Supplemental File:
Lux Post Facto: Learning Portrait Performance Relighting with Conditional Video Diffusion and a Hybrid Dataset

Yiqun Mei[1,2] Mingming He[1] Li Ma[1] Julien Philip[1] Wenqi Xian[1] David M George[1]
Xueming Yu[1] Gabriel Dedic[1] Ahmet Levent Taşel[1] Ning Yu[1] Vishal M. Patel[2] Paul Debevec[1]
[1] Netflix Eyeline Studios [2] Johns Hopkins University

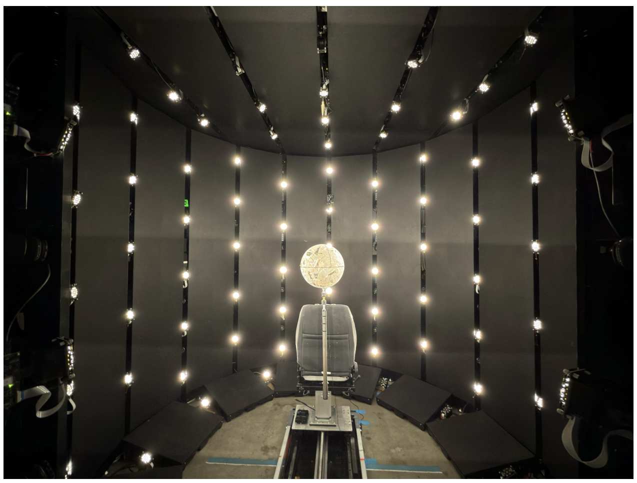


Figure 1. An illustration of our light stage.

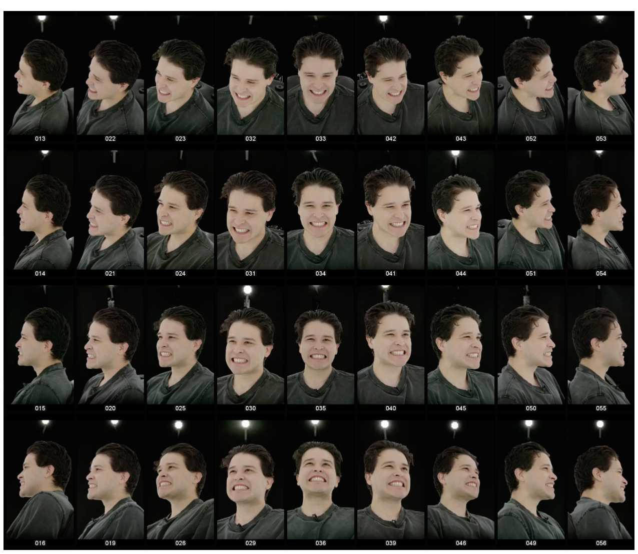


Figure 2. Examples of captured views using 36 frontal cameras.

1. Video Demonstration

We encourage readers to view the provided supplemental video, which contains video results and comparisons, for a more comprehensive illustration of the relighting quality of *Lux Post Facto*.

Figure 3. Visual comparisons for the ablation study on lighting control module. Compared to our design, the common used CLIP-based image encoding [10, 19, 24] cannot accurately capture the lighting intensity and directions in an HDR map and thus fails to enable precise lighting control. In contrast, our approach can produce high-quality lighting effects that follow the given HDR map.

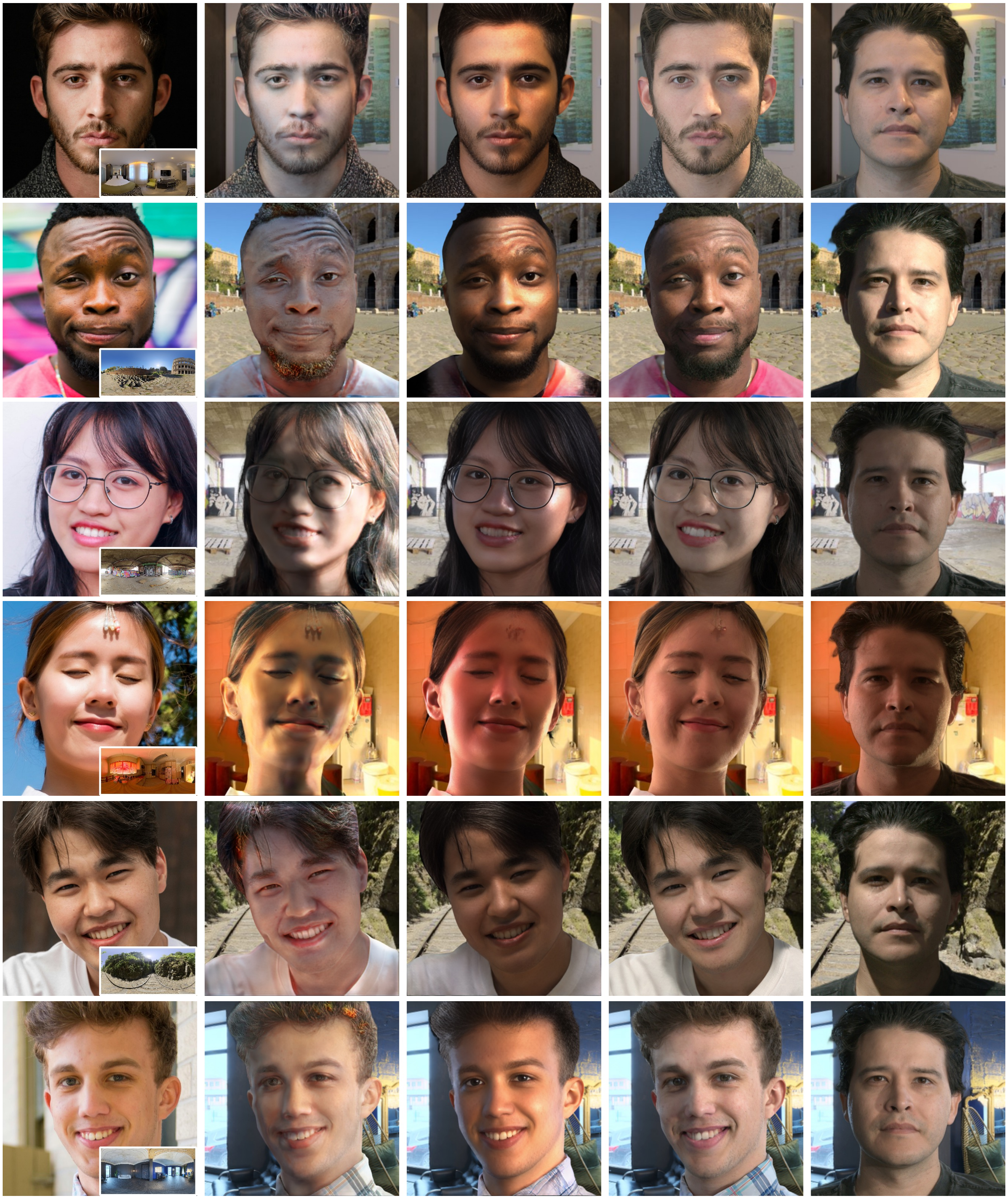
2. Additional Implementation Details

Reference Frame Encoding. *Lux Post Facto* is jointly trained with HDR-based relighting and reference-based appearance copy. To condition the model on a reference frame, we adopt a simple CNN encoder that encodes the image ($512 \times 512 \times 3$) into a feature map of $32 \times 32 \times 768$. We reshape it as a list of image embeddings (*i.e*. 1024×768) and append them after the lighting embeddings. When one conditioning (*e.g*. HDR-based) is used, we deactivate the other conditioning (*e.g*. reference-based) by replacing their embeddings with "null" embeddings.

Figure 4. Visual comparisons with video relighting methods on in-the-wild portrait videos. We compare our method with NVPR [21] and SwitchLight [5].

Image Delighting Model. We use an image delighting model to create paired training samples for the motion-rich dataset $\mathcal{D}_m$. We implement this model based on *Stable Diffusion (SD)* [11]. The model is extended to be spatially conditioned on an input image by adding additional input channels to the first convolution layer of the denoising U-Net, similar to our video model, and the text embeddings are replaced by "null" embeddings. We initialize the model weight from the pre-trained SD 1.5 [14], and supervisely train the model on our static OLAT dataset. We optimize the model towards v-prediction objectives [13] with a learning rate of 1e-5. The training stops after 200K steps.

More Training Details. To support autoregressive inference for long sequence, we randomly sample $T \in [0, 4]$ and replace the first T input frames with ground truth during training. This allows the model to learn to generate subsequent frames based on previous predictions, therefore enhances temporal consistency across prediction windows. In our implementation, we sample $T = 0$ with $p = 0.5$



Input PN-Relighting Holo-Relighting Ours Reference

Figure 5. Visual comparisons on in-the-wild image relighting. We compare our method with PN-Relighting [18] and Holo-Relighting [7]. Both approaches are designed for 512×512 face crops. Therefore, we report results on this region-of-interest for all methods for a fair comparison.

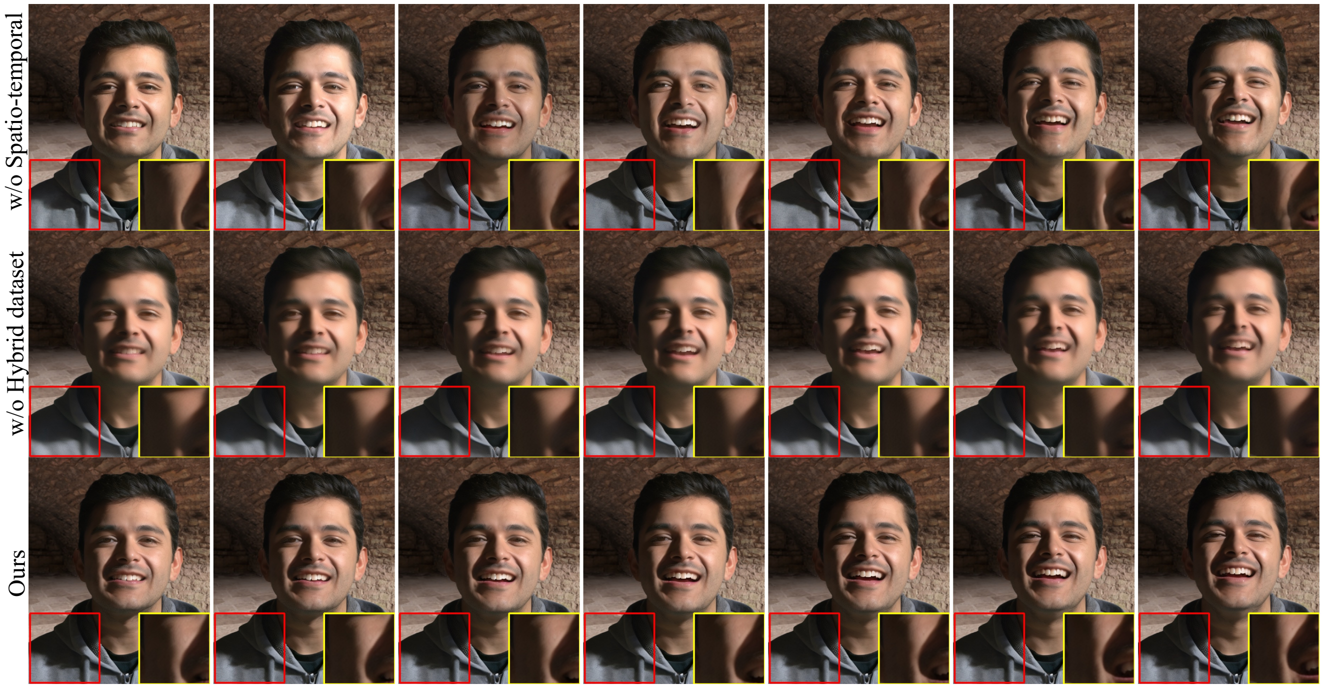


Figure 6. Additional visual evaluation on temporal consistency.

Figure 7. Relighting results under a rotating HDR map.

and other values equally with $p = 0.125$. Our method is implemented using PyTorch and trained on 8 NVIDIA A100 GPUs. During testing, results are generated using DDIM [15] sampler with 30 diffusion steps.

Light Stage and Rendering Details. We capture our static OLAT data using a light stage [3]. Specifically, the stage is configured as a cylindrical rig, equipped with 110 programmable LED lights and 75 Z-CAM e2 cinema cameras. We provide an illustration of the stage in Fig. 1. We use 36 frontal cameras for this project. Examples of the captured views are provided in Fig. 2. The stage has a diameter 2.7m and is 2.5m tall. The OLAT images are captured at 4K resolution. We cropped the upper body region and resize it to a resolution of 512×768 for training. During rendering, we randomly pair each OLAT sequence with multiple HDR maps and obtain lit images using image-based relighting [3, 12]. To diversify the illuminations, we augment an HDR map by randomly rotating it. Following [7], we further add the original OLAT images into our rendered dataset.

Table 1. Quantitative evaluation on temporal consistency.

Methods	NIQE↓	LE↓	LI↓	LPIPS (temp.)↓	WE↓
w/o Spatio-temporal	5.471	0.5542	0.0796	0.0450	0.0013
w/o Hybrid dataset.	6.639	0.5233	0.0387	0.0081	0.0001
Ours	**5.462**	**0.4978**	**0.0350**	**0.0073**	**0.0001**

The use and collection of the OLAT data were reviewed and approved by the Institutional Review Board (IRB) and informed consent was obtained from all participants.

3. More Results for Ablation Study

We provide visual results for the ablation study on lighting control in Fig. 3. As shown, commonly used CLIP-based image encoding [10, 19, 24] cannot enable precise lighting control, whereas our lighting conditioning approach can produce high-fidelity lighting effects that follow the given HDR map.

We also provide additional evaluation on temporal con-

Figure 8. Visual comparisons to background-based relighting method IC-Light [22]. IC-Light produces results with artifacts and struggles with synthesizing precise lighting effects.

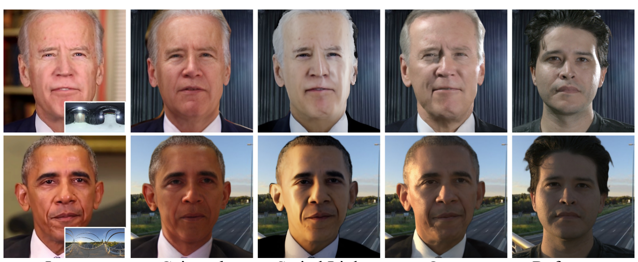


Figure 9. Visual comparisons to Cai *et al*. [2] and SwitchLight [5] on the INSTA dataset [27].

sistency. We conduct ablation study on two key designs: (1) the conditional video diffusion model (spatio-temporal design), which is trained using (2) hybrid dataset training strategy. These two designs together enable temporally consistent and high-quality relighting. We report results in Fig. 6 using a sequence of frames, and Tab. 1 using the image quality metric NIQE [8] and temporal metrics lighting error (LE), light instability (LI), LPIPS [23] between two adjacent frames and warping error (WE). Without spatio-temporal (*i.e*. video) design, the corresponding image diffusion model produces flickering lighting effects (see shadows on shoulder and cheek). With video modeling but without hybrid dataset training, the resulting video model (solely trained on OLAT simulated data $\mathcal{D}_l$) produces temporally smooth but blurry results.

4. Relighting under Rotating HDR Maps

To further demonstrate the effectiveness of the lighting control module, we report relighting results under a rotating HDR map. As shown in Fig. 7, our method can faithfully render lighting effects following the rotated HDR maps.

Table 2. Quantitative comparison with PN-Relighting and IC-Light on our test set.

Methods	LPIPS↓	NIQE↓	PSNR↑	SSIM↑
PN-Relighting [18]	0.2486	7.799	17.15	0.7373
IC-Light [22]	0.2519	6.996	16.12	0.7315
Ours	**0.1158**	**5.653**	**24.62**	**0.8278**

Table 3. Quantitative evaluation on INSTA dataset.

Methods	NIQE↓	LE↓	LI↓	LPIPS (temporal)↓	WE↓
Cai *et al.*	6.582	0.6521	0.1495	0.0206	0.0002
SwitchLight	7.107	0.5987	0.0822	0.0128	0.0001
Ours	**5.953**	**0.5239**	**0.0451**	**0.0092**	**0.0001**

5. More Comparison Results

In Fig. 4, we provide more visual comparisons against video relighting methods [5, 21] on in-the-wild portrait videos. For NVPR [21], we acquire the results from the authors as their code is not available. For SwitchLight [5], we obtain their results by using their commercial application [1].

In Fig. 5, we provide additional comparisons with two state-of-the-art face relighting method PN-Relighting [18] and Holo-Relighting [7]. PN-Relighting also uses the concept of data mixing but for a different goal (i.e. improving image relighting quality and albedo prediction) and via a different self-supervision approach. In contrast, we use data mixing for learning temporal consistent video relighting. The results for Holo-Relighting [7] are acquired from their authors as the source code is not available. Both approaches are designed for 512×512 face crops. Therefore, we report results on this region-of-interest for all methods for a fair comparison. Our methods generate more faithful relighting results, and the produced lighting effects are more consistent to the lighting effects in reference images.

In Fig. 8, we provide additional comparisons with background-based relighting method IC-Light [22] for image relighting. Compared to our method, IC-Light produces artifacts and fails to render precise lighting effects specified in the target HDR map. In Tab. 2, we report quantitative comparison with PN-Relighting [18] and IC-Light [22] on our test set.

In Fig. 9 and Tab. 3, we additionally compare our method with Cai *et al*. [2] and SwitchLight [5] on the INSTA dataset [27]. Note that INSTA dataset is designed for avatar reconstruction rather than evaluating video relighting performance. It may not best reflect the relighting capability as 1. it only contains a small number of subjects with limited input lighting and lack of large motions; 2. the videos are compressed, resulting in smoothed facial details in the input frames. On this dataset, our method achieves the best results both in terms of relighting quality and temporal consistency.

6. Limitations and Future Work

Lux Post Facto is not without limitations. First, although our model can robustly handle most accessories, we found a few challenging cases where accessories, such as the decorative hairpiece shown in Fig. 5 (row 4, column 4), partially occlude the face. In such scenario, the model may not perfectly preserve the accessory's details. This is because our training dataset lacks examples of faces with such occlusions, making it difficult for the model to handle this unseen case effectively. Second, as our model learns to synthesize lighting from the OLAT renderings, it can only generate lighting effects that can be represented by the light stage. Similar to previous methods [5–7, 9, 17], some challenging lighting effects (*e.g*. foreign shadows) cannot be produced by our approach. Third, *Lux Post Facto* relies on video diffusion models to generate relit videos. The iterative nature of the diffusion process makes it challenging to apply our method for real-time applications. Further improving run-time efficiency might be a very interesting direction for future work. Some possible solutions include designing more efficient architectures [25] or exploring distillation techniques [16, 20] to reduce sampling steps. We leave this direction for future work. Finally, due to GPU memory constraint, we train our model at a resolution of 512×768. To support higher-resolution generation, one possible way is to utilize an off-the-shelf super-resolution model (*e.g*. [4, 26]) as a post-processing step. We leave such exploration for future work.

7. Potential Negative Social Impacts

This method is designed to facilitate content creators to create creative and compelling lighting in portrait videos. However, we acknowledge its potential misuse, such as creating deepfakes or misleading videos. Our work is developed to support positive and creative applications. To mitigate misuse of our relighting method, we advocate for responsible usage, clear content labeling and implementing robust detection mechanisms.

References

[1] SwitchLight Desktop Application. Switchlight application. https://www.switchlight.beeble.ai/.

[2] Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6221–6231. IEEE, 2024.

[3] Paul E. Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, pages 145–156. ACM, 2000.

[4] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *CoRR*, abs/2407.07667, 2024.

[5] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 25096–25106. IEEE, 2024.

[6] Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, Hyunjoon Jung, and Vishal M. Patel. Lightpainter: Interactive portrait relighting with freehand scribble. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 195–205. IEEE, 2023.

[7] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M. Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 4263–4273. IEEE, 2024.

[8] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

[9] Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul E. Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4):43:1–43:21, 2021.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[12] Mark Sagar. Reflectance field rendering of human faces for "spider-man 2". In *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2005, Los Angeles, California, USA, July 31 - August 4, 2005, Courses*, page 14. ACM, 2005.

[13] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[14] SD15. Stable diffusionv-1-5. https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5.

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[16] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 32211–32252. PMLR, 2023.

[17] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E. Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79:1–79:12, 2019.

[18] Youjia Wang, Kai He, Taotao Zhou, Kaixin Yao, Nianyi Li, Lan Xu, and Jingyi Yu. Free-view face relighting using a hybrid parametric neural model on a SMALL-OLAT dataset. *Int. J. Comput. Vis.*, 131(4):1002–1021, 2023.

[19] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[20] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6613–6623. IEEE, 2024.

[21] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 782–792. IEEE, 2021.

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Ic-light github page, 2024.

[23] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018.

[24] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Yang Zhao, Yanwu Xu, Zhisheng Xiao, Haolin Jia, and Tingbo Hou. Mobilediffusion: Instant text-to-image generation on mobile devices. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXII*, pages 225–242. Springer, 2024.

[26] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 2535–2545. IEEE, 2024.

[27] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4574–4584. IEEE, 2023.