Comprehensive Relighting: Generalizable and Consistent Monocular Human Relighting and Harmonization Supplementary Material

Junying Wang[1†] Jingyuan Liu[2] Xin Sun[2] Krishna Kumar Singh[2] Zhixin Shu[2]
He Zhang[2] Jimei Yang[3] Nanxuan Zhao[2] Tuanfeng Y. Wang[2] Simon S. Chen[2]
Ulrich Neumann[1] Jae Shin Yoon[2]

[1]University of Southern California [2]Adobe Research [3]Runway

In this document, we provide more details for the method, experiments, dataset, and more qualitative results, as an extension of Sec. 3 and Sec. 4 in the main paper. Please also refer to the video demo for dynamic relighting results, comparison, ablation study, and more results.

A. Method and Experiment Details

We demonstrate that during training, instead of directly using albedo and shading maps, we train with relit images using different lighting augmentations. By leveraging a conditional diffusion model, our approach can implicitly disentangle lighting and appearance from the input image, learning to generate relit images and bypassing the need for a preprocessed de-lighting process.

A.1. Relighting and Harmonization Diffusion Network (Sec. 3.2)

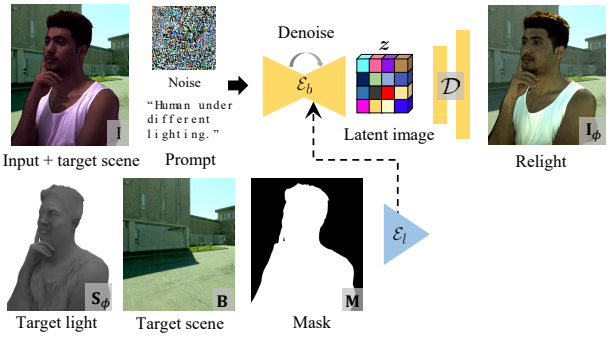

Figure 1. Relighting and Harmonization diffusion model training and denoising pipeline.

As shown in Fig. 1, which includes the diffusion model training process and denoising (sampling) process for our fine-grained relighting. During the training process, we follow the same Stable Diffusion architecture as [1], and both Lighting ControlNet and Motion ControlNet architecture are followed by [19]. Stable Diffusion model adopts a U-Net [12] architecture comprising an encoder, a middle block, and a skip-connected decoder. Each of the encoder and decoder consists of 12 blocks, totaling 25 blocks within the complete model, and each primary block integrates 4 ResNet layers and 2 Vision Transformers (ViTs) with cross-attention and self-attention mechanisms. The ControlNet architecture is applied at each encoder level of the U-Net, featuring a trainable copy of 12 encoding blocks and 1 middle block from the Stable Diffusion model. These 12 encoding blocks includes: 64×64, 32×32, 16×16, 8×8, with each resolution replicated 3 times. The resulting outputs are merged with the 12 skip connections and the single middle block within the U-Net structure. We fine-tune both ControlNet and Stable diffusion module on our relighting dataset.

A.2. Training Dataset (Sec. 4)

In Fig. 4, we visualize the samples of our training dataset. We use two kinds of dataset. One is from the data captured from LightStage where the background images are rendered from a HDR environment map. The ground truth shading, albedo, relighted image, and background captured from a small number of viewpoints (*e.g.*, 6 views) are available. The other one is from the data rendered from a synthetic human model. We render the image of many 3D human models from many views (e.g., 16 views) under different lighting conditions defined by an environment map. We obtain the approximated spherical harmonics coefficients from the environment maps as ground-truth lighting parameters. The ground truths for the mask, albedo, background, and relit images also exist. We kindly note that our training data is relatively smaller compared to other image-based relighting methods as summarized in Fig. 2. For instance, Total Re-

[1†]This work is partially done during an internship at Adobe Research.

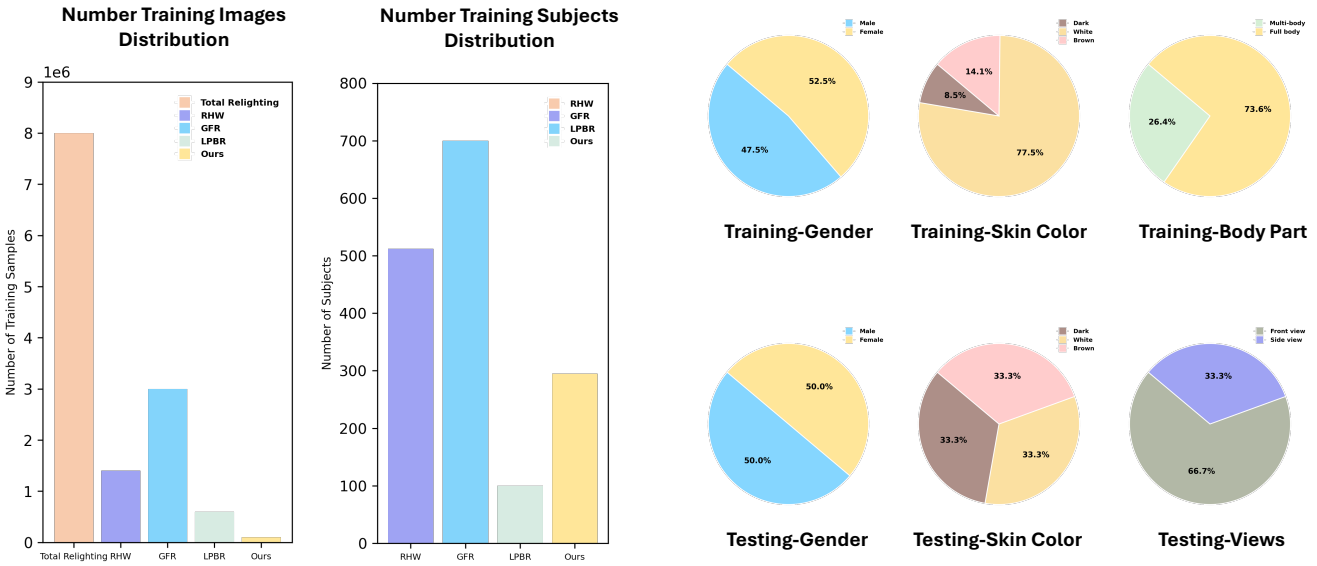

Figure 2. Left side: Training data scale comparisons; Right side: Breakdown of our training and evaluation dataset information.

lighting [7] captures data from 70 diverse subjects. Through extensive lighting augmentation, the dataset expands to include approximately 8 million OLAT training examples; GFR [4] needs 700 subjects and 4,600 HDR maps for training; and LPBR [11] is trained on 100 subjects with OLAT and 2,908 HDR maps, resulting in 600K training samples. Our training data is composed of 100K samples where the detailed data analysis can be found in Fig. 2. We categorize our training data based on gender, skin tone, and body coverage (half-body and full-body). Each subject is captured from 32 viewpoints under varying lighting conditions.

A.3. Add-on Temporal Motion Module Network (Sec. 3.3)

Algorithm 1 Unsupervised Cycle-Training Motion Modeling for Temporal Consistency

Require: Video frames $\mathbf{I}$; decoder $\mathcal{D}_*$
Require: Relit frames $\mathbf{I}_\phi \leftarrow (\mathcal{D}_* \circ \mathcal{E}_b)$
Initialize: Motion encoder $\mathcal{E}_\mathrm{m}$; train step function $\mathbf{T}$
Converged $\leftarrow$ **False**
While not Converged **do**
  $\mathbf{I}^t_\phi \leftarrow \mathcal{D}^*(\mathcal{E}^*_\mathrm{b}(\mathbf{I}^t, \mathcal{E}^*_\mathrm{l}(\{\mathbf{S}^t_\phi, \mathbf{B}^t\}, \mathbf{M}^t)))$
  $\tilde{\mathbf{I}}^t_{t-1} \leftarrow \mathcal{D}^*(\mathcal{E}^*_\mathrm{b}(\mathbf{I}^t_\phi, \mathcal{E}_\mathrm{m}(\mathbf{I}^{t-1}, \mathbf{M}^{t-1})))$
  Converged $\leftarrow \mathbf{T}(\tilde{\mathbf{I}}^t_{t-1}, \mathbf{I}^t)$
end while

We present the cycle-training algorithm for our temporal lighting module in Alg.1, which serves as an additional explanation for Sec. 3.3. Based on the hypothesis: original video sequence inherently contains temporal lighting properties, which can be modeled by a temporal module, conditioned on the relit version. We train an add-on temporal module in an unsupervised way. Before the training process, we require relit video frames, $\mathbf{I}^t \rightarrow \mathbf{I}^t_\phi$. To generate the relit frame we process forward image relighting: $\mathbf{I}^t_\phi \leftarrow \mathcal{D}^*(\mathcal{E}^*_\mathrm{b}(\mathbf{I}^t; \mathcal{E}^*_\mathrm{l}(\{\mathbf{S}^t_\phi, \mathbf{B}^t\}; \mathbf{I}^t, \mathbf{M}^t)))$. During each training iteration, as indicated in: $\tilde{\mathbf{I}}^t_{t-1} \leftarrow \mathcal{D}^*(\mathcal{E}^*_\mathrm{b}(\mathbf{I}^t_\phi; \mathcal{E}_\mathrm{m}(\mathbf{I}^{t-1}, \mathbf{M}^{t-1})))$, we condition on the current relit frame and revert the lighting of the previous frame in the original video back to match that of the original frame.

Implementation details. We train our model on 8 A100 GPUs with a total batch size of 32 (4 batches per GPU) and a learning rate of 2×10^{-6}. In the training phase for Lighting ControlNet, we initialize the Stable diffusion base model using the pre-trained weights from Instruct-Pix2Pix [1], and copy the encoder block weights to serve as the initial weights for the Lighting ControlNet part. Subsequently, we fine-tune both ControlNet and Stable Diffusion module on our relighting dataset

The training of our Motion ControlNet module occurs subsequent to the lighting control training process. During the training phase for motion control, we freeze the weights of the Stable Diffusion base model. Then, we initialize the weights of the Motion ControlNet by copying the encoder block weights from the previously trained lighting Stable Diffusion. Subsequently, we exclusively fine-tune the Motion ControlNet.

During the inference process, we adopt random noise with a resolution of $4 \times 96 \times 96$ as the initial input to generate the final relit image with a resolution of 768×768, and for video testing, we apply the same noise across frame. We apply DDIM [13] sampler with a timestep of 50 to gener-

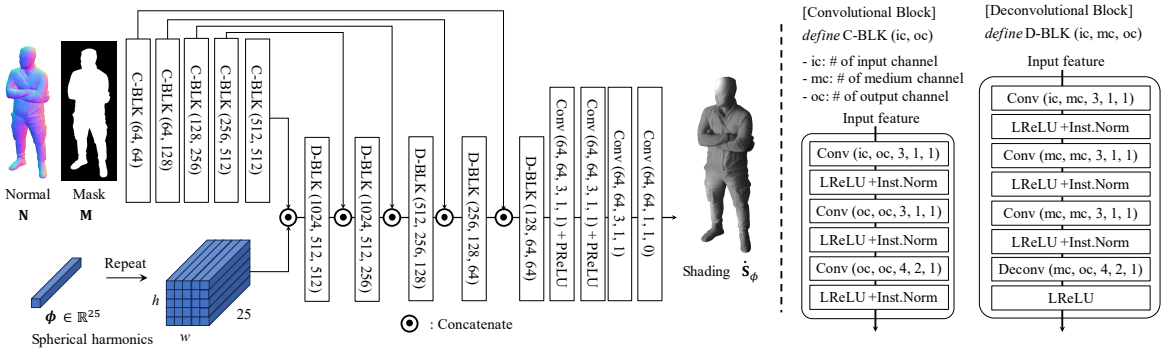

Figure 3. Left: Our shading estimation network, Right: Convolutional and deconvolutional blocks.

ate the final relit image. To utilize frame-by-frame inference with recurrent blending, we extract control features from the 12 encoding blocks of the ControlNet at corresponding resolutions. Subsequently, we perform weighted blending between control feature of previous and current frames.

A.4. Pixel-Aligned Neural Shading (Sec. 3.2)

While coarse shading $\mathbf{S}_\phi$ can be directly computed from Spherical harmonics (SH) lighting parameters, we experimentally found that using $\mathbf{S}_\phi$ obtained from a neural network can improve human relighting and harmonization. Specifically, low-order SH models tend to smooth out fine details, resulting in overly diffuse shading. In contrast, a neural network can recover high-frequency shading variations, enhancing realism by capturing subtle lighting effects. Moreover, the learned shading function improves robustness to normal map inaccuracies, reducing artifacts and better preserving surface details. In this section, we introduce an alternative way of having a coarse shading using a neural network. To this end, we introduce a pixel-aligned lighting estimation function f in Eq. 2 using a conditional Unet framework.

It takes as inputs surface normal map $\mathbf{N}$ and target lighting parameters $\boldsymbol{\phi}$ as conditions, and estimates the shading $\mathbf{S}_\phi$ at each pixel lit by the target lighting. $\mathbf{N}$ is detected from the input image $\mathbf{I}$ using the internal normal detector which is composed of Unet architecture with pyramid vision transformer [15]. It learns many mixtures of ground-truth data similar to [10], and thus, applicable to general scenes and objects. Note that, since f does not take any visual data as inputs, it does not introduce visual domain gaps. We train the $f(\cdot)$ by comparing the input image and its reconstruction from the estimated shading:

$$\mathcal{L}_{\text{recon}} = \sum_i \|\mathbf{I}_{\text{recon}} - \mathbf{I}\|_2^2 = \sum_i \|\mathbf{S}_\phi \odot \mathbf{A}_{\text{GT}} - \mathbf{I}\|_2^2$$

where $\mathbf{I}_{\text{recon}}$ is the reconstructed image based on the multiplication of $\dot{\mathbf{S}}_\phi$ with the ground-truth albedo $\mathbf{A}_{\text{GT}} \in \mathbb{R}^{w \times h \times 3}$. Since we supervise the shading estimation network in the image space, we can utilize other advanced image-based supervision signals that can capture the physical plausibility of the local and global shading as follows:

$$L_{\text{shade}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{v}} \mathcal{L}_{\text{vgg}} + \lambda_{\text{c}} \mathcal{L}_{\text{cGAN}}, \tag{1}$$

where L_{shade} is the entire objective, and λ controls the weight of each loss function. $\mathcal{L}_{\text{vgg}}$ is designed to penalize the difference between the reconstructed image $\mathbf{I}_{\text{recon}}$ and the input $\mathbf{I}$ in the deep feature space [5]. $\mathcal{L}_{\text{cGAN}}$ is the conditional adversarial loss [3] to evaluate the plausibility of the reconstructed shading with respect to the geometric structure where we use $\{\mathbf{N},\ \mathbf{I}\}$ as real and $\{\mathbf{N},\ \mathbf{I}_{\text{recon}}\}$ as fake conditions to the patch discriminator [3].

Coarse Shading Estimation Network. In Fig. 6, we show the general training pipeline for coarse lighting estimation network. Fig. 3 describes the structure of our coarse shading estimation network. It takes as inputs the surface normal, foreground mask, and lighting parameters (*i.e.*, Spherical harmonics); and generates the shading map. An encoder regresses the surface normal and mask to the latent space. In this latent space, the lighting parameters are conditioned where the vector parameters are copied along the spatial direction to fit the same latent space as the one from the encoder. A decoder decodes them to generate a shading map.

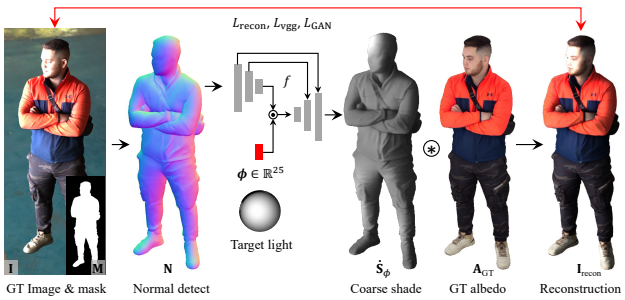

Figure 6. Training pipeline for coarse lighting estimation network.

B. Qualitative Results

B.1. Comparison with other baselines (Sec. 4)

We present the qualitative results of static image testing on our synthetic dataset, compared with other baseline methods: DPR [21], GFR [4] and RHW [14] in Fig. 7. In our evaluation, we perform full-body and multi-person tests on our synthetic testing dataset, integrating background images alongside Spherical harmonics for lighting control. We calculate the average error on the entire testing dataset for a comprehensive and generalizable relighting evaluation. From visual quantitative results, our model shows more realistic relighting results compared to other human relighting models. This demonstrates our model's robust performance across diverse body part tests, indicating higher generalizability.

For evaluation, we validate our model along with other baselines based on the divided categories: gender, and skin color. We present the numerical evaluation in Tab. 1 and 2. From the qualitative results, our method consistently outperforms in all categories.

We further highlight that while all those methods are limited to working on a specific body part (e.g., face or portrait), our method works on general cases including the scene with face, portrait, full body, and multi-person.

We present real data comparison results on the Light-Stage dataset in Fig. 9 and comparisons on in-the-wild images in Fig. 8. Since current state-of-the-art (SOTA) baselines are not designed for comprehensive relighting, their performance varies across different scenarios. In Fig. 8, while DPR performs well for face relighting, its quality significantly deteriorates in half-body scenarios, exhibiting strong artifacts due to domain gaps. Notably, our framework is the first to achieve comprehensive relighting, effectively handling arbitrary body parts, including portraits, half-body, full-body, and multi-body scenarios.

In Fig. 12 and Fig. 13, we present static real image relighting and harmonization comparison results. For harmonization, we use the most recent work, LPBR [11], as one of the baselines: (1) DPR and RHW are only applicable to image relighting with Spherical harmonics for lighting control. For a fair comparison, we tested image relighting with DPR, RHW, and GFR in Fig. 12, using a black background and target lighting parameters. We applied different lighting conditions to various identities, including half-body and full-body images. Although these methods can achieve human relighting, their limited generalizability results in less fidelity during comprehensive testing. (2) Both LPBR and GFR can perform harmonization. We retrained the GFR model with our settings, enabling it to achieve both harmonization and relighting, as shown in Fig. 13. The higher generative prior of LPBR, which also uses a diffusion model, results in noticeable distortions on the human face. Although GFR can achieve both harmonization and relighting, it exhibits obvious color noise.

Method	SH	Bg	Male	Female
RHW	✓	✗	28.89 / 0.950	26.58 / 0.939
DPR	✓	✗	27.63 / 0.972	27.62 / 0.944
GFR	✓	✓	29.32 / 0.926	29.71 / 0.973
Ours	✓	✓	**31.12 / 0.970**	**30.50 / 0.964**

Table 1. Comparison of baseline methods on our full-body synthetic static data, categorized by gender: (PSNR↑ / SSIM↑).

Method	White	Brown	Dark
RHW	28.15 / 0.946	27.37 / 0.944	27.68 / 0.943
DPR	27.44 / 0.956	27.70 / 0.962	27.73 / 0.956
GFR	29.94 / 0.936	29.41 / 0.934	29.10 / **0.978**
Ours	**31.53 / 0.985**	**31.77 / 0.976**	**29.13** / 0.940

Table 2. Comparison of baseline methods on our full-body synthetic static data, categorized by skin color: (PSNR↑ / SSIM↑).

In Fig. 5, we present a new comparison with IC-Light [20], which is the current state-of-the-art for light-aware background harmonization. Both IC-Light and our model are stable diffusion relighting models. IC-Light can generate relit images with text prompts or background harmonization. In the visual results, our harmonization seamlessly blends with the target background while preserving the original identity. While IC-Light also achieves high-quality background harmonization, however, it exhibits greater identity distortion at the same image resolution, particularly in full-body and multi-person scenarios. In Fig. 14, third graph, we show the user preference comparison among our method, LPBR, and IC-Light. Most users selected our method as the best result for all questions.

For video relighting comparison, we present qualitative results in Fig. 11, in the main paper. We show frames relit by our model tested on the synthetic video testing data. The first row shows the composite input (albedo foreground and background). In the second row, we show the ground truth shading, and the third row displays the ground truth relit image. The following rows show our relit frames, followed by those from GFR, RHW, LPBR, and DPR. For real video comparison, please refer to the supplementary demo video.

B.2. More qualitative results

We present additional qualitative results on the DeepFashion dataset [6], as shown in Fig. 15. Given an input image (left side) and target lighting parameters, our model achieves the relighting results (second column). By changing the background image, our model can achieve both background harmonization and relighting, as demonstrated in columns 3 through 7.

Our model can achieve realistic relighting effects given a target lighting, as well as background harmonization and a combination of both. It effectively handles diverse subjects

with varying identities and poses, including both half-body and full-body representations, demonstrating higher generalizability.

B.3. Performance and rendering time

For the generation of the 768x768 pixel resolution image with stable quality, 50 diffusion timesteps are required, leading to around 10 seconds. For video sequences with relighting using a motion module, each frame takes approximately 25 seconds on an A100 GPU. In theory, there is no limit in the number of frames that our model can handle, the video rendering time is highly proportional to the number of frames, requiring around 2 hours for a video clip with 300 frames (768x768).

B.4. User study

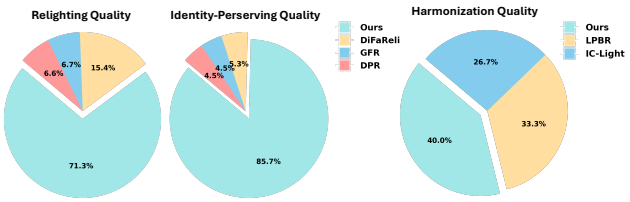

Figure 14. User study results: Preferences between our model and other relighting and harmonization models, including our general object testing.

We performed a user study as shown in Fig. 14. For the relighting model, we used three state-of-the-art methods: DiFaReli [9], GFR [4], and DPR [22]. For the harmonization model, we chose LPBR [11]. Users participated in answering three questions:

- **Q1:** Which result most effectively achieves the relighting?
- **Q2:** Which result most effectively preserves the person's identity (e.g., details and skin)?
- **Q3:** Which result best harmonizes with background scenes?

We summarized the percentage of user preferences and plotted the pie graph as shown in Fig. 14. Overall, users selected our method as the best result for all questions, implying that our method is perceptually effective in achieving reasonable relighting quality, preserving identity, and harmonizing with the background.

C. Limitation and future work

In Fig. 10, we demonstrate some relighting results of the person under shadow and highlights. While our method can suppress shadows from self-occlusion during relighting, we acknowledge that our model shows some weaknesses with strong shadows, especially on human clothes (failure cases in Fig. 10, right side). In fact, these strong shadows can be further suppressed by existing shadow removal models such as [2, 16, 18]. Additionally, incorporating various training data augmentations for hard shadows can be explored as future work to further enhance relighting quality. Our relighting diffusion model requires significant computational time. Recent advancements in diffusion models, such as the One-Step Diffusion Model [17], may further enhance inference efficiency. Significant noise on the detection (*e.g.*, mask and surface normal) affects the temporal coherence, and we admit that our results still have residual flickering. Nevertheless, our approach surpasses other relighting methods in video quality across diverse domains. We believe that advancing video prior models and expanding video datasets will further enhance temporal coherence, which we plan to explore in future work. Our task primarily focuses on human relighting, which limits the model's ability to accurately handle materials associated with general objects such as cars, glass, and metallic surfaces. We acknowledge this limitation and plan to explore this aspect in future work.

D. Broader Impact

As a positive impact, this work can be a useful tool for enhancing the lighting condition of the picture with humans, which can be useful for contents creation in social media. As a negative impact, similar to image synthesis, this work can synthesize human appearance under different lighting that may be used to fabricate fake videos and news.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 2

[2] David Futschik, Kelvin Ritland, James Vecore, Sean Fanello, Sergio Orts-Escolano, Brian Curless, Daniel Sỳkora, and Rohit Pandey. Controllable light diffusion for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8412–8421, 2023. 5

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[4] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *European Conference on Computer Vision*, pages 388–405. Springer, 2022. 2, 4, 5, 13

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[6] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 1096–1104, 2016. 4, 12

[7] 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. 2

[8] Pexels: https://www.pexels.com. 9, 10

[9] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. *arXiv preprint arXiv:2304.09479*, 2023. 5

[10] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

[11] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. *CVPR*, 2024. 2, 4, 5, 13

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[14] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, pages 205–216. Wiley Online Library, 2021. 4

[15] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 3

[16] Joshua Weir, Junhong Zhao, Andrew Chalmers, and Taehyun Rhee. Deep portrait delighting. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 5

[17] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 5

[18] Jae Shin Yoon, Zhixin Shu, Mengwei Ren, Cecilia Zhang, Yannick Hold-Geoffroy, Krishna Kumar Singh, and He Zhang. Generative portrait shadow removal. *ACM Transactions on Graphics (TOG)*, 43(6):1–13, 2024. 5

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[21] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single portrait image relighting. In *International Conference on Computer Vision (ICCV)*, 2019. 4

[22] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 5

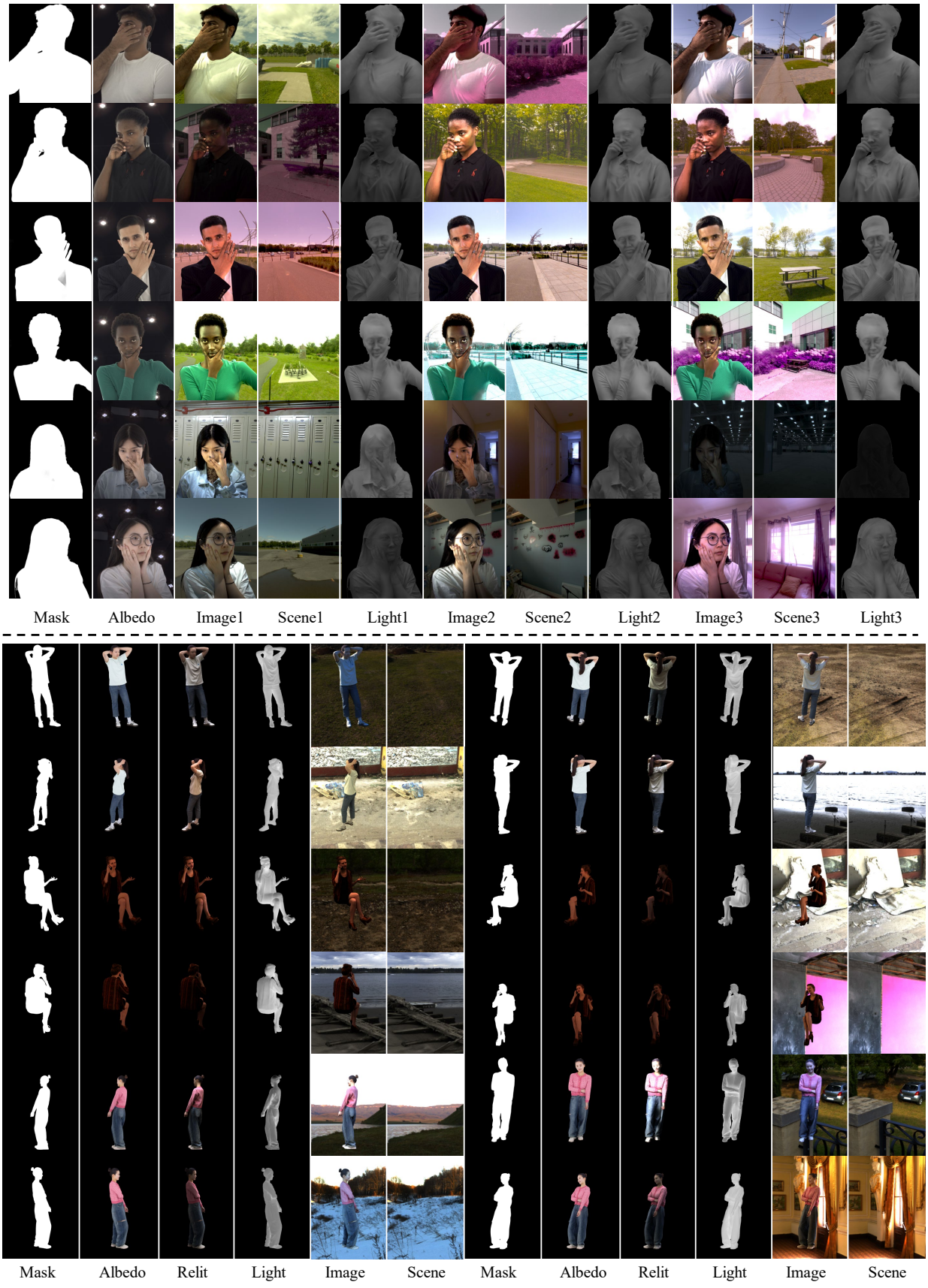

Figure 4. Training samples of the relighting data with half-body portraits (up) and simulation data with full-body images (bottom) .

Figure 5. Comparison with harmonization methods (IC-Light). Left side is multi-person testing, right side is zoom in result.

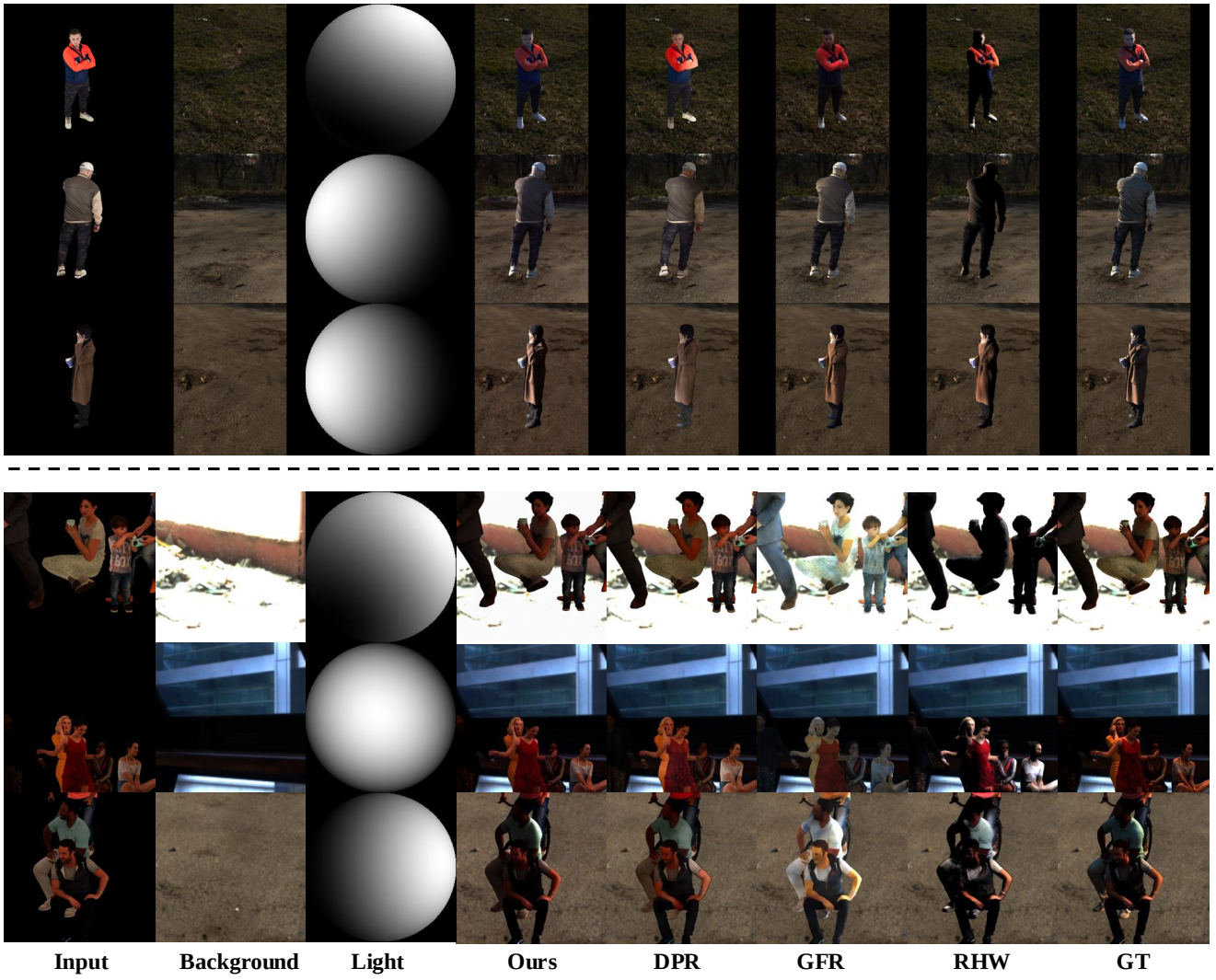

Figure 7. Qualitative comparisons conducted on synthetic data. From top to bottom: full-body testing, multi-person testing. The ground truth data is displayed in the last column.

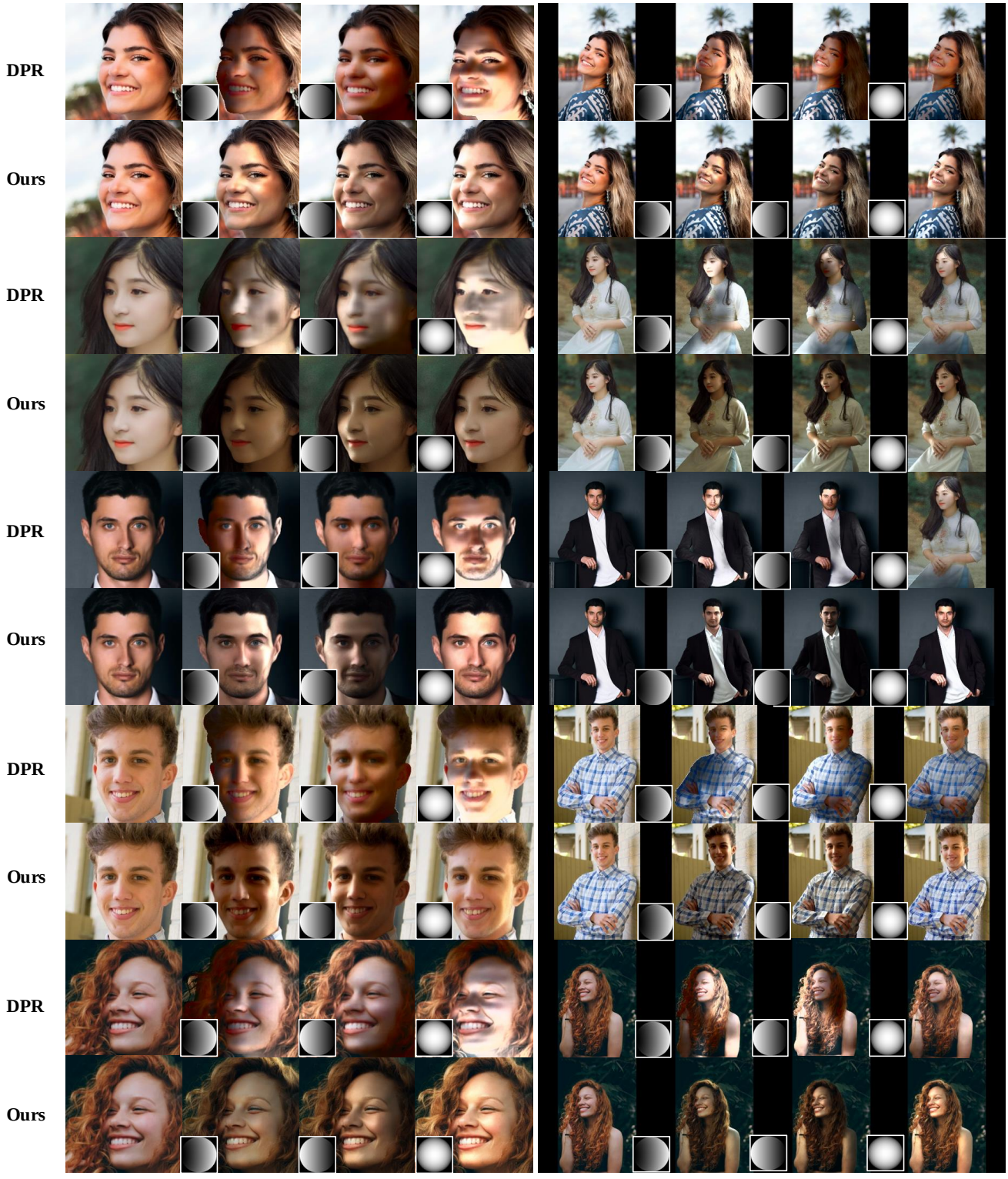

Figure 8. Comparison with DPR on face and half-body relighting on Pexels [8] real images.

Figure 9. Our LigtStage data testing (Left) and comparison with other relighting baselines (Right).

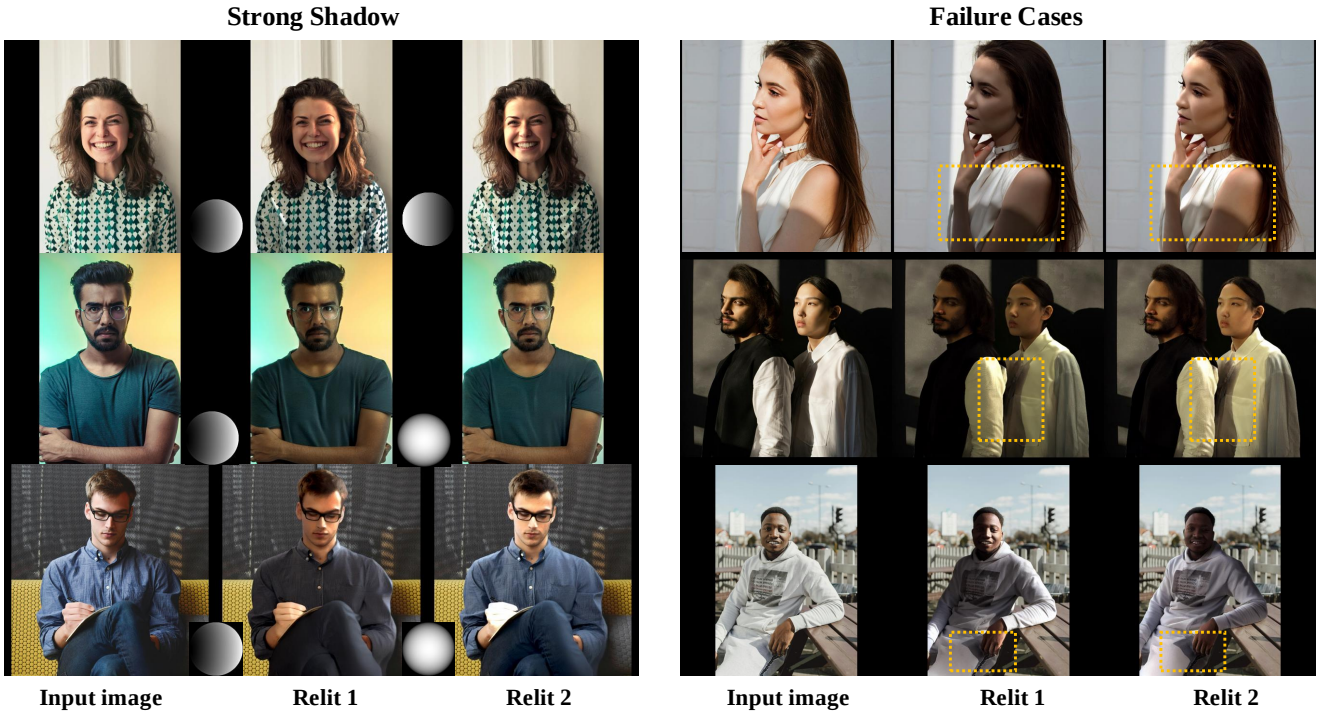

Figure 10. Strong shadow testing results (left) and failure cases (right) on real images from Pexels [8].

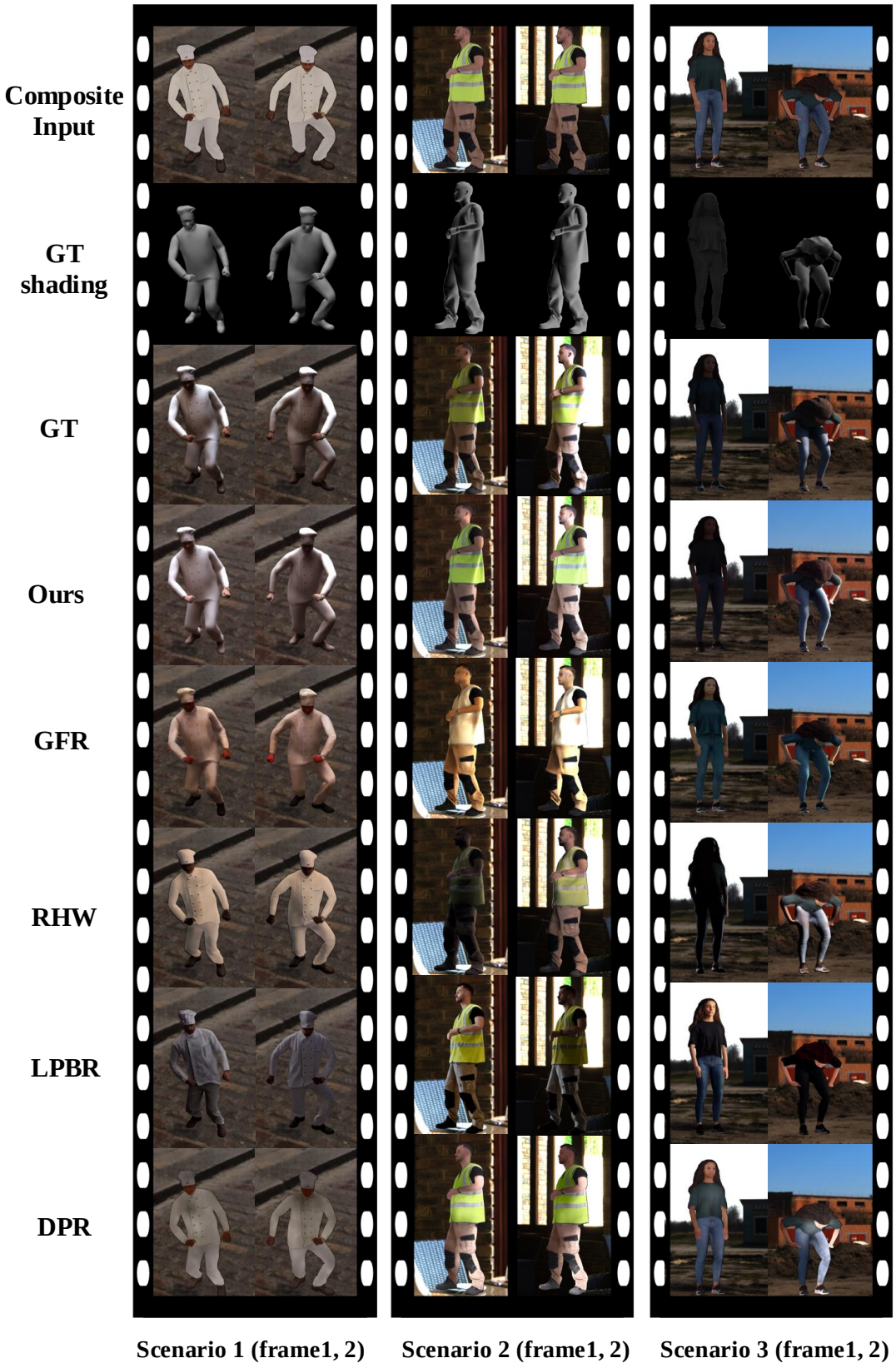

Figure 11. Video relighting comparison results on synthetic testing data: from left to right, we show comparison results for Scenario 1, 2, 3. From top to bottom, the first row shows the composite input (foreground human albedo composited with background image), the second row shows the ground truth (GT) shading, and the third row shows the GT image.

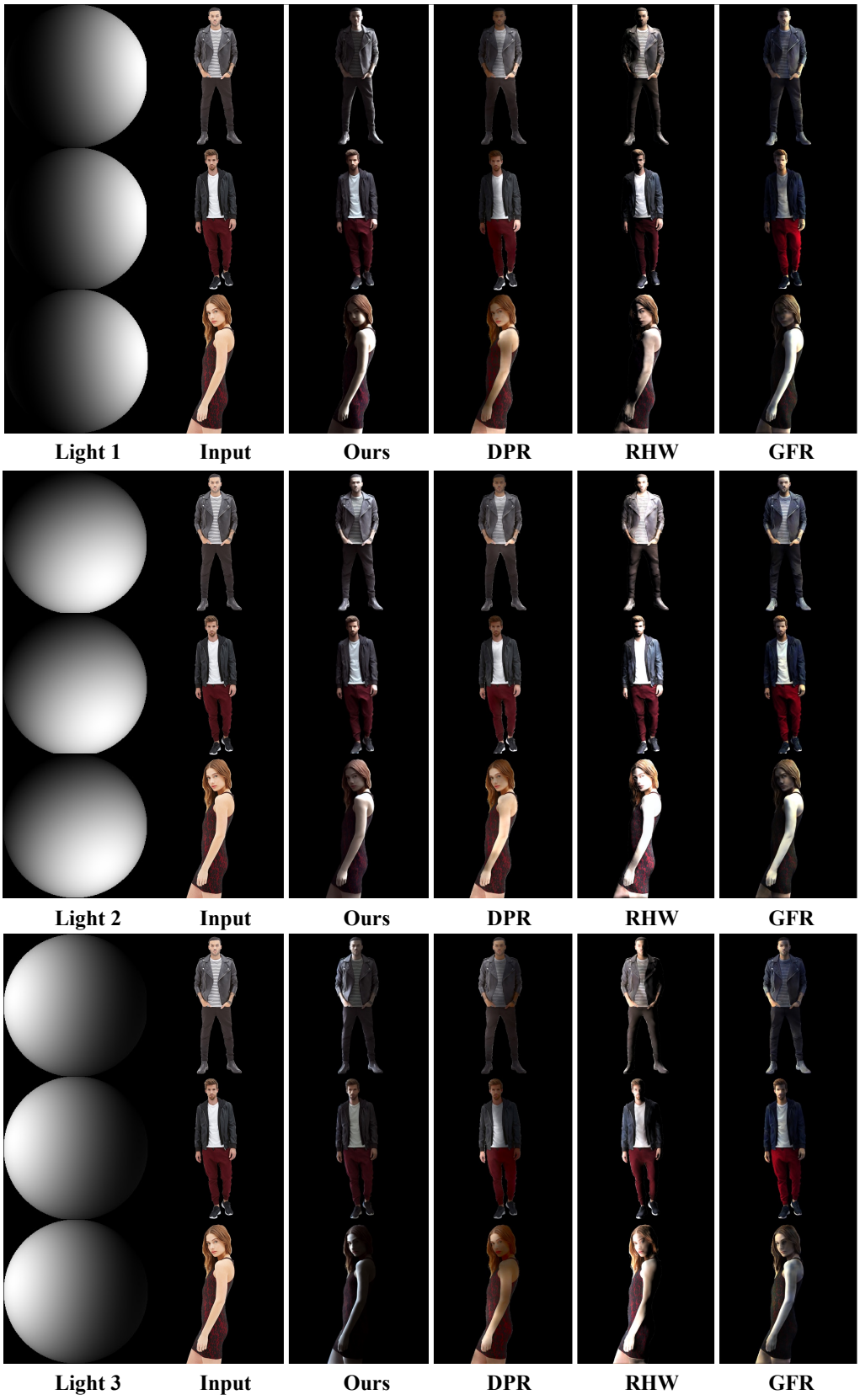

Figure 12. Real image comparisons with other human relighting approaches on the DeepFashion dataset [6]. We test on different identities and body parts (full body, half body). Our model shows consistent and feasible relighting with varying target lighting parameters (Spherical harmonics).

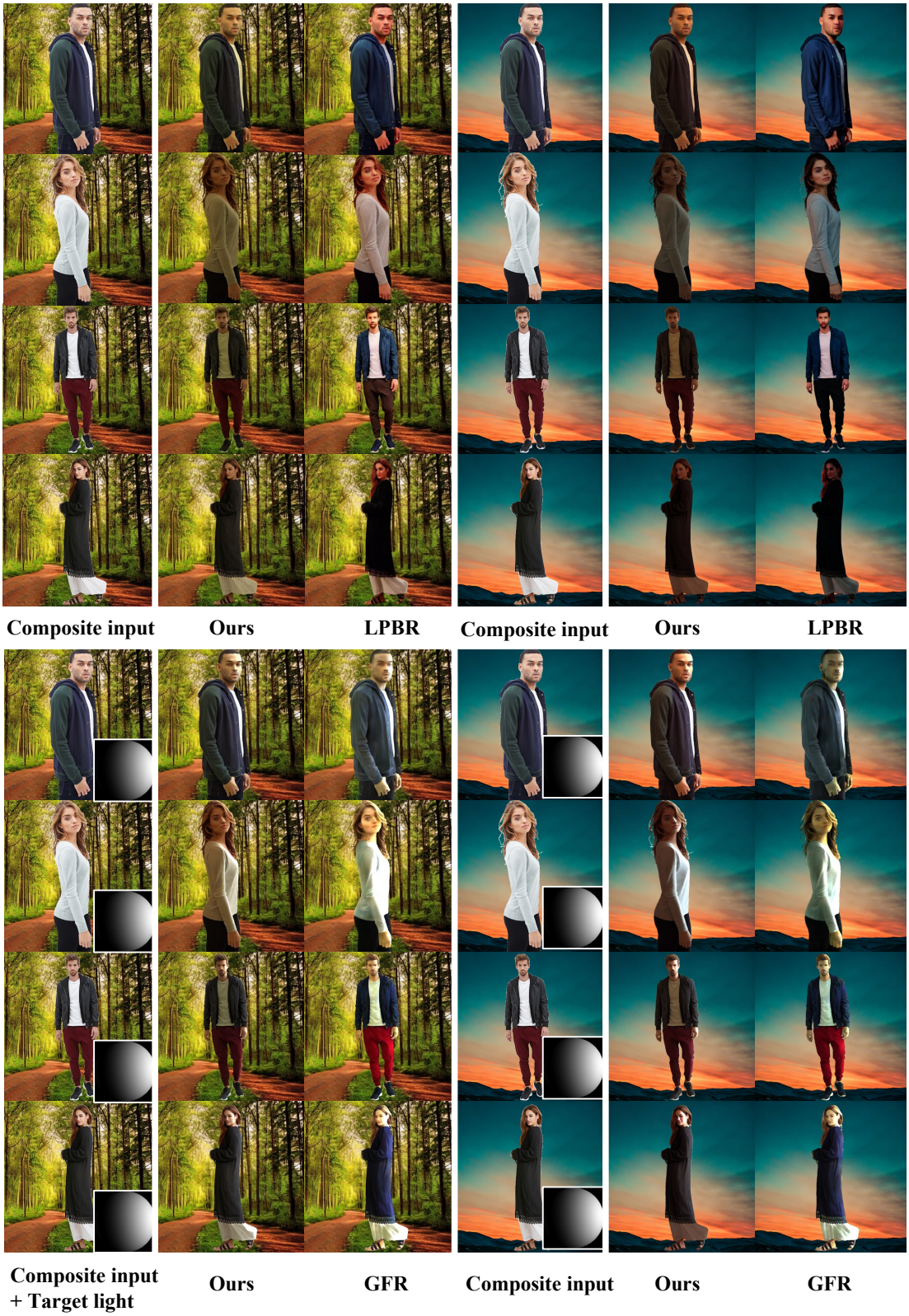

Figure 13. We present real image comparisons with the harmonization method. Given a composite input image, our model can achieve effective harmonization. When provided with target lighting parameters (Spherical harmonics), our model can achieve both background harmonization and relighting. The top section displays the outputs of our background harmonization method compared to the results from [11]. The lower section presents harmonization and relighting comparisons with [4]. Due to the higher generative prior of LPBR, noticeable distortions are present on the human face. Although GFR can achieve both harmonization and relighting, it exhibits obvious color noise.

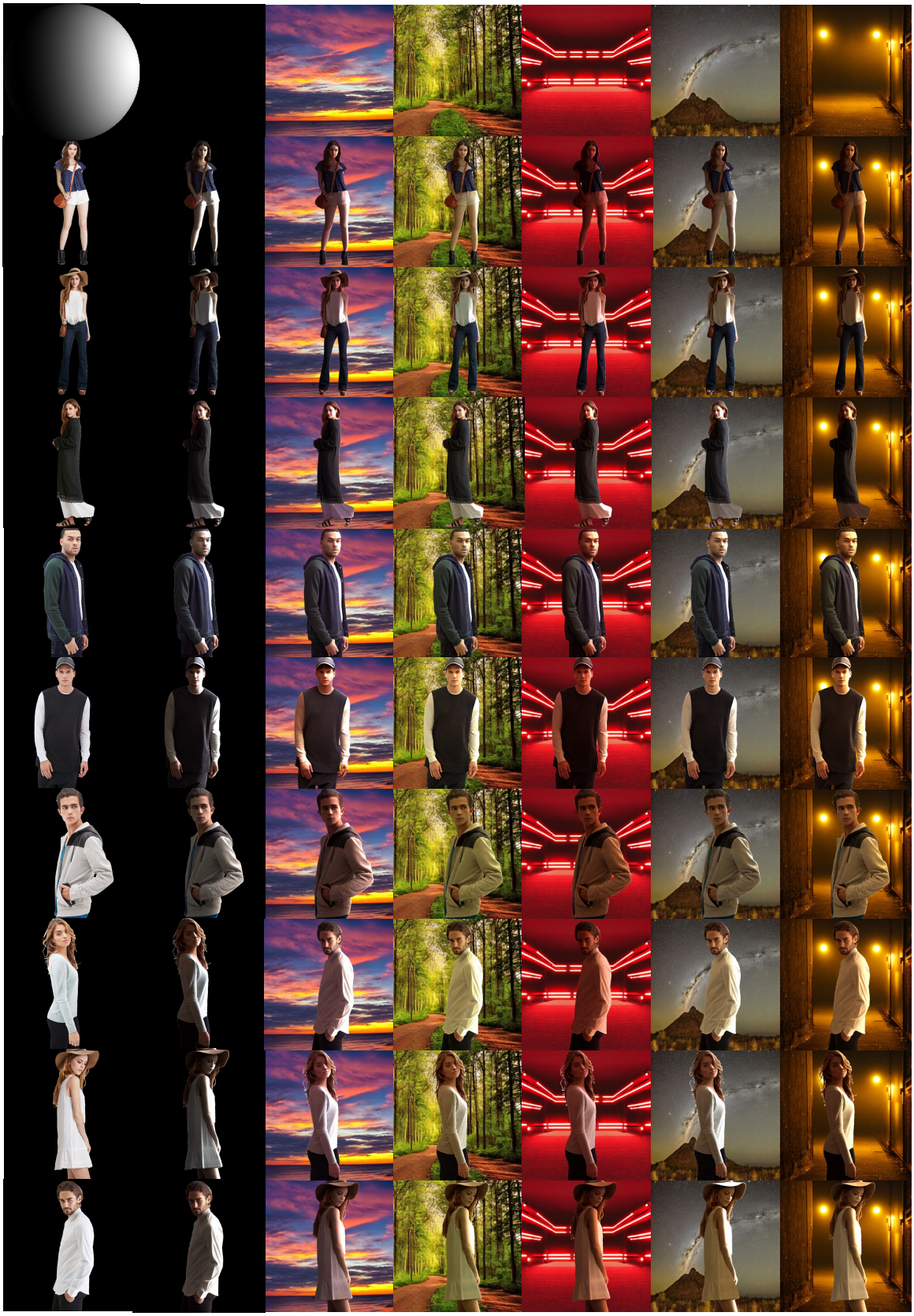

Figure 15. Our model can achieve realistic relighting with lighting 1 and background harmonization.

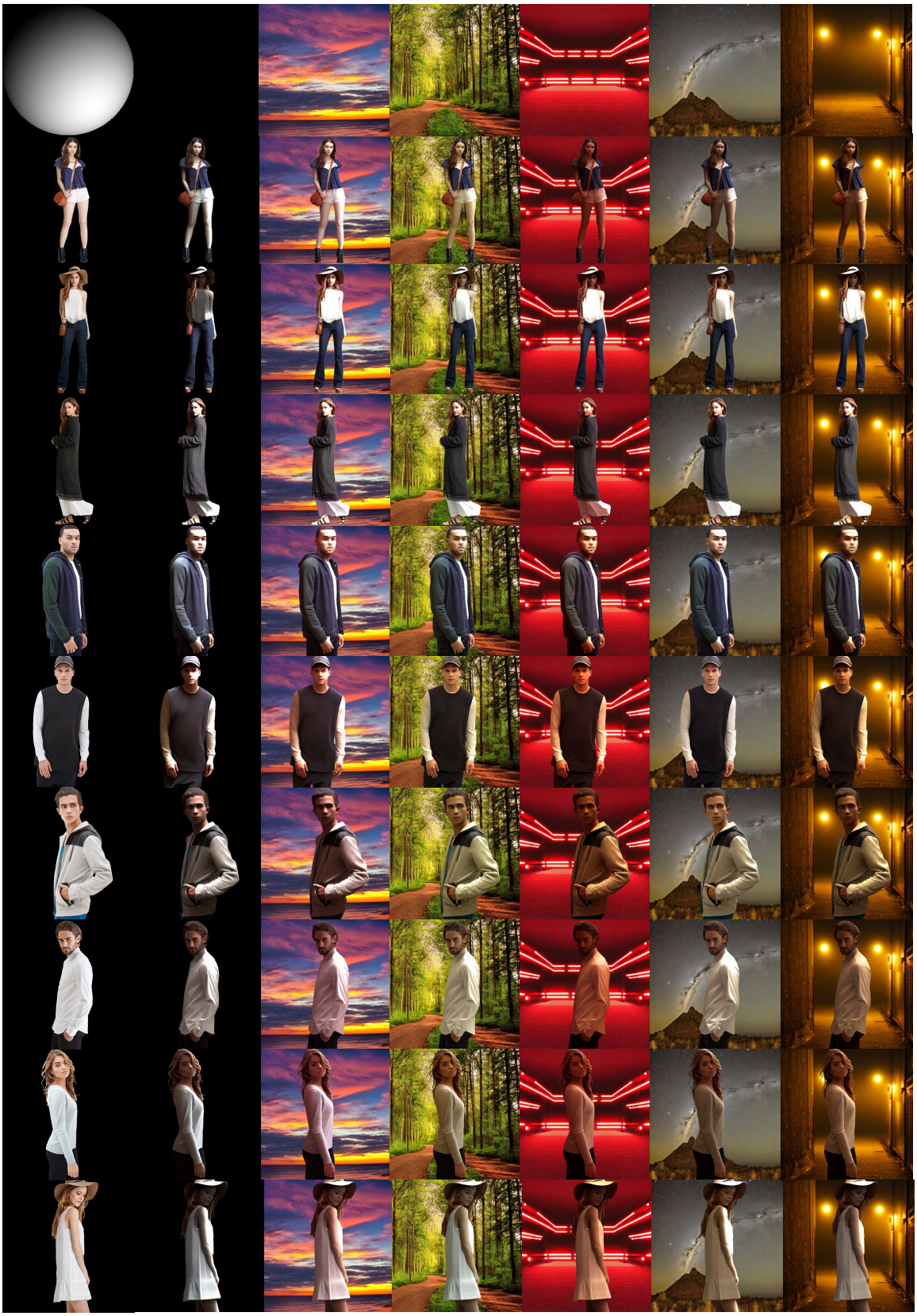

Figure 16. Our model can achieve realistic relighting with lighting 2 and background harmonization.

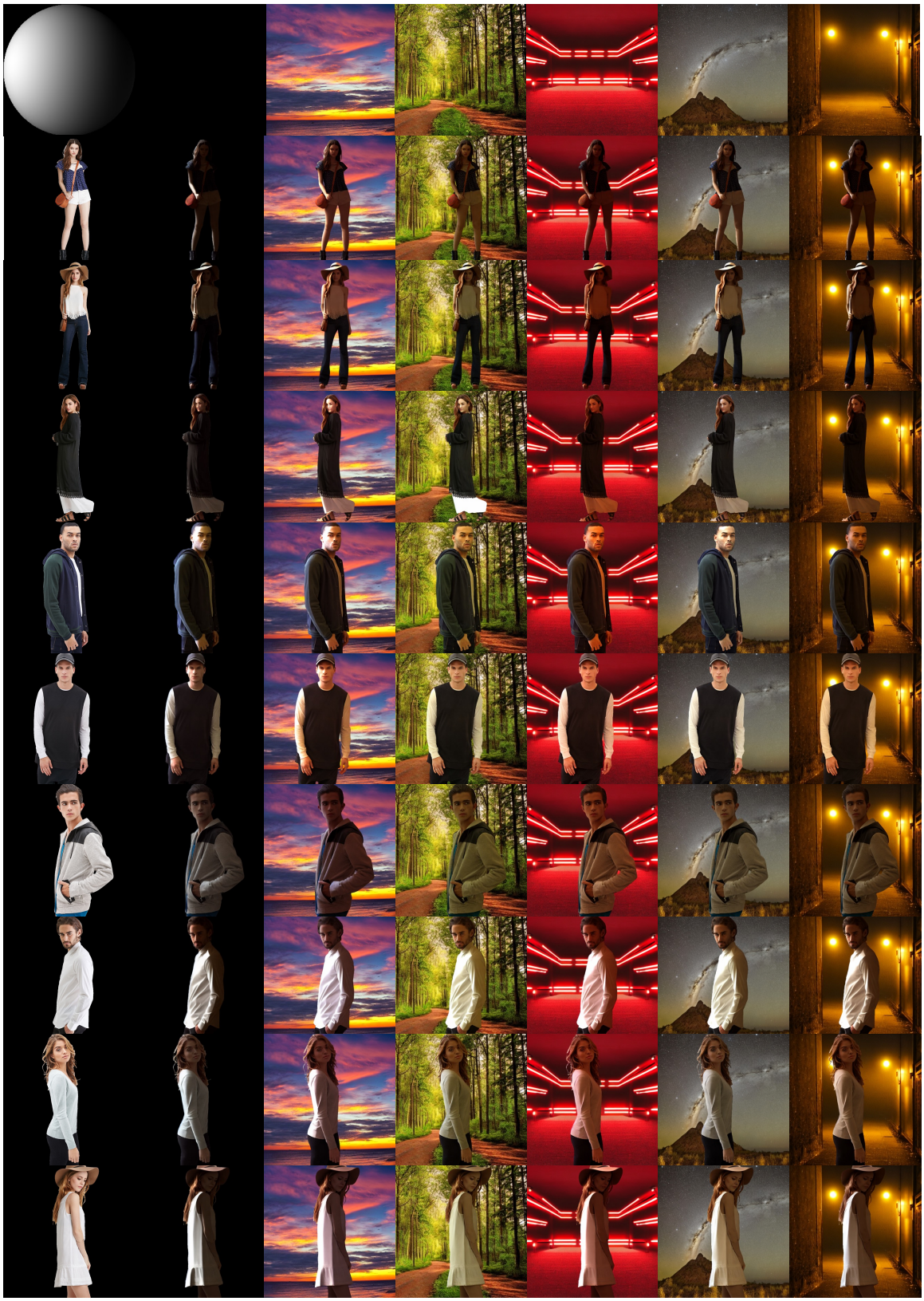

Figure 17. Our model can achieve realistic relighting with lighting 3 and background harmonization.