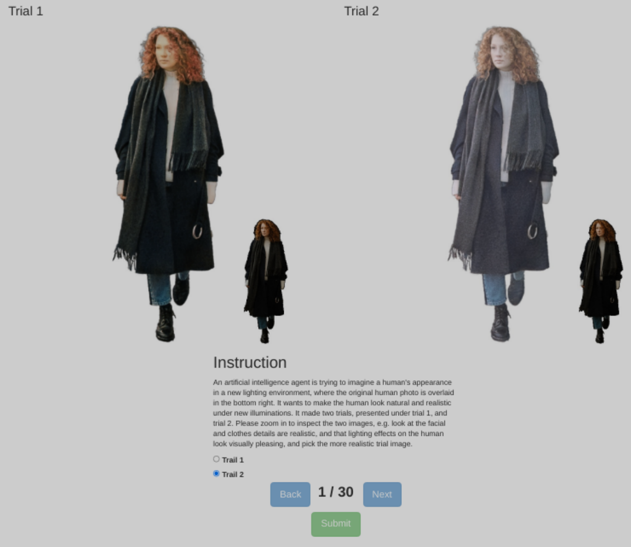
in Fig. 6. Our method produces more realistic specular effects, as evidenced by the silk dress regions. Furthermore, we anticipate that integrating advanced methods like Total Relighting on our dataset will enhance the relighting effects even further, which we leave for future works.

More free-view relighting results. In the main paper Fig. 8 we have compared our novel view synthesis and relighting effects with several baselines and demonstrated significant improvements in rendering quality. The improvements mainly come from the PS priors. Specifically, the two baselines are trained with the MVS images(images captured under white illuminations). In contrast, we train our model with the normal assisted geometry and depth-guided G-buffers, which all rely on the images captured under gradient illuminations. We provide more results in Fig. 15, Fig. 14 and the supplementary video. We believe the rendering and effects can be further improved by applying more powerful designs, which we leave for future works.

F. Single Image Relighting

We show two more testing examples in Fig. 9, where we compare with RH [10] and RW [6] and demonstrate significantly improved relighting effects, thanks to our powerful relighting dataset.

In Fig. 8 we show the user interface of the human study,



Instruction

An artificial intelligence agent is trying to imagine a human's appearance in a new lighting environment, where the original human photo is overlaid in the bottom right. It wants to make the human look natural and realistic under new illuminations. It made two trials, presented under trial 1, and trial 2. Please zoom in to inspect the two images, e.g. look at the facial and clothes details are realistic, and that lighting effects on the human look visually pleasing, and pick the more realistic trial image.

○ Trail 1
● Trail 2

Figure 8. Human Study Interface in AMT. Given the input image (small image in the bottom right), we relight it with our method and a baseline method (either RW or RH) and ask the user to pick up the more natural relit image.

where we invite 9 users to judge each example and use the majority vote to decide the user preference. We compare with RH & RW conduct on 30 examples (6 input images × 5 new lighting conditions), resulting in a total of 540 clicks (2 comparisons × 30 examples × 9 users). We randomize which of the methods is shown on the left vs right to avoid bias in order. The provided instructions are as follows:

An artificial intelligence agent is trying to imagine a human's appearance in a new lighting environment, where the original human photo is overlaid in the bottom right. It wants to make the human look natural and realistic under new illuminations. It made two trials, presented under trial 1, and trial 2. Please zoom in to inspect the two images, e.g. look at the facial and clothes details are realistic, and that lighting effects on the human look visually pleasing, and pick the more realistic trial image.

Among all the examples, more than 80% users choose our method, as shown in the main paper Tbl. 4.

G. Failure Cases

G.1. Artifacts of Depth-guided Reprojection

We use normal maps to train a neural SDF field to represent the geometry. However, in the cases where the SDF field is not accurate, the depth maps integrated from the SDF field will have large errors. As a result, the reprojected texture will have blurry artifacts.

G.2. Disalignments Between Different Illuminations

The formula for normal estimation assumes that images under different illuminations are pixel-aligned. The entire capture process takes around 0.6 second. In some poses, humans cannot keep still during the capture time. Such disalignments between different frames will lead to inaccurate normal maps.

G.3. Normal Estimation Error in Low-reflectance Regions

An inherent limitation of PS methods lies in the accurate estimation of surface normals in areas with dark or low-reflectance regions, such as hair or black clothing, as well as those exhibiting strong texture patterns. These challenging scenarios often yield weaker normal estimation results due to the reduced signal-to-noise ratio and the difficulty of disentangling the complex interactions between surface geometry and reflectance properties. While increasing the number of illumination patterns can potentially improve the estimation accuracy in these regions, it comes at the expense of greater complexity and additional capture efforts, which may not always be feasible in practical applications.

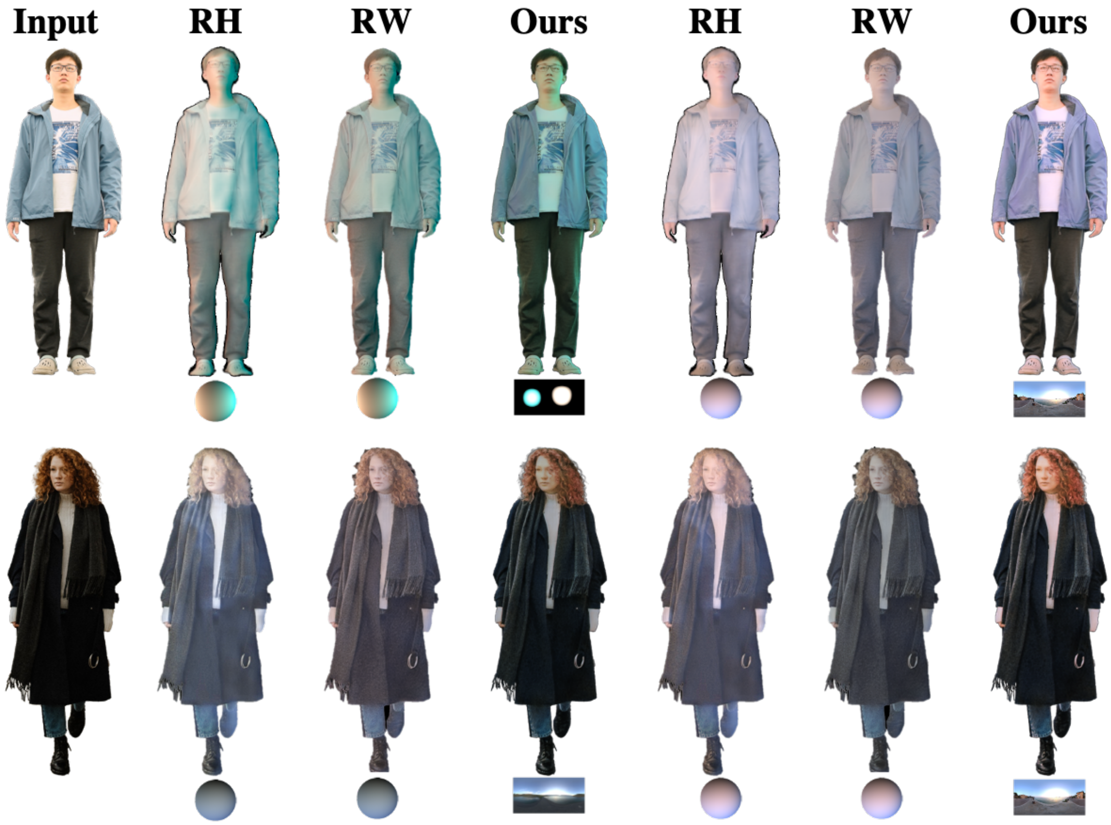


Figure 9. We compare our method with RW [10] and RH [6]. Here we show two examples (two rows). For each input image (the first column), we relight it under two new illuminations (Col.2-4 & Col.5-7). Our method predicts more photorealistic albedo and accurate normal, achieving better relighting effects. Note that RH [10] and RW [6] adopt spherical harmonic lighting (shown with balls) while we use spherical Gaussian environment maps (shown with rectangle images).

References

[1] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *ACM SIGGRAPH*, volume 2012, pages 1–7. vol. 2012, 2012. 4

[2] Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, and Sanja Fidler. Dib-r++: Learning to predict lighting and material with a hybrid differentiable renderer. *Advances in Neural Information Processing Systems*, 34:22834–22848, 2021. 4

[3] Graham Fyffe. Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH'09: Posters*, pages 1–1. 2009. 1

[4] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 1, 2

[5] Christos Kampouris, Stefanos Zafeiriou, and Abhijeet Ghosh. Diffuse-specular separation using binary spherical gradient illumination. In *EGSR (EI&I)*, pages 1–10, 2018. 2

[6] Yoshihiro Kanamori and Yuki Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *arXiv preprint arXiv:1908.02714*, 2019. 5, 7

[7] Chloe LeGendre, Xueming Yu, Dai Liu, Jay Busch, Andrew Jones, Sumanta Pattanaik, and Paul Debevec. Practical multispectral lighting reproduction. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 1

[8] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques*, 2007(9):10, 2007. 1, 2

[9] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 4

[10] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, volume 40, pages 205–216. Wiley Online Library, 2021. 5, 7

[11] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 4

[12] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18643–18652, 2022. 4

[13] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *arXiv preprint arXiv:2209.08468*, 2022. 4

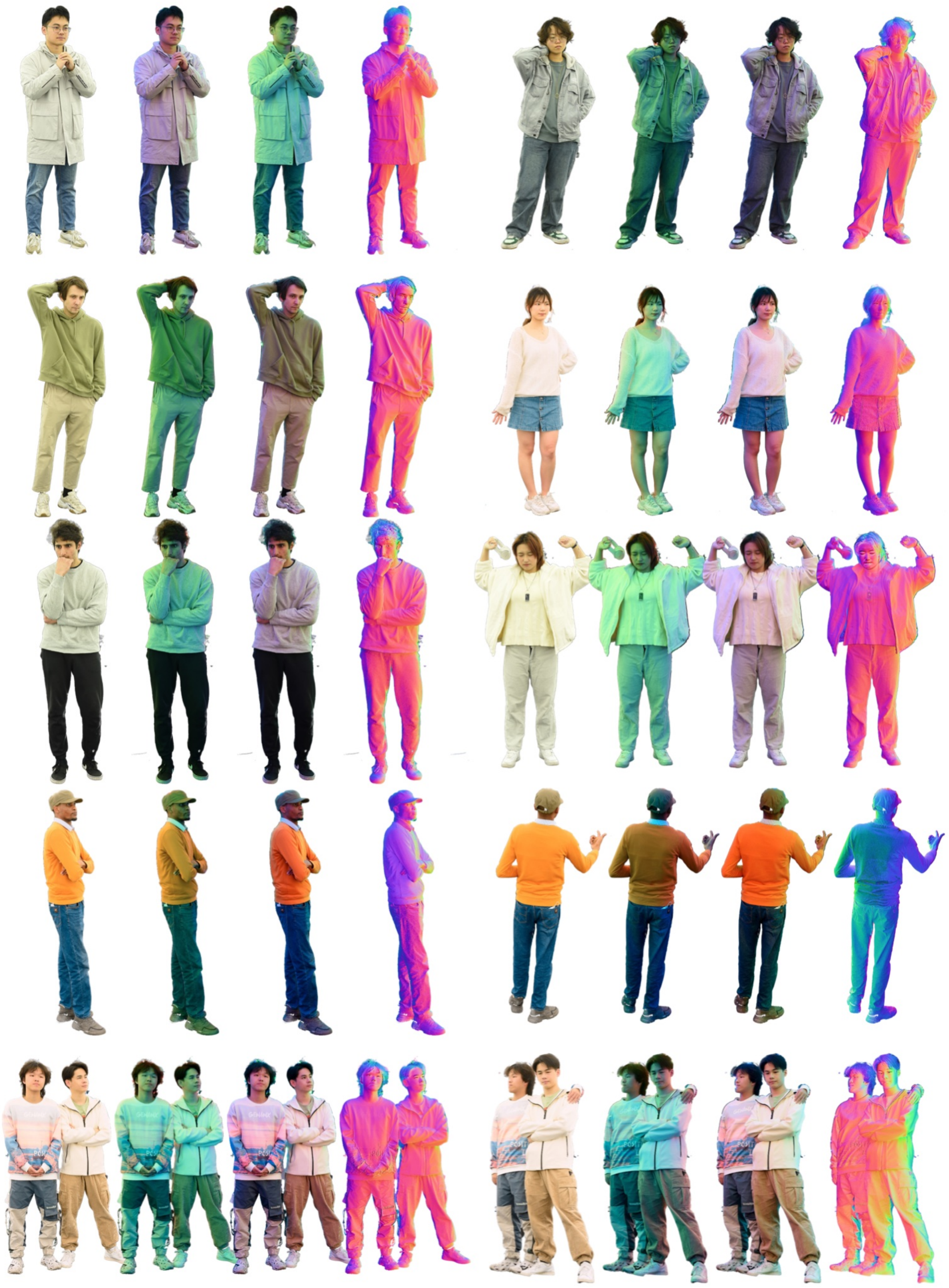


Figure 10. More examples in UltraStage. For each pose, we show their white light image (albedo), color gradient illumination images, and the extracted normal map.

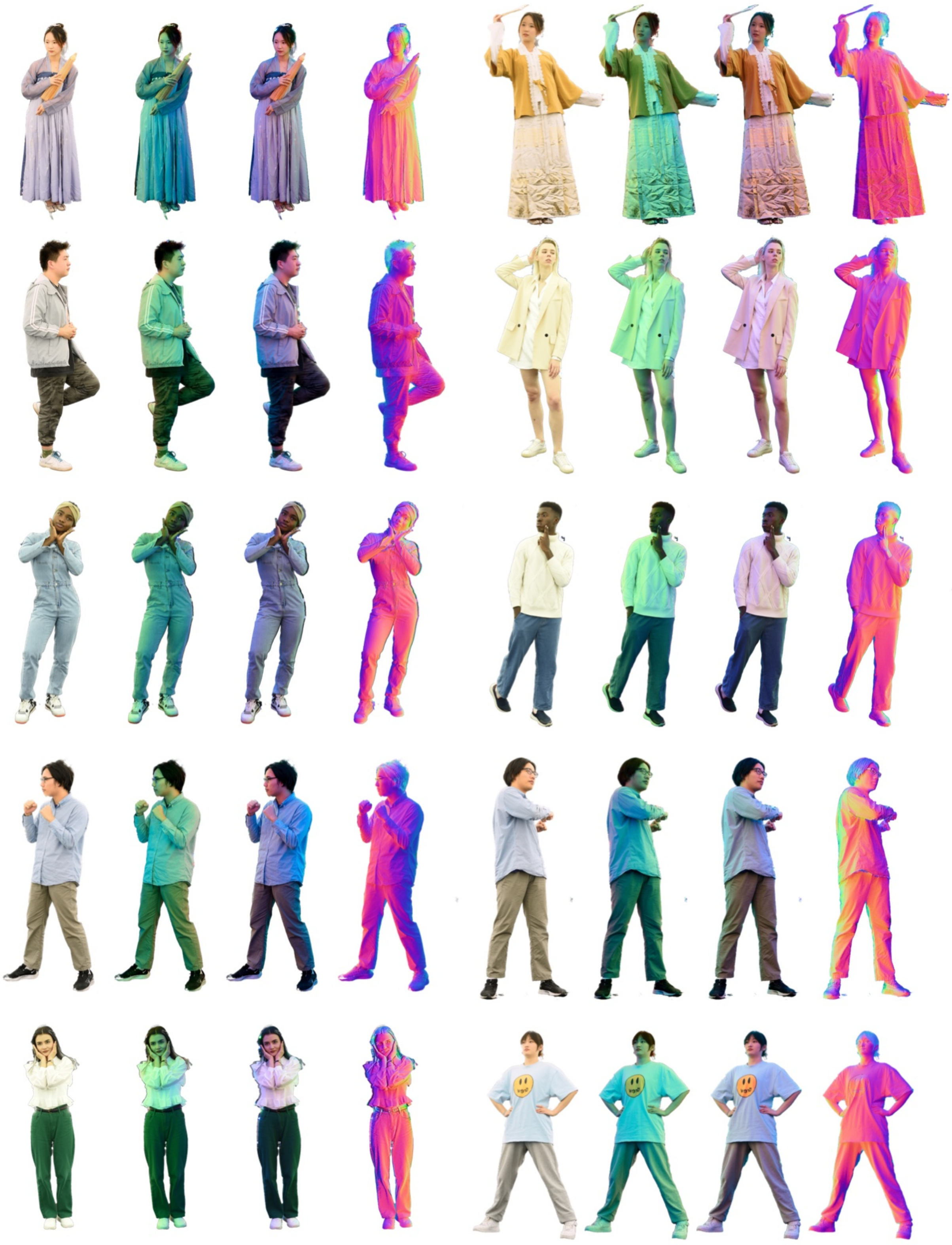


Figure 11. More examples in UltraStage. For each pose, we show their white light image (albedo), color gradient illumination images, and the extracted normal map.

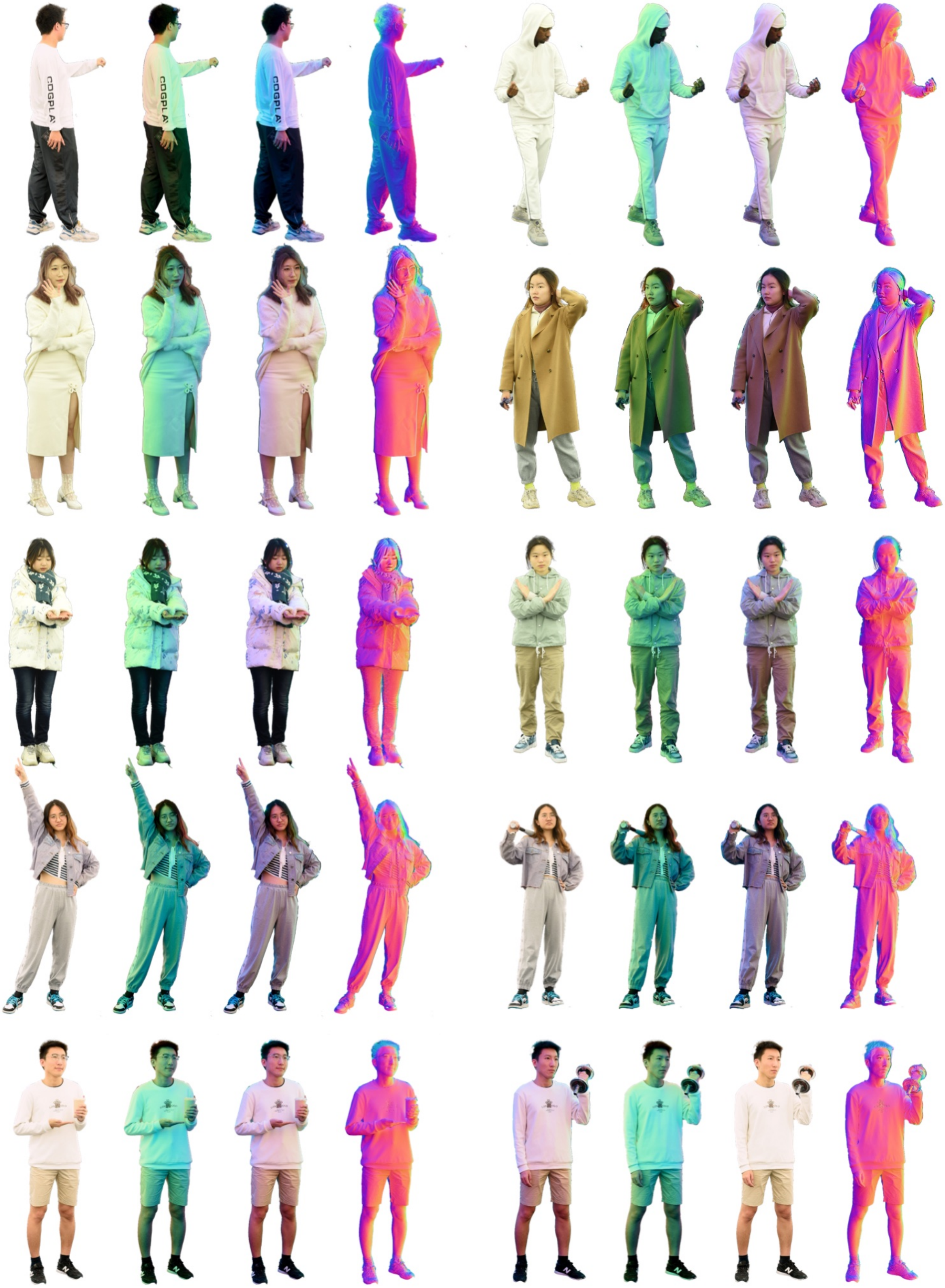


Figure 12. More examples in UltraStage. For each pose, we show their white light image (albedo), color gradient illumination images, and the extracted normal map.

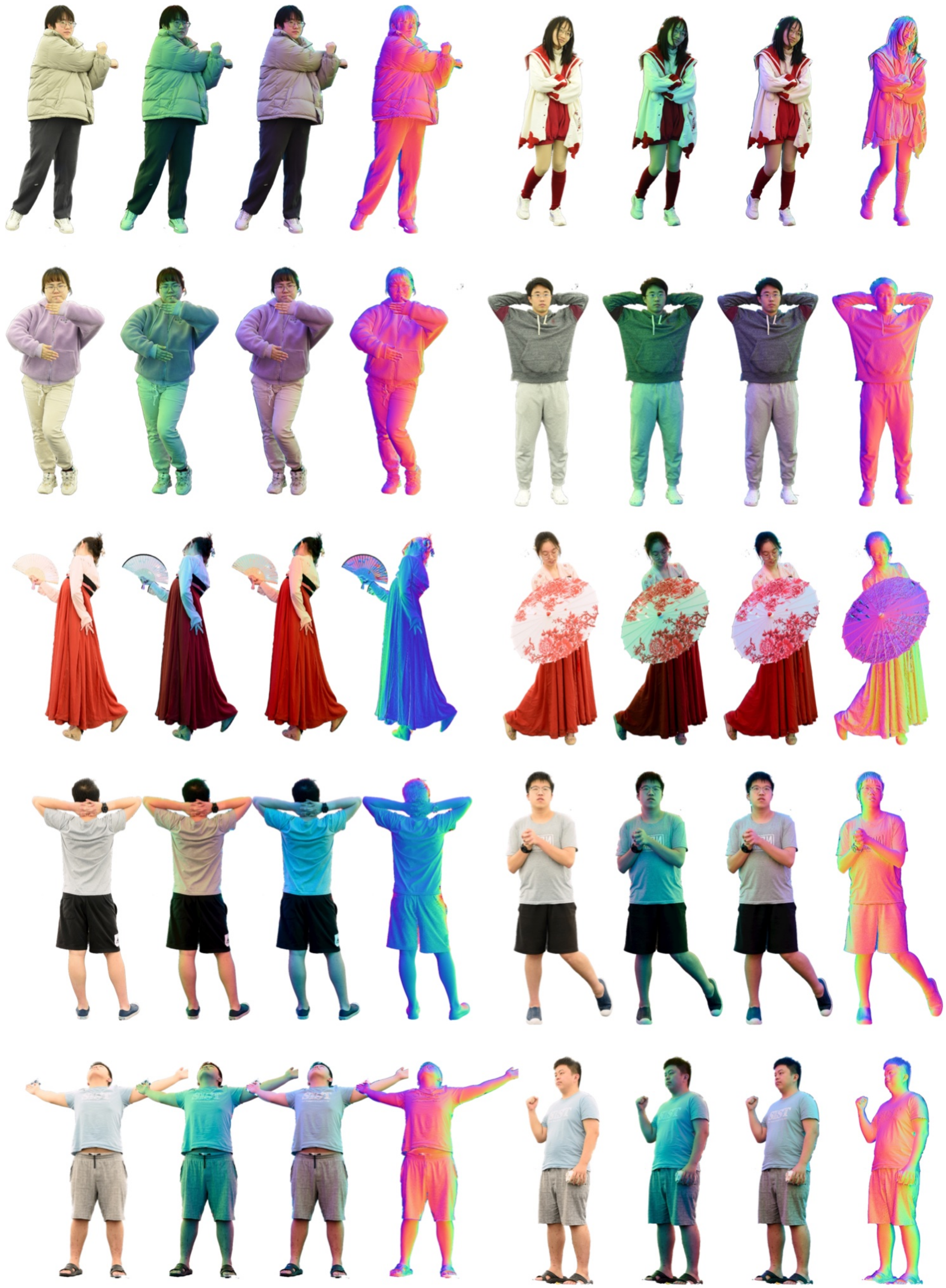


Figure 13. More examples in UltraStage. For each pose, we show their white light image (albedo), color gradient illumination images, and the extracted normal map.

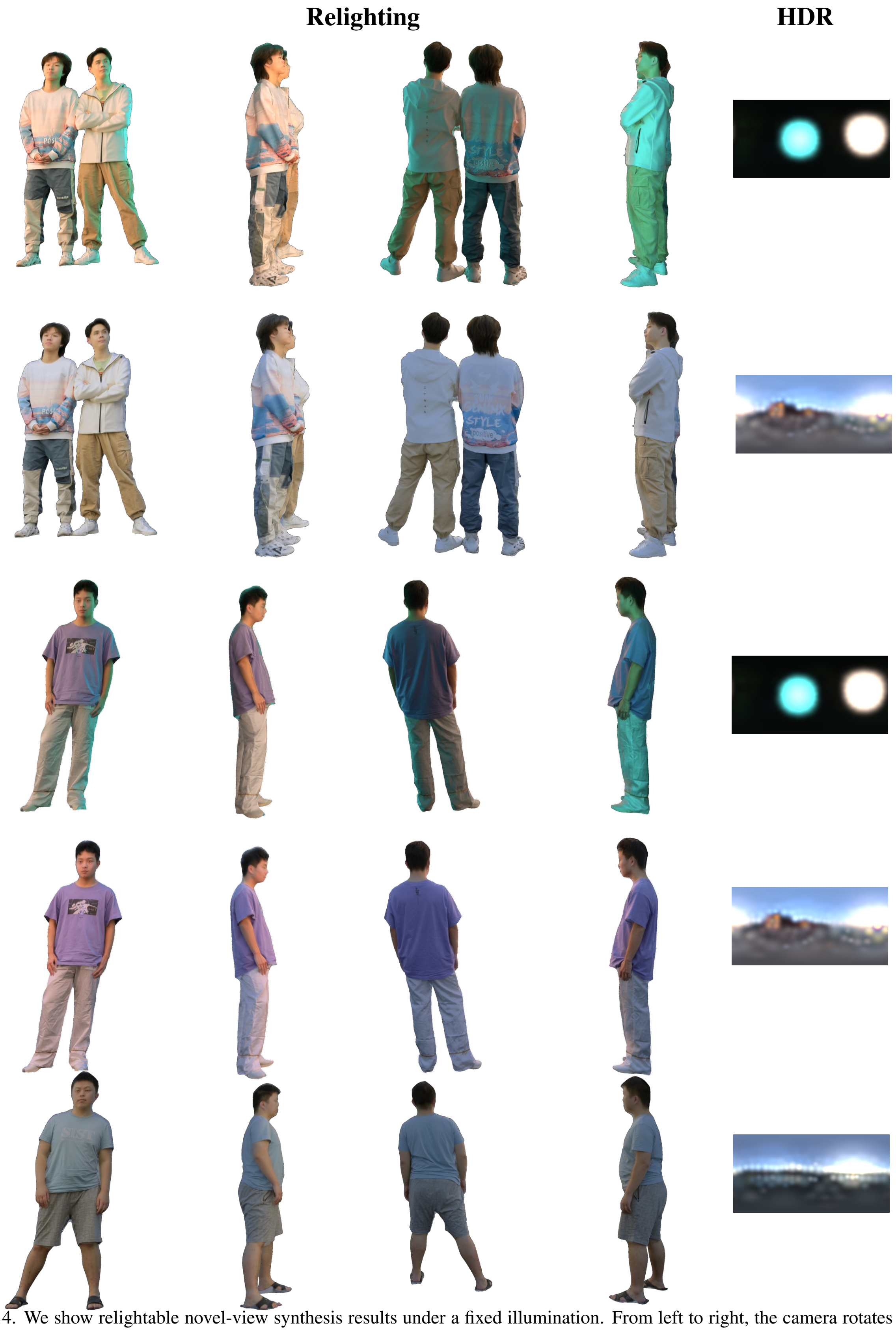


Figure 14. We show relightable novel-view synthesis results under a fixed illumination. From left to right, the camera rotates around the scene. The environment maps are represented by spherical Gaussians.

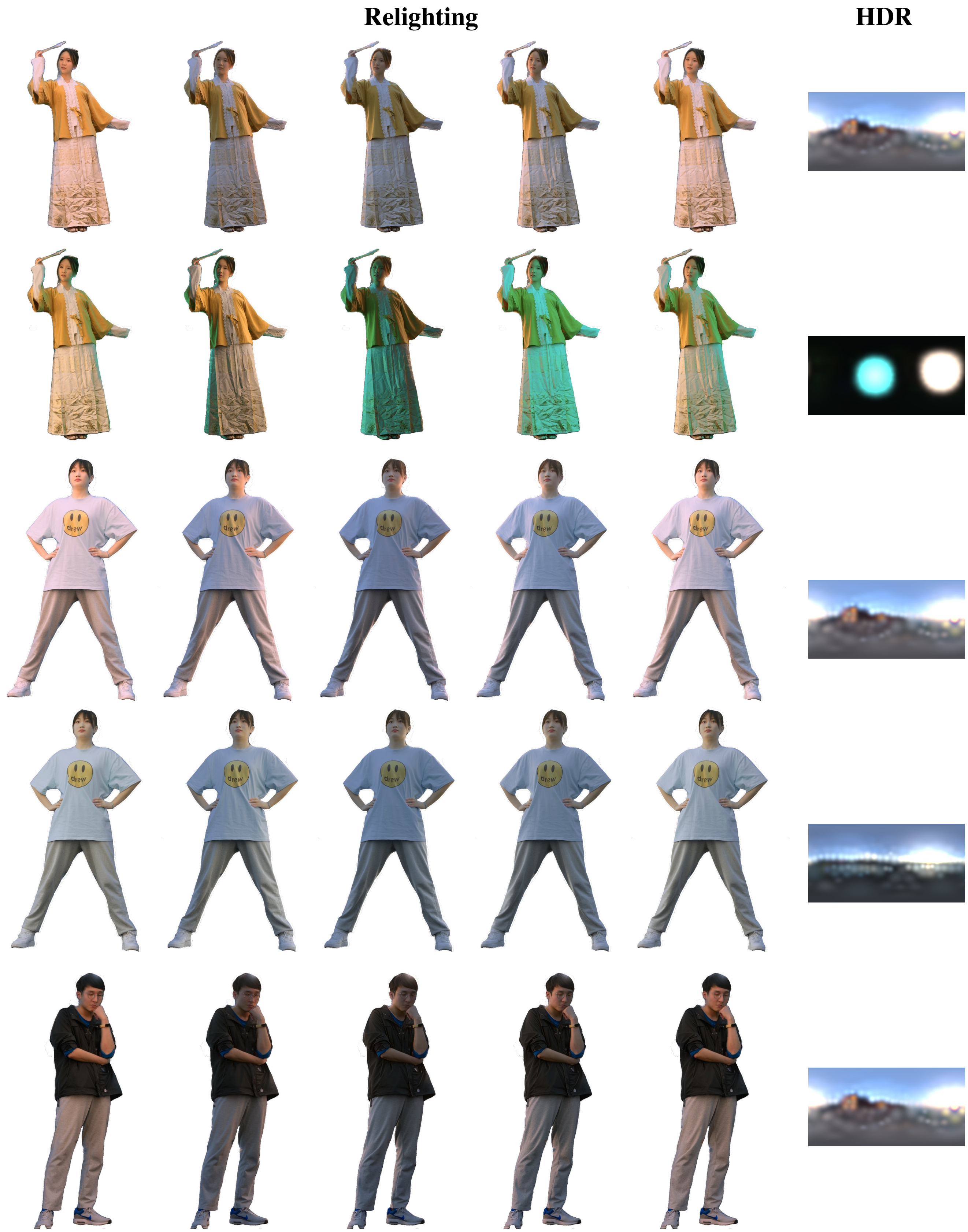


Figure 15. We show the relighting results under a fixed novel viewpoint with dynamic lighting. From left to right, the environment map rotates around.